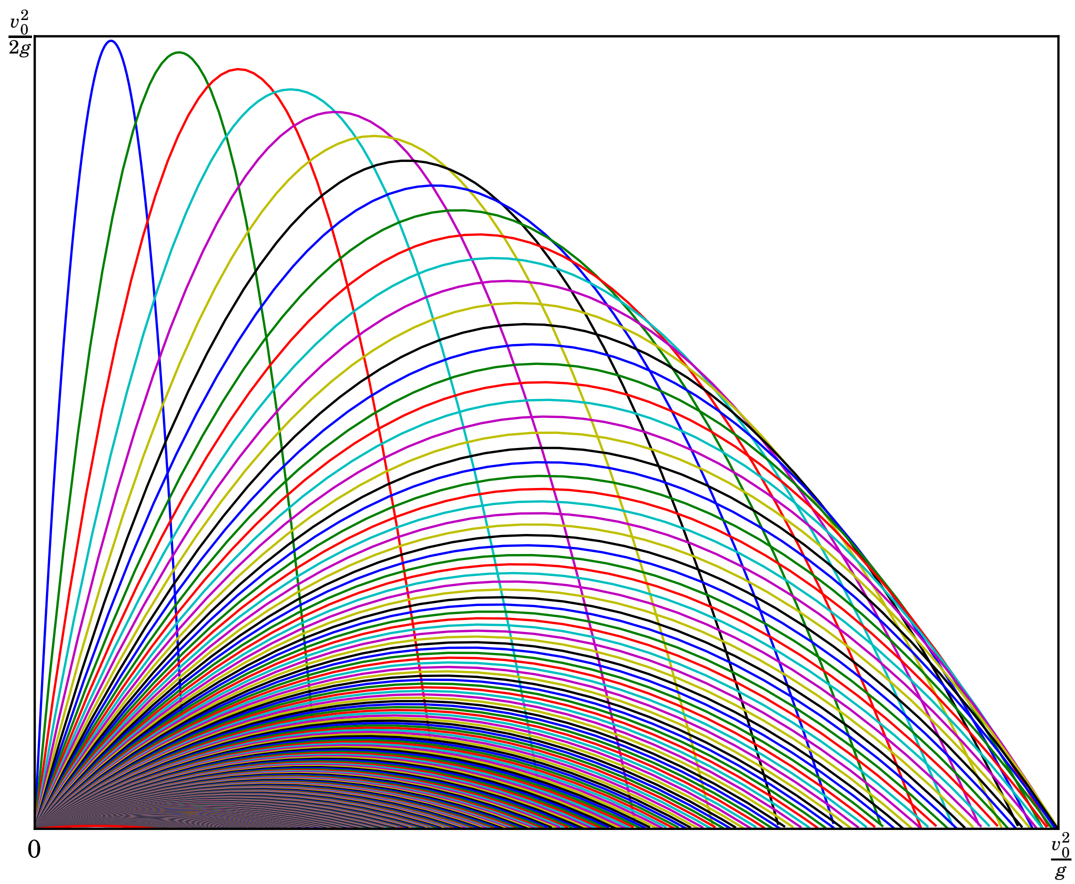
Elementary Calculus



Michael Corral

Elementary Calculus

Michael Corral

Schoolcraft College

About the author:
Michael Corral is an Adjunct Faculty member of the Department of Mathematics at Schoolcraft College. He received a B.A. in Mathematics from the University of California, Berkeley, and received an M.A. in Mathematics and an M.S. in Industrial & Operations Engineering from the University of Michigan.

This text was typeset in LaTeX with the KOMA-Script bundle, using the GNU Emacs text editor on a Fedora Linux system. The graphics were created using TikZ and Gnuplot.

Copyright © 2020 Michael Corral.
Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.

Preface

This book covers calculus of a single variable. It is suitable for a year-long (or two-semester) course, normally known as Calculus I and II in the United States. The prerequisites are high school or college algebra, geometry and trigonometry. The book is designed for students in engineering, physics, mathematics, chemistry and other sciences.

One reason for writing this text was because I had already written its sequel, *Vector Calculus*. More importantly, I was dissatisfied with the current crop of calculus textbooks, which I feel are bloated and keep moving further away from the subject's roots in physics. In addition, many of the intuitive approaches and techniques from the early days of calculus—which I think often yield more insights for students—seem to have been lost.

I agree with the views of the late Russian mathematician V.I. Arnold on teaching mathematics, in particular the idea that "Mathematics is the part of physics where experiments are cheap."[1] The ties to physics are especially important in calculus, so this book tries to introduce new concepts with physical motivations (what other motivations can there be?). The book contains exercises and examples that I hope will adequately prepare students who continue on in physics and engineering.[2]

Perhaps controversially, the book uses infinitesimals, making it a bit of a "throwback" or "retro" calculus text. My justification for this heretical act was purely pedagogical: infinitesimals make learning calculus easier, and their use aligns more with the way students will see calculus in their physics, chemistry and other science classes and textbooks (where infinitesimals are employed liberally). This might ruffle some feathers among mathematical "purists," but they are not the main audience for this book. That said, the book is still compatible with the usual limit-based approach, so an instructor could simply ignore the parts involving infinitesimals and teach the material as he or she normally would. I did not want to be dogmatic, so I used infinitesimals where I thought it made sense, and used limits where appropriate (e.g. in discussing continuity, series). Again, pedagogy was my priority.

The exercises at the end of each section are divided into three categories: A, B and C. The A exercises are mostly of a routine computational nature, the B exercises are slightly more involved, and the C exercises usually require some effort or insight to solve. A crude way of

[1] ARNOLD, V.I., "On Teaching Mathematics", *Russian Math. Surveys* 53 (1998), No. 1, 229-236. An HTML version is at `https://www.uni-muenster.de/Physik.TP/~munsteg/arnold.html`

[2] The book covers some of the types of problems and techniques for solving them that such students will likely encounter. Facility with using named constants (e.g. c, h, T) is also emphasized.

describing A, B and C would be "Easy", "Moderate" and "Challenging", respectively. However, many of the B exercises are easy and not all the C exercises are difficult. Appendix A provides answers and hints to many of the odd-numbered and some of the even-numbered exercises.

A few exercises require the student to write a computer program to solve numerical approximation problems (e.g. numerical methods for approximating definite integrals). Algorithms are presented in pseudocode, with code implementations in various languages (primarily Java, but also Python, Octave, Sage). I hope the code comments will help the reader figure out what is being done, regardless of familiarity with those languages. Students are free to implement solutions using the language of their choice. There are no dedicated "calculator exercises," as those have been rendered pointless by modern computing (with which students need to become acquainted).

Stylistically I made a conscious effort to break from an unfortunate but all too common mode of writing in mathematics texts, lamented in the preface of a physics book: "Nothing is more repellent to normal human beings than the clinical succession of definitions, axioms, and theorems generated by the labours of pure mathematicians."[3] I have been guilty of that sin myself, but I have changed my ways and banished all traces of that sort of thing from this book. So you won't find Definition 1.2, Theorem 3.3, Corollary 4.6, Lemma 5.7, Axiom 1B, etc. Instead, I tried to borrow the best of the styles from the physics and foreign languages textbooks I enjoyed so much in college. I also deliberately avoided what the author Gore Vidal called the "we-ness" that prevails in academic writing. There is no good reason for the "royal we" in a textbook, and it comes off as a bit pompous, so *we* won't use it.

This book is released under the GNU Free Documentation License (GFDL), which allows others to not only copy and distribute the book but also to modify it. For more details, see the included copy of the GFDL. So that there is no ambiguity on this matter, anyone can make as many copies of this book as desired and distribute it as desired, without needing my permission. The PDF version will always be freely available to the public at no cost (go to `http://www.mecmath.net/calculus`). Feel free to contact me at `mcorral@schoolcraft.edu` for any questions on this or any other matter regarding the book. I welcome your feedback.

Schoolcraft College MICHAEL CORRAL
December 2020

[3] ZIMAN, J.M., *Elements of Advanced Quantum Theory*, Cambridge, U.K.: Cambridge University Press, 1969.

Contents

The Greek Alphabet

Letters		Name	Letters		Name	Letters		Name
A	α	alpha	I	ι	iota	P	ρ	rho
B	β	beta	K	κ	kappa	Σ	σ	sigma
Γ	γ	gamma	Λ	λ	lambda	T	τ	tau
Δ	δ	delta	M	μ	mu	Υ	υ	upsilon
E	ϵ	epsilon	N	ν	nu	Φ	ϕ	phi
Z	ζ	zeta	Ξ	ξ	xi	X	χ	chi
H	η	eta	O	o	omicron	Ψ	ψ	psi
Θ	θ	theta	Π	π	pi	Ω	ω	omega

Mathematical Notation

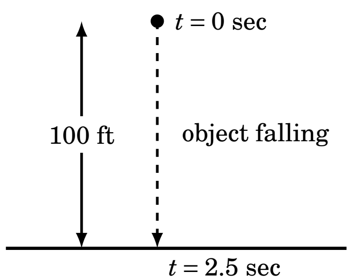
Symbol	Meaning	Example
$\Rightarrow$	if...then; implies	$\lvert x\rvert > 1 \Rightarrow x^2 > 1$
$\Leftrightarrow$	if and only if; two-way implication	$\lvert x\rvert > 1 \Leftrightarrow x^2 > 1$
iff	if and only if; two-way implication	$\lvert x\rvert > 1$ iff $x^2 > 1$
$\not\Rightarrow$	does not imply	$\lvert x\rvert > 1 \not\Rightarrow x > 1$
$\exists$	there exists	$\exists$ a number $c > 0$
$\nexists$	there does not exist	$\nexists$ x such that $x^2 < 0$
$\exists!$	there exists a unique	$\exists!$ x such that $2x - 1 = 3$
$\forall$	for every	$\forall x \geq 0$, $\sqrt{x}$ is a real number
$\equiv$	is identically equal to	$f \equiv 0 \Rightarrow f(x) = 0$ for all x
$\propto$	is proportional to	$y \propto x^2 \Rightarrow y = kx^2$ for some k
$\subseteq$	is a subset of	$\{0,1\} \subseteq \{0,1,2\}$
$\in$	is an element of	$1 \in \{1,2,3\}$
$\notin$	is not an element of	$1 \notin \{2,3\}$
$\cup$	union of sets	$\{0,1\} \cup \{2,3\} = \{0,1,2,3\}$
$\cap$	intersection of sets	$\{0,1\} \cap \{1,2\} = \{1\}$
$\varnothing$	empty set	$\{0,1\} \cap \{2,3\} = \varnothing$
$\therefore$	therefore	$\therefore$ n must exist

CHAPTER 1

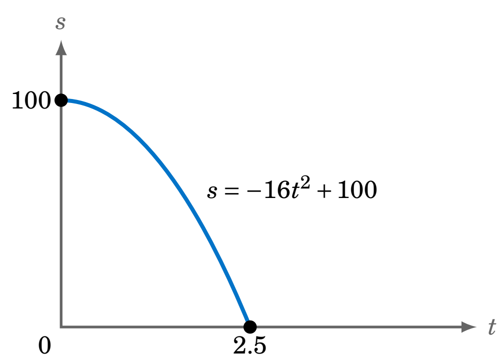
The Derivative

1.1 Introduction

Calculus can be thought of as the analysis of curved shapes.[1] Its development grew out of attempts to solve physical problems. For example, suppose that an object at rest 100 ft above the ground is dropped. Ignoring air resistance and wind, the object will fall straight down until it hits the ground (see Figure 1.1.1(a)). As will be proved later, t seconds after being dropped the object will be $s = s(t) = -16t^2 + 100$ ft above the ground. The object will thus hit the ground after 2.5 seconds (when $s = 0$). While the object's *path* is a straight line, the graph of its position s above the ground as a function of time t is curved, part of a parabola (see Figure 1.1.1(b)).



(a) Path of the object



(b) Position s as a function of time t

Figure 1.1.1 An object dropped from 100 ft above the ground

How fast is the object moving before it hits the ground? This is where calculus comes in. The solution, presented now, will motivate much of this chapter.

[1]It is more than that, of course, but that definition puts us in good company: the first European textbook on calculus, written by the French mathematician Guillaume de l'Hôpital in 1696, was titled *Analyse des Infiniment Petits pour l'Intelligence des Lignes Courbes* (which translates as *Analysis of the Infinitely Small for Understanding Curved Lines*). That book (in French) can be obtained freely in electronic form at `https://archive.org`

First, the object travels 100 ft in 2.5 seconds, so its **average speed** in that time is

$$\frac{\text{distance traveled}}{\text{time elapsed}} = \frac{100\text{ ft}}{2.5\text{ seconds}} = 40\text{ ft/s},$$

and its **average velocity** in that time is

$$\frac{\text{change in position}}{\text{change in time}} = \frac{\text{final position} - \text{initial position}}{\text{end time} - \text{start time}} = \frac{0\text{ ft} - 100\text{ ft}}{2.5\text{ sec} - 0\text{ sec}} = -40\text{ ft/s}.$$

Unlike speed, velocity takes direction into account. Thus, the object's downward motion means it has negative velocity. Positive velocity implies upward motion.

Using the idea of average velocity over an interval of time, there is a natural way to define the object's **instantaneous velocity** at a particular *instant* of time t:

1. Find the average velocity over an interval of time.

2. Let the interval become smaller and smaller indefinitely, shrinking to a point t. If the average velocity over that smaller and smaller interval approaches some value, call that value the instantaneous velocity at time t.

Figure 1.1.2 below shows how to choose the interval: for any time t between 0 and 2.5, use the interval $[t, t+\Delta t]$, where Δt (pronounced "delta t") is a small positive number. So Δt is the change in time over the interval; denote by Δs the change in the position s over that interval.

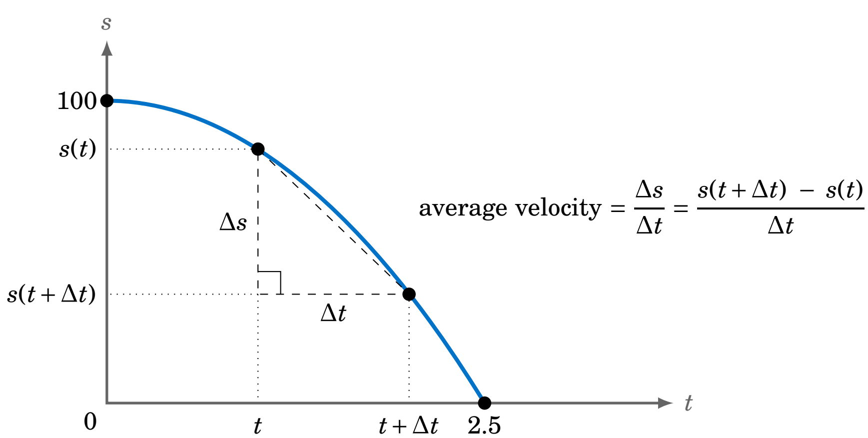


Figure 1.1.2 Average velocity $\frac{\Delta s}{\Delta t}$ over the interval $[t, t+\Delta t]$

The average velocity of the object over the interval $[t, t+\Delta t]$ is $\frac{\Delta s}{\Delta t}$, so since $s(t) = -16t^2 + 100$:

$$\begin{aligned}
\frac{\Delta s}{\Delta t} &= \frac{s(t+\Delta t) - s(t)}{\Delta t} \\
&= \frac{-16(t+\Delta t)^2 + 100 - (-16t^2 + 100)}{\Delta t} \\
&= \frac{-16t^2 - 32t\Delta t - 16(\Delta t)^2 + 100 + 16t^2 - 100}{\Delta t} \\
&= \frac{-32t\Delta t - 16(\Delta t)^2}{\Delta t} = \frac{\cancel{\Delta t}(-32t - 16\Delta t)}{\cancel{\Delta t}} \\
&= -32t - 16\Delta t\ ,
\end{aligned}$$

Now let the interval $[t, t+\Delta t]$ get smaller and smaller indefinitely—that is, let Δt get closer and closer to 0. Then the average velocity $\frac{\Delta s}{\Delta t} = -32t - 16\Delta t$ gets closer and closer to $-32t - 0 = -32t$. Thus, *the object has instantaneous velocity* $-32t$ *at time* t. This calculation can be interpreted as taking the **limit** of $\frac{\Delta s}{\Delta t}$ *as* Δt *approaches* 0, written as follows:

$$\begin{aligned}
\text{instantaneous velocity at } t &= \text{limit of average velocity over } [t, t+\Delta t] \text{ as } \Delta t \text{ approaches to } 0 \\
&= \lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t} \\
&= \lim_{\Delta t \to 0} (-32t - 16\Delta t) \\
&= -32t - 16(0) \\
&= -32t
\end{aligned}$$

Notice that Δt is not replaced by 0 in the ratio $\frac{\Delta s}{\Delta t}$ until *after* doing as much cancellation as possible. Notice also that the instantaneous velocity of the object varies with t, as it should (why?). In particular, at the instant when the object hits the ground at time $t = 2.5$ sec, the instantaneous velocity is $-32(2.5) = -80$ ft/s.

If this makes sense so far, then you understand the crux of the idea of what a limit is and how to calculate a limit. The instantaneous velocity $v(t) = -32t$ is called the **derivative** of the position function $s(t) = -16t^2 + 100$. Calculating derivatives, analyzing their properties, and using them to solve various problems are part of *differential calculus*.

What does this have to do with curved shapes? Instantaneous velocity is a special case of an **instantaneous rate of change** of a function; in this case the instantaneous rate of change of the position (height above the ground) of the object. Similar to how the rate of change of a line is its slope, the instantaneous rate of change of a general curve represents the **slope of the curve**. For example, the parabola $s(t) = -16t^2 + 100$ has slope $-32t$ for all t. Note that the slope of this curve varies (as a function of t), unlike the slope of a straight line.

Finding the area inside curved regions is another type of problem that calculus can solve. The basic idea is to use simpler regions—rectangles—whose areas are known, then use those to approximate the area inside the curved region. One such method is to draw more and more rectangles of diminishing widths *inside* the curved region,[2] so that the sums of their areas approach the area of the curved region. Figure 1.1.3 shows an example with four rectangles to approximate the area under a curve $y = f(x)$ over an interval $[a,b]$ on which $f(x) \geq 0$.

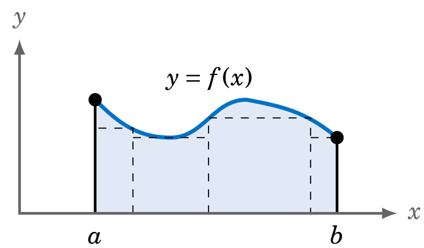


Figure 1.1.3 The area of a curved region

The limit of these sums of rectangular areas is called an **integral**. The study and application of integrals are part of *integral calculus*. Perhaps the most remarkable result in calculus is that there is a connection between derivatives and integrals—the *Fundamental Theorem of Calculus*, discovered in the 17^{th} century, independently, by the two men who invented calculus as we know it: English physicist, astronomer and mathematician Isaac Newton (1642-1727) and German mathematician and philosopher Gottfried Wilhelm von Leibniz (1646-1716).

Calculus makes extensive use of *infinite sequences and series*. An **infinite series** is just a sum of an infinite number of terms. For example, it will be shown later in the text that

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots , \tag{1.1}$$

where the sum on the right involves an infinite number of terms. A **power series** is a particular type of infinite series applied to functions; it can be thought of as a polynomial of infinite degree. For example, the trigonometric function $\sin x$ does not appear to be a polynomial. But it turns out that $\sin x$ has a power series representation as

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots , \tag{1.2}$$

where again the sum continues infinitely, and the formula holds for all x (in radians).

The idea of replacing a function by its power series played an important role throughout the development of calculus, and is a powerful technique in many applications.

All the functions in this text will be functions of a single real variable—that is, the values that the variable can take are real numbers. Below is some standard notation for commonly-used sets of numbers:

[2]It will be shown later (in Chapter 5) that the rectangles do not have to be completely inside the region.

$\mathbb{N}$ = the set of all **natural** numbers, i.e. the set of nonnegative integers: $0, 1, 2, 3, 4, \ldots$

$\mathbb{Z}$ = the set of all integers: $0, \pm 1, \pm 2, \pm 3, \pm 4, \ldots$

$\mathbb{Q}$ = the set of all **rational** numbers $\frac{m}{n}$, where m and n are integers, with $n \neq 0$

$\mathbb{R}$ = the set of all real numbers

Note that $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$.

The set of real numbers consists of the rational numbers together with numbers that are not rational, called **irrational numbers**. For example, $\sqrt{2}$ is irrational. That is, 2 is not the square of a rational number. In fact, if the square of a rational number q were an integer, then q itself would have to be an integer: write q as m/n, where m and n are positive integers with no common positive integer divisors other than 1. Since $q^2 = m^2/n^2$ simply duplicates the integer divisors of m and n, then q^2 can be an integer only if $n = 1$, i.e. q is an integer. Clearly 2 is not the square of an integer, and thus it cannot be the square of a rational number. This argument also shows that $\sqrt{3}$, $\sqrt{5}$, $\sqrt{6}$, $\sqrt{7}$, $\sqrt{8}$, $\sqrt{10}$, and so on, are irrational.[3]

It turns out that there are far more irrational numbers–and hence real numbers—than rational numbers. In fact, whereas the rational numbers can be listed in a sequence (i.e. first, second, third, etc.), the set of real numbers cannot.[4] For example, in the closed interval $[0,1]$ there is no "next" real number after the number 0. Thus, some infinite sets are larger than others—$\mathbb{R}$ is larger than $\mathbb{Q}$. Intervals such as $[0,1]$ or $\mathbb{R}$ itself are examples of a *continuum* of objects, i.e. no gaps exist.[5] A famous unsolved problem in mathematics—the *Continuum Hypothesis*—is whether an infinite set exists that is larger in size than $\mathbb{Q}$ but smaller than $\mathbb{R}$.

Infinity is an important notion in calculus. Whether it is the idea of *infinitely large* or *infinitesimally small*, calculus attempts to give the idea some mathematical meaning (typically by way of limits).[6] The mathematical use of infinity has been a subject of philosophical debate.[7]

Though several centuries old, calculus was the beginning of *modern* mathematics. *Classical* mathematics (e.g. algebra, geometry, trigonometry)—whose origins date back to the ancient Babylonians, Egyptians, and Greeks—was concerned mostly with the study of *static* quantities. Calculus produced a way to analyze *dynamic* (i.e. changing) quantities. The period from the 17th through the 19th century also saw revolutionary advances in physics, chemistry, biology and other sciences. The birth of calculus was one part of that qualitative leap.

---

[3] This argument is due to the British philosopher Bertrand Russell (1872-1970). For an alternative proof that $\sqrt{2}$ is irrational, see pp. 97-98 in GELFAND, I.M. AND A. SHEN, *Algebra*, Boston: Birkhäuser, 1993.

[4] For a proof see Ch.1 in KAMKE, E., *Theory of Sets*, New York: Dover Publications, Inc., 1950.

[5] For a study of the structure of the real number system, see BURRILL, C.W., *Foundations of Real Numbers*, New York: McGraw-Hill Book Company, 1967.

[6] Not everyone agrees that calculus does this satisfactorily. For example, for an alternative development of basically the same material in "standard" calculus but without the use of limits—called *infinitesimal analysis*—see KEISLER, H.J., *Elementary Calculus: An Infinitesimal Approach*, Boston: Prindle, Weber & Schmidt, 1976.

[7] For example, see the essays by L. E. J. Brouwer, Hermann Weyl and David Hilbert in HEIJENOORT, J. VAN, *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931*, Cambridge, MA: Harvard University Press, 1967.

Exercises

A

For Exercises 1-4, suppose that an object moves in a straight line such that its position s after time t is the given function $s = s(t)$. Find the instantaneous velocity of the object at a general time $t \geq 0$. You should mimic the earlier example for the instantaneous velocity when $s = -16t^2 + 100$.

1. $s = t^2$ **2.** $s = 9.8t^2$ **3.** $s = -16t^2 + 2t$ **4.** $s = t^3$

5. By equation (1.1), $\pi = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots\right)$, where the n^{th} term in the sum inside the parentheses is $\frac{(-1)^{n+1}}{2n-1}$ (starting at $n = 1$).[8] So the first approximation of π using this formula is $\pi \approx 4(1) = 4.0$, and the second approximation is $\pi \approx 4\left(1 - \frac{1}{3}\right) = 8/3 \approx 2.66667$. Continue like this until two consecutive approximations have 3 as the first digit before the decimal point. How many terms in the sum did this require? Be careful with rounding off in the approximations.

B

6. In elementary geometry you learned that the area inside a circle of radius $r > 0$ is πr^2 (that formula will be proved later in the text). So in particular, let C be a circle of radius 1. Then the area inside C is π. That area can be approximated by *Eudoxus' method of exhaustion*.[9] The idea is to inscribe *regular polygons* inside the circle, i.e. the vertexes of the polygons touch C. Recall from geometry that a polygon is regular if its sides are of equal length. By increasing the number of sides of the polygons, the areas inside the polygons will approach the area (π) of C. This was an early attempt at using what is now called a limit.[10]

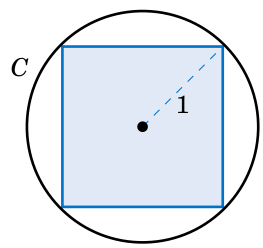


Figure 1.1.4 Inscribed square

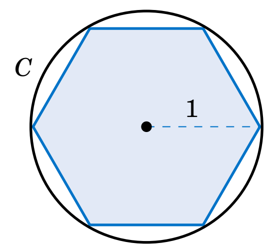


Figure 1.1.5 Inscribed regular hexagon

(a) Inscribe a square inside C, as in Figure 1.1.4. Show that the area inside the square is 2. This is a poor approximation of $\pi = 3.14159265...$, obviously.

(b) Inscribe a regular hexagon (6-sided) inside C, as in Figure 1.1.5. Show that the area inside the hexagon is $\frac{3\sqrt{3}}{2} \approx 2.59807621$. This is a slightly better—though still poor—approximation of π.

(c) Inscribe a regular dodecagon (12-sided) inside C. Show that the area inside the dodecagon is 3. It thus takes 12 sides for the approximation to get the first digit of π correct.

(d) Inscribe a regular 100-sided polygon inside C. Show that the area inside this polygon is approximately 3.13952598. This is getting closer to π.

(e) Show that the general formula for the area inside a regular n-sided polygon inscribed inside C is $\frac{n}{2} \sin\left(\frac{360°}{n}\right)$. *(Hint: The double-angle identity* $\sin 2\theta = 2 \sin\theta \cos\theta$ *might help.)*

---

[8] This, by the way, is a terrible formula for calculating π; getting just the 3.14 part requires 119 terms in the sum!

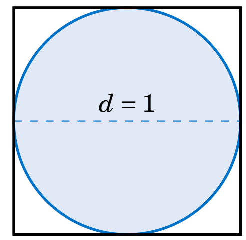
[9] Originally due to another ancient Greek mathematician, Antiphon (ca. 430 B.C.)

[10] The great ancient Greek mathematician, physicist and astronomer Archimedes (ca. 287-212 B.C.) used this method, together with circumscribed regular polygons, to calculate π.

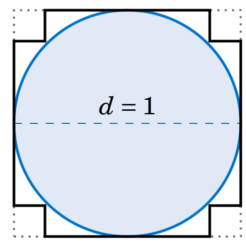
C

7. What is the flaw in the following "proof" that $\pi = 4$?:

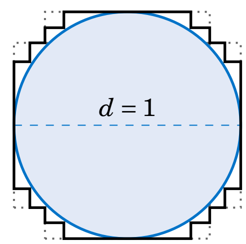
Step 1: Draw a square around a circle of diameter $d = 1$. The circumference of the circle is thus $\pi d = \pi$, and the perimeter of the square is 4.



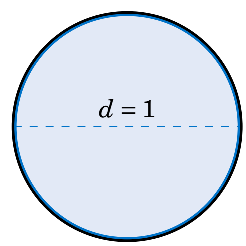
Step 2: Remove corners from the square as shown in the picture on the right, so that four new corners touch the circle. Notice that the perimeter of the resulting polygon is still 4, since the lengths of the removed corner pieces are duplicated in the new polygon, so that the lengths of all the vertical sides add up to 2 while the lengths of all the horizontal sides add up to 2.



Step 3: Remove corners from the polygon in Step 2, as shown in the picture on the right, so that eight new corners touch the circle. The perimeter of the resulting polygon is again still 4.



Step 4: Continue this procedure indefinitely, with each successive polygon still having a perimeter of 4 and becoming increasingly indistinguishable from the circle. Since the perimeters of the polygons always equal 4 and approach the circle's circumference (π), then π must equal 4.



8. An infinite set is **countable** if its members can be put into a one-to-one correspondence with the members of $\mathbb{N}$, the set of natural numbers $(0,1,2,3,4,\ldots)$. Clearly $\mathbb{N}$ is itself countable. The set $\mathbb{Z}$ of all integers is also countable, by means of the following one-to-one correspondence with $\mathbb{N}$:

$\mathbb{N}$	0	1	2	3	4	5	6	7	8	...
$\mathbb{Z}$	0	1	-1	2	-2	3	-3	4	-4	...

Show that $\mathbb{Q}$ (the set of all rational numbers) is countable. *(Hint: The above correspondence for $\mathbb{Z}$ is an infinite list in one dimension (the horizontal direction). For $\mathbb{Q}$ think two-dimensionally.)*

1.2 The Derivative: Limit Approach

The following definition generalizes the example from the previous section (concerning instantaneous velocity) to a general function $f(x)$:

The **derivative** of a real-valued function $f(x)$, denoted by $f'(x)$, is

$$f'(x) = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} \tag{1.3}$$

for x in the domain of f, provided that the limit exists.[11]

For a general function $f(x)$, the derivative $f'(x)$ represents the **instantaneous rate of change** of f at x, i.e. the rate at which f changes at the "instant" x. For the limit part of the definition only the intuitive idea of how to take a limit—as in the previous section—is needed for now. Notice that the above definition makes the derivative f' itself a function of the variable x. The function f' can be evaluated at specific values of x, or you can write its general formula $f'(x)$.

The (instantaneous) velocity of an object as the derivative of the object's position as a function of time is only one physical application of derivatives. There are many other examples:

Field	Function	Derivative
Physics	position	velocity
	velocity	acceleration
	momentum	force
	work	power

Field	Function	Derivative
Physics	angular momentum	torque
Engineering	electric charge	electric current
	magnetic flux	induced voltage
Economics	profit	marginal profit

The limit definition can be used for finding the derivatives of simple functions.

Example 1.1

Find the derivative of the function $f(x) = 1$.

Solution: By definition, $f(x) = 1$ for all x, so:

$$\begin{aligned} f'(x) &= \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{1 - 1}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{0}{\Delta x} \\ &= \lim_{\Delta x \to 0} 0 \\ f'(x) &= 0 \end{aligned}$$

[11] Recall that the *domain* of f is the set of all numbers x such that $f(x)$ is defined.

Notice in the above example that replacing Δx by 0 was unnecessary when taking the limit, since the ratio $\frac{f(x+\Delta x) - f(x)}{\Delta x}$ simplified to 0 *before* taking the limit, and the limit of 0 is 0 regardless of what Δx approaches. In fact, the answer—namely, $f'(x) = 0$ for all x—should have been obvious without any calculations: the function $f(x) = 1$ is a *constant function*, so its value (1) never changes , and thus its rate of change is always 0. Hence, its derivative is 0 everywhere. Replacing the constant 1 by *any* constant yields the following important result:

The derivative of any constant function is 0.

The above discussion shows that the calculation in Example 1.1 was unnecessary. Consider another example where no calculation is required to find the derivative: the function $f(x) = x$. The graph of this function is just the line $y = x$ in the xy-plane, and the rate of change of a line is a constant, called its **slope**. The line $y = x$ has a slope of 1, so the derivative of $f(x) = x$ is $f'(x) = 1$ for all x. The formal calculation of the derivative, though unnecessary, verifies this:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{(x+\Delta x) - x}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta x}{\Delta x} = \lim_{\Delta x \to 0} 1 = 1$$

Recall that a function whose graph is a line is called a **linear function**. For a general linear function $f(x) = mx + b$, where m is the slope of the line and b is its y-intercept, the same argument as above for $f(x) = x$ yields the following result:

The derivative of any linear function is the slope of the line itself:
If $f(x) = mx + b$ then $f'(x) = m$ for all x.

The function $f(x) = 1$ from Example 1.1 is the special case where $m = 0$ and $b = 1$; its graph is a horizontal line, so its slope (and hence its derivative) is 0 for all x. Likewise, the function $f(x) = 2x - 1$ represents a line of slope $m = 2$, so its derivative is 2 for all x. Figure 1.2.1 shows these and other linear functions $y = f(x)$.

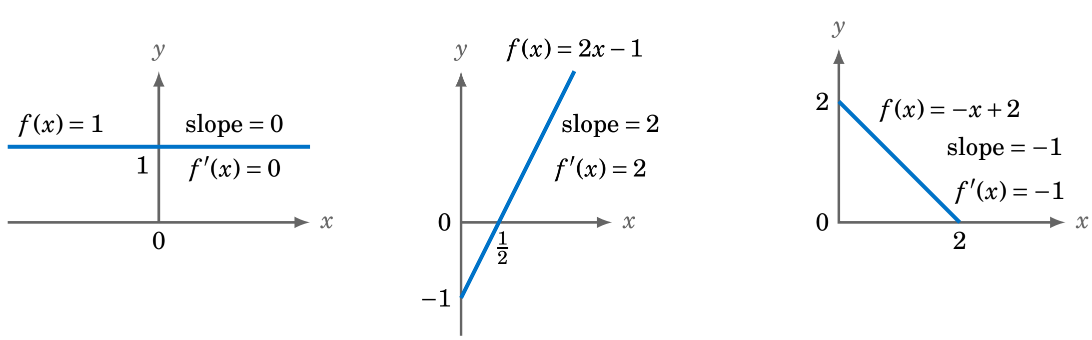


Figure 1.2.1 Slopes and derivatives of lines

Linear functions have a constant derivative—the constant being the slope of the line. The converse turns out to be true: a function with a constant derivative must be a linear function.[12] What types of functions do not have constant derivatives? The previous section discussed such a function: the parabola $s(t) = -16t^2 + 100$, whose derivative $s'(t) = -32t$ is clearly not a constant function. In general, functions that represent curves (i.e. not straight lines) do not change at a constant rate—that is *precisely* what makes them curved. So such functions do not have a constant derivative.

Example 1.2

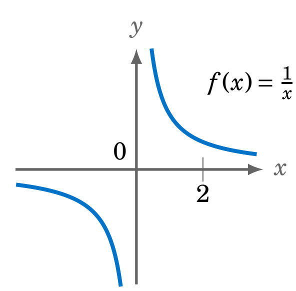
Find the derivative of the function $f(x) = \dfrac{1}{x}$. Also, find the instantaneous rate of change of f at $x = 2$.

Solution: For all $x \neq 0$, the derivative $f'(x)$ is:

$$\begin{aligned}
f'(x) &= \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{\dfrac{1}{x+\Delta x} - \dfrac{1}{x}}{\Delta x} \rightarrow \frac{0}{0} \quad \text{, so simplify the ratio before plugging in } \Delta x = 0, \\
&= \lim_{\Delta x \to 0} \frac{\dfrac{x - (x+\Delta x)}{(x+\Delta x)x}}{\Delta x} \quad \text{(after getting a common denominator)} \\
&= \lim_{\Delta x \to 0} \frac{-\cancel{\Delta x}}{\cancel{\Delta x}(x+\Delta x)x} \\
&= \lim_{\Delta x \to 0} \frac{-1}{(x+\Delta x)x} = \frac{-1}{(x+0)x} \\
f'(x) &= -\frac{1}{x^2}
\end{aligned}$$

The instantaneous rate of change of f at $x = 2$ is just the derivative $f'(x)$ evaluated at $x = 2$, that is, $f'(2) = -\frac{1}{2^2} = -\frac{1}{4}$.

Notice that the instantaneous rate of change $f'(2) = -\frac{1}{4}$ in the above example is a negative number. This should make sense, since the function $f(x) = \frac{1}{x}$ is changing in the negative direction at $x = 2$; that is, $f(x)$ is *decreasing* in value at $x = 2$. This is plain to see from the graph of $f(x) = \frac{1}{x}$ shown on the right. In fact, for all $x \neq 0$ the function $f(x) = \frac{1}{x}$ is decreasing as x grows.[13] This is reflected in the derivative $f'(x) = -\frac{1}{x^2}$ being negative for all $x \neq 0$. In general, a negative derivative means that the function is decreasing, while a positive derivative means that it is increasing.



[12]This will be proved in Chapter 5.

[13]In this text, the rate of change of $f(x)$ is *always* taken in the direction of increasing x, i.e. in the positive x direction.

The problem with using the limit definition to find the derivative of a curved function is that the calculations require more work, as the above example shows. As the functions become more complicated those calculations can become difficult or even impossible. And though limits have not yet been defined formally, for now the intuitively obvious idea of limits suffices, namely:

For a real number a and a real-valued function $f(x)$, say that the *limit* of $f(x)$ as x approaches a equals the number L, written as

$$\lim_{x\to a} f(x) = L\,,$$

if $f(x)$ approaches L as x approaches a.
Equivalently, this means that $f(x)$ can be made as close as you want to L by choosing x close enough to a. Note that x can approach a from any direction.

Below are some simple rules for limits, which will be proved later:

Rules for Limits: Suppose that a is a real number and that $f(x)$ and $g(x)$ are real-valued functions such that $\lim_{x\to a} f(x)$ and $\lim_{x\to a} g(x)$ both exist. Then:

(a) $\displaystyle \lim_{x\to a}\,(f(x)+g(x)) = \left(\lim_{x\to a} f(x)\right) + \left(\lim_{x\to a} g(x)\right)$

(b) $\displaystyle \lim_{x\to a}\,(f(x)-g(x)) = \left(\lim_{x\to a} f(x)\right) - \left(\lim_{x\to a} g(x)\right)$

(c) $\displaystyle \lim_{x\to a}\,(k\cdot f(x)) = k\cdot\left(\lim_{x\to a} f(x)\right)$ for any constant k

(d) $\displaystyle \lim_{x\to a}\,(f(x)\cdot g(x)) = \left(\lim_{x\to a} f(x)\right)\cdot\left(\lim_{x\to a} g(x)\right)$

(e) $\displaystyle \lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}$, if $\displaystyle\lim_{x\to a} g(x) \neq 0$

The above rules say that the limit of sums, differences, constant multiples, products, and quotients is the sum, difference, constant multiple, product, and quotient, respectively, of the limits. This seems intuitively obvious.

These rules can be used for finding other expressions for the derivative. The quantity Δx represents a small number—positive or negative—that approaches 0, but it is common in mathematics texts to use the letter h instead:[14]

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \tag{1.4}$$

[14]Physics texts typically prefer the delta notation, since Δx represents a small change in some physical quantity x.

Another formulation is to set $h = w - x$ in formula (1.4), which yields

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{w-x\to 0} \frac{f(x+(w-x)) - f(x)}{w - x},$$

so that

$$\boxed{f'(x) = \lim_{w\to x} \frac{f(w) - f(x)}{w - x}} \tag{1.5}$$

since $w - x$ approaches 0 if and only if w approaches x. Another formulation replaces h by $-h$:

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{-h\to 0} \frac{f(x+-h) - f(x)}{-h} = \lim_{-h\to 0} \frac{-(f(x) - f(x-h))}{-h},$$

and thus

$$\boxed{f'(x) = \lim_{h\to 0} \frac{f(x) - f(x-h)}{h}} \tag{1.6}$$

since $-h$ approaches 0 if and only if h approaches 0. The above formulations did not use the Limit Rules, but the following result does:

> Suppose that $f'(x)$ exists. Then
>
> $$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x-h)}{2h}. \tag{1.7}$$

Proof: Since $f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{f(x) - f(x-h)}{h}$ by formulas (1.4) and (1.6), then Limit Rule (c) shows that

$$\frac{1}{2}f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{2h} = \lim_{h\to 0} \frac{f(x) - f(x-h)}{2h}.$$

Now use the idea that $a - b = (a - c) + (c - b)$ for all a, b, and c to write:

$$\begin{aligned}
\lim_{h\to 0} \frac{f(x+h) - f(x-h)}{2h} &= \lim_{h\to 0} \frac{(f(x+h) - f(x)) + (f(x) - f(x-h))}{2h} \\
&= \lim_{h\to 0} \frac{f(x+h) - f(x)}{2h} + \lim_{h\to 0} \frac{f(x) - f(x-h)}{2h} \quad \text{(by Limit Rule (a))} \\
&= \frac{1}{2}\cdot f'(x) + \frac{1}{2}\cdot f'(x) \\
&= f'(x)
\end{aligned}$$

QED

As an example of using these different formulations, recall that a function f is **even** if $f(-x) = f(x)$ for all x in the domain of f, and f is **odd** if $f(-x) = -f(x)$ for all x in its domain. For example, x^2, x^4, and $\cos x$ are even functions; x, x^3, and $\sin x$ are odd functions. The following result is often useful:

> The derivative of an even function is an odd function.
> The derivative of an odd function is an even function.

To prove the first statement—the second is an exercise—suppose that f is an even function and that $f'(x)$ exists for all x in its domain. Then

$$\begin{aligned}
f'(-x) &= \lim_{h\to 0} \frac{f(-x+h) - f(-x)}{h} && \text{by formula (1.4) with } x \text{ replaced by } -x \\
&= \lim_{h\to 0} \frac{f(-(x-h)) - f(-x)}{h} \\
&= \lim_{h\to 0} \frac{f(x-h) - f(x)}{h} && \text{since } f \text{ is even} \\
&= \lim_{h\to 0} \frac{-(f(x) - f(x-h))}{h} \\
&= -\lim_{h\to 0} \frac{f(x) - f(x-h)}{h} && \text{by Limit Rule (c), so} \\
f'(-x) &= -f'(x) && \text{by formula (1.6),}
\end{aligned}$$

which shows that f' is an odd function.

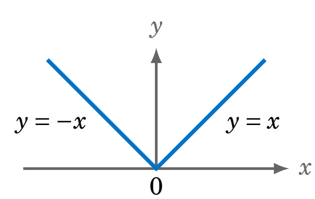
Derivatives do not always exist, as the following example shows.

Example 1.3

Let $f(x) = |x|$. Show that $f'(0)$ does not exist.

Solution: Recall that the *absolute value function* $f(x) = |x|$ is defined as

$$f(x) = |x| = \begin{cases} x & \text{if } x \ge 0 \\ -x & \text{if } x < 0 \end{cases}$$



$f(x) = |x|$

The graph consists of two lines meeting at the origin. For $x \ge 0$ the graph is the line $y = x$, which has slope 1. For $x \le 0$ the graph is the line $y = -x$, which has slope -1. These lines agree in *value* ($y = 0$) at $x = 0$, but their *slopes* do not agree in value at $x = 0$. Therefore the derivative of f does not exist at $x = 0$, since the derivative of a curve is just its slope. A more "formal" proof (which amounts to the same argument) is outlined in the exercises.

If the derivative $f'(x)$ exists then f is **differentiable** at x. A **differentiable function** is one that is differentiable at every point in its domain. For example, $f(x) = x$ is a differentiable function, but $f(x) = |x|$ is not differentiable at $x = 0$. The act of calculating a derivative is called **differentiation**. For example, differentiating the function $f(x) = x$ yields $f'(x) = 1$.

Exercises

Note: For all exercises, you can use anything discussed so far (including previous exercises).

A

For Exercises 1-11, find the derivative of the given function $f(x)$ for all x (unless indicated otherwise).

1. $f(x) = 0$ **2.** $f(x) = 1 - 3x$ **3.** $f(x) = (x+1)^2$ **4.** $f(x) = 2x^2 - 3x + 1$

5. $f(x) = \frac{1}{x+1}$, for all $x \neq -1$ **6.** $f(x) = \frac{-1}{x+1}$, for all $x \neq -1$ **7.** $f(x) = \frac{1}{x^2}$, for all $x \neq 0$

8. $f(x) = \sqrt{x}$, for all $x > 0$ *(Hint: Rationalize the numerator in the definition of the derivative.)*

9. $f(x) = \sqrt{x+1}$, for all $x > -1$ **10.** $f(x) = \sqrt{x^2+1}$ **11.** $f(x) = \sqrt{x^2+3x+4}$

12. In Exercise 8 the point $x = 0$ was excluded when calculating $f'(x)$, even though $x = 0$ is in the domain of $f(x) = \sqrt{x}$. Can you explain why $x = 0$ was excluded?

B

13. Show that for all functions f such that $f'(x)$ exists, $f'(x) = \lim\limits_{w \to x} \dfrac{f(x) - f(w)}{x - w}$.

14. True or false: If f and g are differentiable functions on an interval (a,b) and $f(x) < g(x)$ for all x in (a,b), then $f'(x) < g'(x)$ for all x in (a,b). If true, prove it; if false, give a counterexample. Would your answer change if the restriction of x to (a,b) were removed and all real x were used instead?

15. Show that the derivative of an odd function is an even function.

C

For Exercises 16-21, assuming that $f'(x)$ exists, prove the given formula.

16. $f'(x) = \lim\limits_{h \to 0} \dfrac{f(x+2h) - f(x-2h)}{4h}$ **17.** $f'(x) = \lim\limits_{h \to 0} \dfrac{f(x+3h) - f(x-3h)}{6h}$

18. $f'(x) = \lim\limits_{h \to 0} \dfrac{f(x+2h) - f(x-3h)}{5h}$ **19.** $f'(x) = \lim\limits_{h \to 0} \dfrac{f(x+ah) - f(x-bh)}{(a+b)h}$ $(a, b > 0)$

20. $\lim\limits_{w \to x} \dfrac{w f(x) - x f(w)}{w - x} = f(x) - x f'(x)$ **21.** $\lim\limits_{w \to x} \dfrac{w^2 f(x) - x^2 f(w)}{w - x} = 2x f(x) - x^2 f'(x)$

22. Show that $f(x) = |x|$ is not differentiable at $x = 0$, using formula (1.4) for the derivative. Here you will have to use a part of the definition which has not been used yet: as h approaches 0, h can be either positive or negative. Consider those two cases in showing that the limit is not defined at $x = 0$.

23. Suppose that $f(a+b) = f(a)f(b)$ for all a and b, and $f'(0)$ exists. Show that $f'(x)$ exists for all x.

1.3 The Derivative: Infinitesimal Approach

Traditionally a function f of a variable x is written as $y = f(x)$. The *dependent* variable y is considered a function of the *independent* variable x. This allows taking the derivative of y **with respect to** x, i.e. the derivative of y as a function of x, denoted by $\frac{dy}{dx}$. This is simply a different way of writing $f'(x)$, and is just one of many ways of denoting the derivative:

Notation for the derivative of $y = f(x)$: The following are all equivalent:

$$\frac{dy}{dx}\ ,\quad f'(x)\ ,\quad \frac{d}{dx}(f(x))\ ,\quad y'\ ,\quad \dot{y}\ ,\quad \dot{f}(x)\ ,\quad \frac{df}{dx}\ ,\quad Df(x)$$

The notation $\frac{dy}{dx}$ appears to denote a fraction: a quantity dy divided by a quantity dx. It turns out that the derivative really *can* be thought of in that way, as a ratio of **infinitesimals**. In fact, this was the way in which derivatives were used by the founders of calculus—Newton and, in particular, Leibniz.[15] Even today, this is often the way in which derivatives are thought of and used in fields outside of mathematics, such as physics, engineering, and chemistry, perhaps due to its more intuitive nature.

The concept of infinitesimals used here is based on the *nilsquare infinitesimal* approach developed by J.L. Bell[16], namely:

A number δ is an **infinitesimal** if the conditions (a)-(d) hold:

(a) $\delta \neq 0$

(b) if $\delta > 0$ then δ is smaller than any positive real number

(c) if $\delta < 0$ then δ is larger than any negative real number

(d) $\delta^2 = 0$ (and hence all higher powers of δ, such as δ^3 and δ^4, are also 0)

Note: Any infinitesimal multiplied by a nonzero real number is also an infinitesimal, while 0 times an infinitesimal is 0.

The above definition says that infinitesimals are numbers which are closer to 0 than any positive or negative number without being zero themselves, and raising them to powers greater than or equal to 2 makes them 0. So infinitesimals are not real numbers.[17] This is not a problem, since calculus deals with other numbers, such as infinity, which are not real. An infinitesimal can be thought of as an infinitely small number arbitrarily close to 0 but not 0.

[15]It was Leibniz who created the notation $\frac{dy}{dx}$. For this reason $\frac{dy}{dx}$ is called the *Leibniz notation* for the derivative. Newton used the *dot notation* $\dot{y}$, which has fallen out of favor with mathematicians but is still used by many physicists, especially when the independent variable represents time. Newton called derivatives *fluxions*. The *prime notation* f' is due to the French mathematician and physicist Joseph Louis Lagrange (1736-1813).

[16]BELL, J.L., *A Primer of Infinitesimal Analysis*, Cambridge, U.K.: Cambridge University Press, 1998.

[17]In an equivalent treatment, infinitesimals are part of the *hyperreal number system*. See KEISLER, H.J., *Elementary Calculus: An Infinitesimal Approach*, Boston: Prindle, Weber & Schmidt, 1976.

This might seem like a strange notion, but it really is not all that different from the limit notion where, say, you let Δx *approach* 0 but not necessarily let it *equal* 0.[18]

As for the square of a nonzero infinitesimal being 0, think of how a calculator handles the squares of small numbers. For example, most calculators can display 10^{-8} as 0.00000001, and will even let you add 1 to that to get 1.00000001. But when you square 10^{-8} and add 1 to it, most calculators will display the sum as simply 1. The calculator treats the square of 10^{-8}, namely 10^{-16}, as a number so small compared to 1 that it is effectively zero.[19]

Notice a major difference between 0 and an infinitesimal δ: $2\cdot 0$ and 0 are the same, but 2δ and δ are distinct. This holds for any nonzero constant multiple, not just the number 2.

The derivative $\frac{dy}{dx}$ of a function $y = f(x)$ can now be defined in terms of infinitesimals:

> Let dx be an infinitesimal such that $f(x+dx)$ is defined. Then $dy = f(x+dx) - f(x)$ is also an infinitesimal, and the derivative of $y = f(x)$ at x is the ratio of dy to dx:
>
> $$\frac{dy}{dx} = \frac{f(x+dx) - f(x)}{dx} \tag{1.8}$$

The basic idea is that dx is an infinitesimally small change in the variable x, producing an infinitesimally small change dy in the value of $y = f(x)$.

Example 1.4

Show that the derivative of $y = f(x) = x^2$ is $\frac{dy}{dx} = 2x$.

Solution: For any real number x,

$$\begin{aligned}
\frac{dy}{dx} &= \frac{f(x+dx) - f(x)}{dx} \\
&= \frac{(x+dx)^2 - x^2}{dx} \\
&= \frac{\cancel{x^2} + 2x\,dx + (dx)^2 - \cancel{x^2}}{dx} \\
&= \frac{2x\,dx + 0}{dx} \qquad \text{since } dx \text{ is an infinitesimal} \Rightarrow (dx)^2 = 0 \\
&= \frac{2x\,\cancel{dx}}{\cancel{dx}} \\
&= 2x
\end{aligned}$$

[18]The infinitesimal approach was first developed in an axiomatic manner in the landmark book ROBINSON, A., *Non-Standard Analysis*, Amsterdam: North-Holland, 1966. Robinson showed that for all practical purposes calculus can be developed without resorting to limits, with equivalent results.

[19]Calculators do this for display reasons—most can show only 10-12 digits. Try this experiment on your calculator: Add 10^{30}, $-(10^{30})$, and 1 in two different ways: $(10^{30} + -(10^{30})) + 1$, and $10^{30} + (-(10^{30}) + 1)$. The first way will give you the correct answer 1, but the second way yields 0. So addition is not always associative on calculators!

You might have noticed that the above example did not involve limits, and that the derivative $2x$ represents a real number (i.e. no infinitesimals appear in the final answer); this will always be the case. Infinitesimals possess another useful property:

Microstraightness Property: For the graph of a differentiable function, any part of the curve with infinitesimal length is a straight line segment.

In other words, *at the infinitesimal level differentiable curves are straight*. The idea behind this is simple. At various points on a nonstraight differentiable curve $y = f(x)$ the distances along the curve between the points are not quite the same as the lengths of the line segments joining the points. For example, in Figure 1.3.1 the distance s measured along the curve from the point A to the point B is not the same as the length of the line segment $\overline{AB}$ joining A to B.

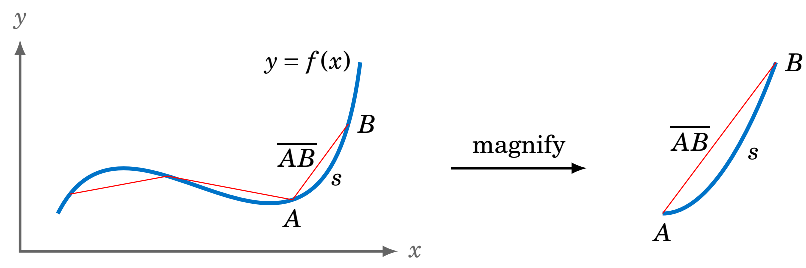
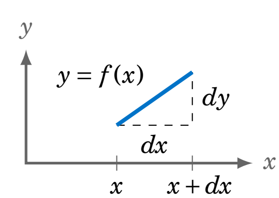


Figure 1.3.1 Curved *real-valued* distance $s \neq$ length of $\overline{AB}$, but *infinitesimal* $s =$ length of $\overline{AB}$

However, as the points A and B get closer to each other, the difference between that part of the curve joining A to B and the line segment $\overline{AB}$ becomes less noticeable. That is, the curve is *almost* linear when A and B are close. The Microstraightness Property simply goes one step further and says that the curve actually *is* linear when the distance s between the points is infinitesimal (so that s equals the length of $\overline{AB}$ at the infinitesimal level).

At first this might seem nonsensical. After all, how could any nonstraight part of a curve be straight? You have to remember that an infinitesimal is an abstraction—it does not exist physically. A curve $y = f(x)$ is also an abstraction, which exists in a purely mathematical sense, so its geometric properties at the "normal" scale do not have to match those at the infinitesimal scale (which can be defined in any way, provided the properties at that scale are consistent).

This abstraction finally reveals what an instantaneous rate of change is: the average rate of change over an infinitesimal interval. Moving an infinitesimal amount dx away from a value x produces an infinitesimal change dy in a differentiable function $y = f(x)$. The average rate of change of $y = f(x)$ over the infinitesimal interval $[x, x+dx]$ is thus $\frac{dy}{dx}$, i.e. the slope—rise over run—of the straight line segment represented by the curve $y = f(x)$ over that interval, as in the figure on the right.[20]



[20]Notice that the figure implies that the Pythagorean Theorem does not apply to infinitesimal triangles. This will be discussed in Chapter 8.

The Microstraightness Property can be extended to **smooth** curves—that is, curves without sharp edges or cusps. For example, circles and ellipses are smooth, but polygons are not.

The properties of infinitesimals can be applied to determine the derivatives of the sine and cosine functions. Consider a circle of radius 1 with center O and points A and B on the circle such that the line segment $\overline{AB}$ is a diameter. Let C be a point on the circle such that the angle $\angle BAC$ has an infinitesimal measure dx (in radians) as in Figure 1.3.2(a).

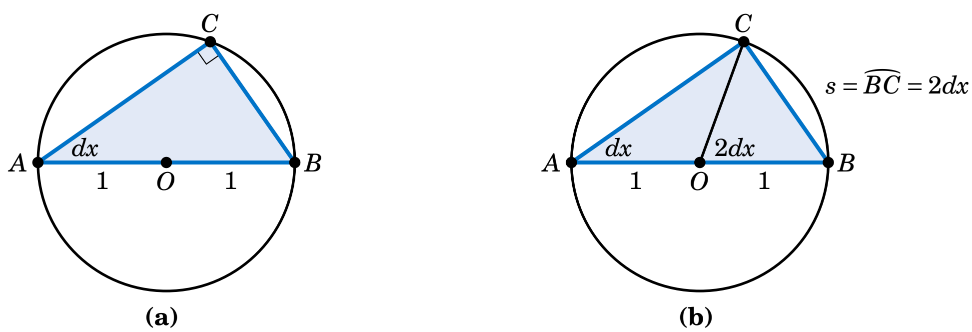


Figure 1.3.2 Circle O: $BC = 2\sin dx$, $\angle BOC = 2\angle BAC$

By *Thales' Theorem* from elementary geometry, the angle $\angle ACB$ is a right angle. Thus:

$$\sin dx \;=\; \frac{BC}{AB} \;=\; \frac{BC}{2} \quad\Rightarrow\quad BC \;=\; 2\sin dx$$

Figure 1.3.2(b) shows that $\angle OAC + \angle OCA + \angle AOC = \pi$. Thus, $1 = OC = OA \Rightarrow \angle OCA = \angle OAC = dx \Rightarrow \angle AOC = \pi - dx - dx = \pi - 2dx \Rightarrow \angle BOC = 2dx$. By the arc length formula from trigonometry, the length s of the *arc* $\widehat{BC}$ along the circle from B to C is the radius times the central angle $\angle BOC$: $s = \widehat{BC} = 1 \cdot 2dx = 2dx$. But by Microstraightness, $\widehat{BC} = BC$, and thus:

$$2\sin dx \;=\; BC \;=\; \widehat{BC} \;=\; 2dx \quad\Rightarrow\quad \boxed{\sin dx \;=\; dx}$$

Since dx is an infinitesimal, $(dx)^2 = 0$. So since $\sin^2 dx + \cos^2 dx = 1$, then:

$$\cos^2 dx \;=\; 1 \;-\; \sin^2 dx \;=\; 1 \;-\; (dx)^2 \;=\; 1 \;-\; 0 \;=\; 1 \quad\Rightarrow\quad \boxed{\cos dx \;=\; 1}$$

The derivative of $y = \sin x$ is then:

$$\begin{aligned}\frac{d}{dx}(\sin x) \;&=\; \frac{dy}{dx} \;=\; \frac{\sin(x+dx) \;-\; \sin x}{dx}\\ &=\; \frac{(\sin x \cos dx \;+\; \sin dx \cos x) \;-\; \sin x}{dx} \quad \text{by the sine addition formula}\\ &=\; \frac{\cancel{(\sin x)(1)} \;+\; dx \cos x \;-\; \cancel{\sin x}}{dx} \;=\; \frac{\cancel{dx} \cos x}{\cancel{dx}} \quad \text{, and thus:}\end{aligned}$$

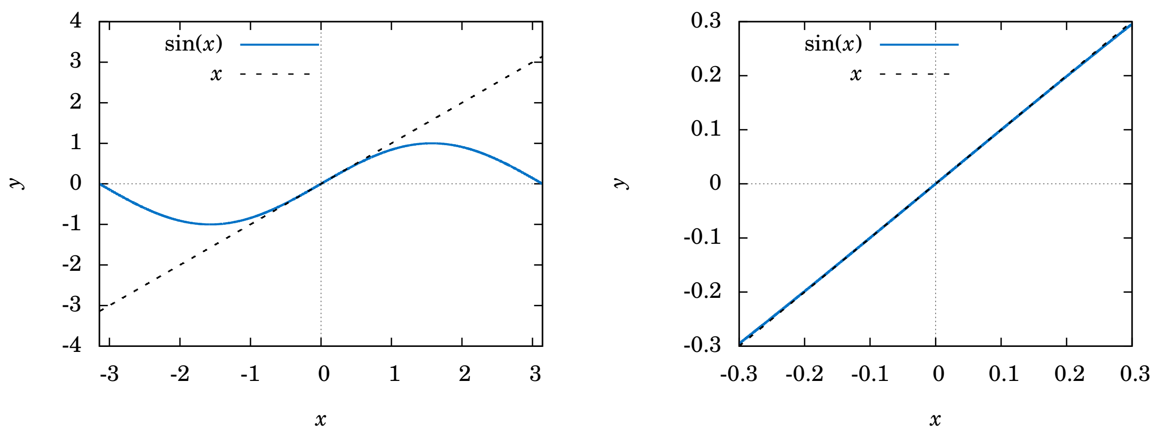
$$\boxed{\frac{d}{dx}(\sin x) \;=\; \cos x}$$

A similar argument (left as an exercise) using the cosine addition formula shows:

$$\boxed{\frac{d}{dx}(\cos x) \;=\; -\sin x}$$

One of the intermediate results proved here bears closer examination. Namely, $\sin dx = dx$ for an infinitesimal angle dx measured in radians. At first, it might seem that this cannot be true. After all, an infinitesimal dx is thought of as being infinitely close to 0, and $\sin 0 = 0$, so you might expect that $\sin dx = 0$. But this is not the case. The formula $\sin dx = dx$ says that in an infinitesimal interval around 0, the function $y = \sin x$ is identical to the line $y = x$ (not the line $y = 0$). This, in turn, suggests that for real-valued x close to 0, $\sin x \approx x$.

This indeed turns out to be the case. The free graphing software Gnuplot[21] can display the graphs of $y = \sin x$ and $y = x$. Figure 1.3.3(a) below shows how those graphs compare over the interval $[-\pi, \pi]$. Outside the interval $[-1, 1]$ there is a noticeable difference.



(a) **(b)**

Figure 1.3.3 $\sin dx = dx$: Comparing $y = \sin x$ and $y = x$ near $x = 0$

Figure 1.3.3(b) shows that there is virtually no difference in the graphs even over the non-infinitesimal interval $[-0.3, 0.3]$. So $\sin x \approx x$ is a good approximation when x is close to 0, that is, when $|x| \ll 1$ (the symbol $\ll$ means "much less than"). This approximation is used in many applications in engineering and physics when the angle x is assumed to be small.

Notice something else suggested by the relation $\sin dx = dx$: there is a fundamental difference at the infinitesimal level between a line of slope 1 ($y = x$) and a line of slope 0 ($y = 0$). In a *real* interval $(-a, a)$ around $x = 0$ the difference between the two lines can be made as small as desired by choosing $a > 0$ small enough. But in an infinitesimal interval $(-\delta, \delta)$ around $x = 0$ there is unbridgeable gulf between the two lines. This is the crucial difference in $\sin dx$ being equal to dx rather than 0.

[21]Available at `http://www.gnuplot.info`.

Notice also that the value of a function at an infinitesimal may itself be an infinitesimal (e.g. $\sin dx = dx$) or a real number (e.g. $\cos dx = 1$).

For a differentiable function $f(x)$, $\frac{df}{dx} = f'(x)$ and so multiplying both sides by dx yields the important relation:

$$df = f'(x)\,dx \tag{1.9}$$

Note that both sides of the above equation are infinitesimals for each value of x in the domain of f', since $f'(x)$ would then be a real number.

The notion of an infinitesimal was fairly radical at the time (and still is). Some mathematicians embraced it, e.g. the outstanding Swiss mathematician Leonhard Euler (1707-1783), who produced a large amount of work using infinitesimals. But it was *too* radical for many mathematicians (and philosophers[22]), enough so that by the 19th century some mathematicians (notably Augustin Cauchy and Karl Weierstrass) felt the need to put calculus on what they considered a more "rigorous" footing, based on limits.[23] Yet it was precisely the notion of an infinitesimal which lent calculus its *modern* character, by showing the power and usefulness of such an abstraction (especially one that did not obey the rules of classical mathematics).

Exercises

A

For Exercises 1-9, let dx be an infinitesimal and prove the given formula.

1. $(dx+1)^2 = 2dx+1$

2. $(dx+1)^3 = 3dx+1$

3. $(dx+1)^{-1} = 1-dx$

4. $\tan dx = dx$

5. $\sin 2dx = 2dx$

6. $\cos 2dx = 1$

7. $\sin 3dx = 3dx$

8. $\cos 3dx = 1$

9. $\sin 4dx = 4dx$

10. Is $\cot dx$ defined for an infinitesimal dx? If so, then find its value. If not, then explain why.

11. In the proof of the derivative formulas for $\sin x$ and $\cos x$, the equation $\cos^2 dx = 1$ was solved to give $\cos dx = 1$. Why was the other possible solution $\cos dx = -1$ ignored?

B

12. Show that $\frac{d}{dx}(\cos x) = -\sin x$.

13. Show that $\frac{d}{dx}(\cos 2x) = -2\sin 2x$. *(Hint: Use Exercises 5 and 6.)*

C

14. Show that $\frac{d}{dx}(\tan x) = \sec^2 x$. *(Hint: Use Exercise 4.)*

[22]The English philosopher George Berkeley (1685-1753) famously derided infinitesimals as "the ghosts of departed quantities" in his book *The Analyst* (1734), which had the disquieting subtitle "A Discourse Addressed to an Infidel Mathematician" (directed at Newton).

[23]However, the limit approach turns out, ultimately, to be equivalent to the infinitesimal approach. In essence, only the terminology is different.

1.4 Derivatives of Sums, Products and Quotients

So far the derivatives of only a few simple functions have been calculated. The following rules will make it easier to calculate derivatives of more functions:

Rules for Derivatives: Suppose that f and g are differentiable functions of x. Then:

Sum Rule: $\dfrac{d}{dx}(f+g) = \dfrac{df}{dx} + \dfrac{dg}{dx}$

Difference Rule: $\dfrac{d}{dx}(f-g) = \dfrac{df}{dx} - \dfrac{dg}{dx}$

Constant Multiple Rule: $\dfrac{d}{dx}(cf) = c \cdot \dfrac{df}{dx}$ for any constant c

Product Rule: $\dfrac{d}{dx}(f \cdot g) = f \cdot \dfrac{dg}{dx} + g \cdot \dfrac{df}{dx}$

Quotient Rule: $\dfrac{d}{dx}\left(\dfrac{f}{g}\right) = \dfrac{g \cdot \frac{df}{dx} - f \cdot \frac{dg}{dx}}{g^2}$

The above rules can be written using the prime notation for derivatives:

Sum Rule: $(f+g)'(x) = f'(x) + g'(x)$

Difference Rule: $(f-g)'(x) = f'(x) - g'(x)$

Constant Multiple Rule: $(cf)'(x) = c \cdot f'(x)$ for any constant c

Product Rule: $(f \cdot g)'(x) = f(x) \cdot g'(x) + g(x) \cdot f'(x)$

Quotient Rule: $\left(\dfrac{f}{g}\right)'(x) = \dfrac{g(x) \cdot f'(x) - f(x) \cdot g'(x)}{(g(x))^2}$

The proof of the Sum Rule is straightforward. Since $\frac{df}{dx}$ and $\frac{dg}{dx}$ both exist then:

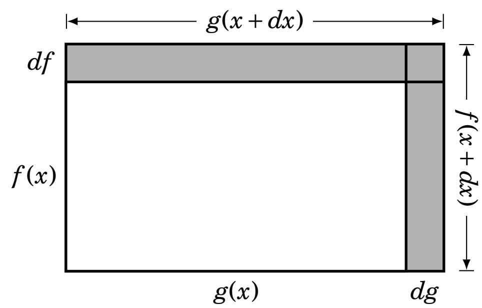
$$\begin{aligned}
\frac{d}{dx}(f+g) &= \frac{(f+g)(x+dx) - (f+g)(x)}{dx} = \frac{f(x+dx) + g(x+dx) - (f(x) + g(x))}{dx} \\
&= \frac{f(x+dx) - f(x) + g(x+dx) - g(x)}{dx} = \frac{f(x+dx) - f(x)}{dx} + \frac{g(x+dx) - g(x)}{dx} \\
&= \frac{df}{dx} + \frac{dg}{dx} \qquad \checkmark
\end{aligned}$$

The proofs of the Difference and Constant Multiple Rules are similar and are left as exercises.

Note that by the Product Rule, in general the derivative of a product is **not** the product of the derivatives. That is, $\frac{d(f\cdot g)}{dx} \neq \frac{df}{dx}\cdot\frac{dg}{dx}$. This should be obvious from some earlier examples. For instance, if $f(x)=x$ and $g(x)=1$ then $(f\cdot g)(x) = x\cdot 1 = x$ so that $\frac{d(f\cdot g)}{dx} = 1$, but $\frac{df}{dx}\cdot\frac{dg}{dx} = 1\cdot 0 = 0$.

There is a proof of the Product Rule similar to the proof of the Sum Rule (see Exercise 20), but there is a more geometric way of seeing why the formula holds, described below.

Construct a rectangle whose perpendicular sides have lengths $f(x)$ and $g(x)$ for some x, as in the drawing on the right. Change x by some infinitesimal amount dx, which produces infinitesimal changes df and dg in $f(x)$ and $g(x)$, respectively. Assume those changes are positive and extend the original rectangle by those amounts, creating a larger rectangle with perpendicular sides $f(x+dx)$ and $g(x+dx)$. Then



$$\begin{aligned}
d(f\cdot g) &= (f\cdot g)(x+dx) - (f\cdot g)(x)\\
&= f(x+dx)\cdot g(x+dx) - f(x)\cdot g(x)\\
&= (\text{area of outer rectangle}) - (\text{area of original rectangle})\\
&= \text{sum of the areas of the three shaded inner rectangles}\\
&= f(x)\cdot dg + g(x)\cdot df + df\cdot dg\\
&= f(x)\cdot dg + g(x)\cdot df + (f'(x)\,dx)\cdot(g'(x)\,dx)\\
&= f(x)\cdot dg + g(x)\cdot df + (f'(x)g'(x))\cdot(dx)^2\\
&= f(x)\cdot dg + g(x)\cdot df + (f'(x)g'(x))\cdot 0\\
d(f\cdot g) &= f(x)\cdot dg + g(x)\cdot df \quad \text{, so dividing both sides by } dx \text{ yields}
\end{aligned}$$

$$\frac{d(f\cdot g)}{dx} = f(x)\cdot\frac{dg}{dx} + g(x)\cdot\frac{df}{dx} \quad \checkmark$$

To prove the Quotient Rule, let $y = \frac{f}{g}$, so $f = g\cdot y$. If y were a differentiable function of x, then the Product Rule would give

$$\frac{df}{dx} = \frac{d(g\cdot y)}{dx} = g\cdot\frac{dy}{dx} + y\cdot\frac{dg}{dx} = g\cdot\frac{dy}{dx} + \frac{f}{g}\cdot\frac{dg}{dx} \quad\Rightarrow\quad g\cdot\frac{dy}{dx} = \frac{df}{dx} - \frac{f}{g}\cdot\frac{dg}{dx}$$

and so dividing both sides by g and getting a common denominator gives

$$\frac{dy}{dx} = \frac{1}{g}\cdot\frac{df}{dx} - \frac{f}{g^2}\cdot\frac{dg}{dx} = \frac{g\cdot\frac{df}{dx} - f\cdot\frac{dg}{dx}}{g^2} \quad \checkmark$$

A simple mnemonic device for remembering the Quotient Rule is: write $\frac{f}{g}$ as $\frac{\text{HI}}{\text{HO}}$—so that HI represents the "high" (numerator) part of the quotient and HO represents the "low" (denominator) part—and think of dHI and dHO as the derivatives of HI and HO, respectively. Then $\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{\text{HO}\cdot d\text{HI} - \text{HI}\cdot d\text{HO}}{\text{HO}^2}$, pronounced as "ho-dee-hi minus hi-dee-ho over ho-ho."

Example 1.5

Use the Quotient Rule to show that $\frac{d}{dx}(\tan x) = \sec^2 x$.

Solution: Since $\tan x = \frac{\sin x}{\cos x}$ then:

$$\frac{d}{dx}(\tan x) = \frac{d}{dx}\left(\frac{\sin x}{\cos x}\right) = \frac{(\cos x)\cdot\frac{d}{dx}(\sin x) - (\sin x)\cdot\frac{d}{dx}(\cos x)}{\cos^2 x} = \frac{(\cos x)\cdot(\cos x) - (\sin x)\cdot(-\sin x)}{\cos^2 x}$$

$$= \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x$$

Example 1.6

Use the Quotient Rule to show that $\frac{d}{dx}(\sec x) = \sec x \tan x$.

Solution: Since $\sec x = \frac{1}{\cos x}$ then:

$$\frac{d}{dx}(\sec x) = \frac{d}{dx}\left(\frac{1}{\cos x}\right) = \frac{(\cos x)\cdot\frac{d}{dx}(1) - 1\cdot\frac{d}{dx}(\cos x)}{\cos^2 x} = \frac{(\cos x)\cdot 0 - 1\cdot(-\sin x)}{\cos^2 x}$$

$$= \frac{\sin x}{\cos^2 x} = \frac{1}{\cos x}\cdot\frac{\sin x}{\cos x} = \sec x \tan x$$

Similar to the above examples, the derivatives of $\cot x$ and $\csc x$ can be found using the Quotient Rule (left as exercises). The derivatives of all six trigonometric functions are:

$$\frac{d}{dx}(\cos x) = -\sin x \qquad \frac{d}{dx}(\sec x) = \sec x \tan x$$

$$\frac{d}{dx}(\sin x) = \cos x \qquad \frac{d}{dx}(\csc x) = -\csc x \cot x$$

$$\frac{d}{dx}(\tan x) = \sec^2 x \qquad \frac{d}{dx}(\cot x) = -\csc^2 x$$

Note that the Sum and Difference Rules can be applied to sums and differences, respectively, of not just two functions but any finite (integer) number of functions. For example, for three differentiable functions f_1, f_2, and f_3,

$$\frac{d}{dx}(f_1+f_2+f_3) = \frac{df_1}{dx} + \frac{d}{dx}(f_2+f_3) \quad \text{by the Sum Rule}$$

$$= \frac{df_1}{dx} + \frac{df_2}{dx} + \frac{df_3}{dx} \quad \text{by the Sum Rule again.}$$

Continuing like this for four functions, then five functions, and so forth, the Sum and Difference Rules combined with the Constant Multiple Rule yield the following formula:

For $n \geq 1$ differentiable functions $f_1, \ldots, f_n$ and constants $c_1, \ldots, c_n$:

$$\frac{d}{dx}\left(c_1 f_1 + \cdots + c_n f_n\right) = c_1 \frac{df_1}{dx} + \cdots + c_n \frac{df_n}{dx} \tag{1.10}$$

Note that the above formula includes differences, by using negative constants. The formula also shows that differentiation is a *linear operation*, which makes $\frac{d}{dx}$ a *linear operator*. The idea is that $\frac{d}{dx}$ "operates" on differentiable functions by taking their derivatives with respect to the variable x. The sum $c_1 f_1 + \cdots + c_n f_n$ is called a *linear combination* of functions, and the derivative of that linear combination can be taken term by term, with the constant multiples taken outside the derivatives.

A special case of the above formula is replacing the functions $f_1, \ldots, f_n$ by nonnegative powers of x, making the sum a polynomial in x. In previous sections the derivatives of a few polynomials—such as x and x^2—were calculated. For the derivative of a general polynomial, the following rule is needed:

Power Rule: $\dfrac{d}{dx}\left(x^n\right) = n\,x^{n-1}$ for any integer n

There are several ways to prove this formula; one such way being a **proof by induction**, which in general means using the following principle:[24]

Principle of Mathematical Induction
A statement $P(n)$ about integers $n \geq k$ is true for all $n \geq k$ if:

1. $P(k)$ is true.
2. If $P(n)$ is true for some integer $n \geq k$ then $P(n+1)$ is true.

The idea behind mathematical induction is simple: if a statement is true about some initial integer k (Step 1 above) and if the statement being true for *some* integer implies it is true for the *next* integer (Step 2 above), then the statement being true for k implies it is true for $k+1$, which in turn implies it is true for $k+2$, which implies it is true for $k+3$, and so forth, making it true for *all* integers $n \geq k$.

Typically the initial integer k will be 0 or 1. To prove the Power Rule for all integers, first use induction to prove the rule for all nonnegative integers $n \geq 0$, using $k = 0$ for the initial integer. For the proof by induction, let $P(n)$ be the statement: $\frac{d}{dx}(x^n) = n\,x^{n-1}$.

1. Show that $P(0)$ is true.
   That means showing that the Power Rule holds for $n = 0$, i.e. $\frac{d}{dx}\left(x^0\right) = 0\,x^{0-1} = 0$. But $x^0 = 1$ is a constant, so its derivative is 0. ✓

[24]As Bertrand Russell noted, the name is really a misnomer: it is actually a *definition* of the natural numbers rather than a principle, and induction technically has a different meaning.

2. Assuming $P(n)$ is true for some $n \ge 0$, show that $P(n+1)$ is true.
   Assuming that $\frac{d}{dx}(x^n) = n x^{n-1}$, show that $\frac{d}{dx}\left(x^{n+1}\right) = (n+1)x^{(n+1)-1} = (n+1)x^n$. It was shown in Section 1.2 that $\frac{d}{dx}(x) = 1$, so:

$$\begin{aligned}
\frac{d}{dx}\left(x^{n+1}\right) &= \frac{d}{dx}\left(x \cdot x^n\right) \\
&= x \cdot \frac{d}{dx}\left(x^n\right) + x^n \cdot \frac{d}{dx}(x) \quad \text{(by the Product Rule)} \\
&= x \cdot n x^{n-1} + x^n \cdot 1 \quad \text{(by the assumption that } P(n) \text{ is true)} \\
&= n x^n + x^n = (n+1)x^n \quad \checkmark
\end{aligned}$$

Thus, by induction, the Power Rule is true for all nonnegative integers $n \ge 0$.

To show that the Power Rule is true for all negative integers $n < 0$, write $n = -m$, where m is positive (namely, $m = |n|$). Then:

$$\begin{aligned}
\frac{d}{dx}\left(x^n\right) &= \frac{d}{dx}\left(x^{-m}\right) = \frac{d}{dx}\left(\frac{1}{x^m}\right) = \frac{x^m \cdot \frac{d}{dx}(1) - 1 \cdot \frac{d}{dx}(x^m)}{(x^m)^2} \quad \text{(by the Quotient Rule)} \\
&= \frac{x^m \cdot 0 - 1 \cdot m x^{m-1}}{x^{2m}} \quad \text{(by the Power Rule for positive integers)} \\
&= -m x^{m-1-2m} = -m x^{-m-1} = n x^{n-1} \quad \checkmark
\end{aligned}$$

Thus, the Power Rule is true for all integers, which completes the proof. **QED**

Example 1.7

Find the derivative of $f(x) = x^4 - 4x^3 + 6x^2 - 4x + 1$.

Solution: Differentiate the polynomial term by term and use the Power Rule:

$$\begin{aligned}
\frac{df}{dx} &= \frac{d}{dx}\left(x^4 - 4x^3 + 6x^2 - 4x + 1\right) \\
&= \frac{d}{dx}\left(x^4\right) - 4 \cdot \frac{d}{dx}\left(x^3\right) + 6 \cdot \frac{d}{dx}\left(x^2\right) - 4 \cdot \frac{d}{dx}(x) + \frac{d}{dx}(1) \\
&= 4x^{4-1} - 4 \cdot 3x^{3-1} + 6 \cdot 2x^{2-1} - 4 \cdot 1 + 0 \\
&= 4x^3 - 12x^2 + 12x - 4
\end{aligned}$$

In general, the derivative of a polynomial of degree $n \ge 0$ is given by:

For any constants $a_0, \ldots, a_n$ with $n \ge 0$:

$$\frac{d}{dx}\left(a_n x^n + a_{n-1}x^{n-1} \cdots + a_2 x^2 + a_1 x + a_0\right) = n a_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + 2a_2 x + a_1$$

A way to remember the Power Rule is: bring the exponent down in front of the variable then reduce the variable's original exponent by 1. This works even for negative exponents.

Example 1.8

Find the derivative of $f(t) = 3t^{100} - \frac{2}{t^{100}}$.

Solution: Differentiate term by term:

$$\frac{df}{dt} = \frac{d}{dt}\left(3t^{100} - \frac{2}{t^{100}}\right) = \frac{d}{dt}\left(3t^{100} - 2t^{-100}\right) = 3\cdot 100t^{99} - 2\cdot\left(-100t^{-101}\right) = 300t^{99} + \frac{200}{t^{101}}$$

Exercises

A

For Exercises 1-14, use the rules from this section to find the derivative of the given function.

1. $f(x) = x^2 - x - 1$

2. $f(x) = x^8 + 2x^4 + 1$

3. $f(x) = \dfrac{2x^6}{3} - \dfrac{3}{2x^6}$

4. $f(x) = \dfrac{\sin x + \cos x}{4}$

5. $f(x) = x \sin x$

6. $f(x) = x^2 \tan x$

7. $f(x) = \dfrac{\sin x}{x}$

8. $f(x) = \dfrac{\sin x}{x^2}$

9. $f(t) = \dfrac{2t}{1+t^2}$

10. $g(t) = \dfrac{1-t^2}{1+t^2}$

11. $f(x) = \dfrac{ax+b}{cx+d}$ (a, b, c, d are constants)

12. $F(r) = -\dfrac{Gm_1m_2}{r^2}$ (G, m_1, m_2 are constants)

13. $A(r) = \pi r^2$

14. $V(r) = \dfrac{4}{3}\pi r^3$

15. Show that $\frac{d}{dx}(\cot x) = -\csc^2 x$.

16. Show that $\frac{d}{dx}(\csc x) = -\csc x \cot x$.

B

17. Prove the Difference Rule.

18. Prove the Constant Multiple Rule.

19. Use the Product Rule to show that for three differentiable functions f, g, and h, the derivative of their product is $(fgh)' = f'gh + fg'h + fgh'$.

C

20. Provide an alternative proof of the Product Rule for two differentiable functions f and g of x:

(a) Show that $(df)(dg) = 0$.

(b) By definition, the derivative of the product $f \cdot g$ is

$$\frac{d}{dx}(f \cdot g) = \frac{f(x+dx)\cdot g(x+dx) - f(x)\cdot g(x)}{dx}.$$

Use that formula along with part (a) to show that $\frac{d}{dx}(f \cdot g) = f \cdot \frac{dg}{dx} + g \cdot \frac{df}{dx}$. *(Hint: Recall that $df = f(x+dx) - f(x)$.)*

1.5 The Chain Rule

From what has been discussed so far it might be tempting to think that the derivative of a function like $\sin 2x$ is simply $\cos 2x$, since the derivative of $\sin x$ is $\cos x$. It turns out that is not correct:

$$\begin{aligned}
\frac{d}{dx}(\sin 2x) &= \frac{d}{dx}(2\sin x\cos x) \quad \text{(by the double-angle formula for sine)}\\
&= 2\,\frac{d}{dx}(\sin x\cos x) \quad \text{(by the Constant Multiple Rule)}\\
&= 2\left(\sin x\cdot\frac{d}{dx}(\cos x) + \cos x\cdot\frac{d}{dx}(\sin x)\right) \quad \text{(by the Product Rule)}\\
&= 2\,(\sin x\cdot(-\sin x) + \cos x\cdot\cos x)\\
&= 2\,(\cos^2 x - \sin^2 x)\\
&= 2\cos 2x \quad \text{(by the double-angle formula for cosine)}
\end{aligned}$$

So the derivative of $\sin 2x$ is $2\cos 2x$, *not* $\cos 2x$.

In other words, you cannot simply replace x by $2x$ in the derivative formula for $\sin x$. Instead, regard $\sin 2x$ as a *composition* of two functions: the sine function and the $2x$ function. That is, let $f(u) = \sin u$, where the variable u itself represents a function of another variable x, namely $u(x) = 2x$. So since f is a function of u, and u is a function of x, then f is a function of x, namely: $f(x) = \sin 2x$. Since f is a differentiable function of u, and u is a differentiable function of x, then $\frac{df}{du}$ and $\frac{du}{dx}$ both exist (with $\frac{df}{du} = \cos u$ and $\frac{du}{dx} = 2$), and multiplying the derivatives shows that f is a differentiable function of x:

$$\frac{df}{\cancel{du}}\cdot\frac{\cancel{du}}{dx} = \frac{df}{dx} \quad \text{since the infinitesimals } du \text{ cancel, so}$$

$$(\cos u)\cdot 2 = \frac{df}{dx} \quad\Rightarrow\quad \frac{df}{dx} = 2\cos u = 2\cos 2x$$

The above argument holds in general, and is known as the *Chain Rule*:

Chain Rule: If f is a differentiable function of u, and u is a differentiable function of x, then f is a differentiable function of x, and its derivative with respect to x is:

$$\frac{df}{dx} = \frac{df}{du}\cdot\frac{du}{dx}$$

Notice how simple the proof is—the infinitesimals du cancel.[25]

[25]Some textbooks give dire warnings to *not* think that du is an actual quantity that can be canceled. However, you can safely ignore those warnings, because du is just an infinitesimal and hence *can* be canceled!

The Chain Rule should make sense intuitively. For example, if $\frac{df}{du} = 4$ then that means f is increasing 4 times as fast as u, and if $\frac{du}{dx} = 3$ then u is increasing 3 times as fast as x, so overall f should be increasing $12 = 4 \cdot 3$ times as fast as x, exactly as the Chain Rule says.

Example 1.9

Find the derivative of $f(x) = \sin(x^2 + x + 1)$.

Solution: The idea is to make a *substitution* $u = x^2 + x + 1$ so that $f(x) = \sin u$. By the Chain Rule,

$$\begin{aligned} \frac{df}{dx} &= \frac{df}{du} \cdot \frac{du}{dx} \\ &= \frac{d}{du}(\sin u) \cdot \frac{d}{dx}(x^2 + x + 1) \\ &= (\cos u) \cdot (2x + 1) \\ &= (2x + 1)\cos(x^2 + x + 1) \end{aligned}$$

after replacing u by its definition as a function of x in the last step; the final answer for the derivative should be in terms of x, not u.

In the Chain Rule you can think of the function in question as the composition of an "outer" function f and an "inner" function u; first take the derivative of the "outer" function then multiply by the derivative of the "inner" function. Think of the "inner" function as a box into which you can put any function of x, and the "outer" function being a function of that empty box.

For instance, for the function $f(x) = \sin(x^2 + x + 1)$ in the previous example, think of the "outer" function as $\sin \square$, where $\square = x^2 + x + 1$ is the "inner" function, so that

$$\begin{aligned} f(x) &= \sin(x^2 + x + 1) \\ &= \sin \square \\ \frac{df}{dx} &= (\cos \square) \cdot \frac{d}{dx} \square \\ &= \left(\cos \boxed{x^2 + x + 1}\right) \cdot \frac{d}{dx} \boxed{x^2 + x + 1} \\ &= (2x + 1)\cos(x^2 + x + 1) \end{aligned}$$

Example 1.10

Find the derivative of $f(x) = (2x^4 - 3\cos x)^{10}$.

Solution: Here the "outer" function is $f(\square) = \square^{10}$ and the "inner" function is $\square = u = 2x^4 - 3\cos x$:

$$\frac{df}{dx} = \frac{df}{du} \cdot \frac{du}{dx} = 10\,\square^9 \cdot \frac{d}{dx}(\square) = 10\,(2x^4 - 3\cos x)^9\,(8x^3 + 3\sin x)$$

Recall that the composition $f \circ g$ of two functions f and g is defined as $(f \circ g)(x) = f(g(x))$. Using prime notation the Chain Rule can be written as:

Chain Rule: If g is a differentiable function of x, and f is a differentiable function on the range of g, then $f \circ g$ is a differentiable function of x, and its derivative with respect to x is:

$$(f \circ g)'(x) \;=\; f'(g(x)) \cdot g'(x)$$

Using the Chain Rule, the Power Rule can be extended to include exponents that are rational numbers:[26]

$$\frac{d}{dx}\left(x^r\right) \;=\; r\,x^{r-1} \quad \text{for any rational number } r$$

To prove this, let $r = m/n$, where m and n are integers with $n \neq 0$. Then $y = x^r = x^{m/n} = (x^m)^{1/n}$, so that $y^n = x^m$. Taking the derivative with respect to x of both sides of this equation gives

$$\frac{d}{dx}\left(y^n\right) \;=\; \frac{d}{dx}\left(x^m\right) \quad \text{, so evaluating the left side by the Chain Rule gives}$$

$$n\,y^{n-1} \cdot \frac{dy}{dx} \;=\; m x^{m-1}$$

$$n\left(x^{m/n}\right)^{n-1} \cdot \frac{dy}{dx} \;=\; m x^{m-1}$$

$$\frac{dy}{dx} \;=\; \frac{m x^{m-1}}{n x^{m-(m/n)}} \;=\; \frac{m}{n}\,x^{m-1-(m-(m/n))} \;=\; \frac{m}{n}\,x^{(m/n)-1} \;=\; r\,x^{r-1} \quad \checkmark$$

Example 1.11

Find the derivative of $f(x) = \sqrt{x}$.

Solution: Since $\sqrt{x} = x^{1/2}$ then by the Power Rule:

$$\frac{df}{dx} \;=\; \frac{d}{dx}\left(x^{1/2}\right) \;=\; \frac{1}{2}\,x^{1/2-1} \;=\; \frac{1}{2}\,x^{-1/2} \;=\; \frac{1}{2\sqrt{x}}$$

Example 1.12

Find the derivative of $f(x) = \frac{2}{3\sqrt{x}}$.

Solution:

$$\frac{df}{dx} \;=\; \frac{d}{dx}\left(\frac{2}{3}\,x^{-1/2}\right) \;=\; \frac{2}{3} \cdot \frac{-1}{2}\,x^{-3/2} \;=\; -\frac{1}{3\,x^{3/2}}$$

[26] It will be shown in Chapter 2 how to define any real number as an exponent. The Power Rule extends to that case as well.

Exercises

A

For Exercises 1-18, find the derivative of the given function.

1. $f(x) = (1 - 5x)^4$

2. $f(x) = 5(x^3 + x - 1)^4$

3. $f(x) = \sqrt{1 - 2x}$

4. $f(x) = (1 - x^2)^{\frac{3}{2}}$

5. $f(x) = \dfrac{\sqrt{x}}{x + 1}$

6. $f(x) = \dfrac{\sqrt{x} + 1}{\sqrt{x} - 1}$

7. $f(t) = \left(\dfrac{1 - t}{1 + t}\right)^4$

8. $f(x) = \left(\dfrac{x^2 + 1}{x - 1}\right)^6$

9. $f(x) = \sin^2 x$

10. $f(x) = \cos\left(\sqrt{x}\right)$

11. $f(x) = 3\tan(5x)$

12. $f(x) = A\cos(\omega x + \phi)$ (A, ω, ϕ are constants)

13. $f(x) = \sec(x^2)$

14. $f(x) = \sin^2\left(\dfrac{1}{1-x}\right) + \cos^2\left(\dfrac{1}{1-x}\right)$

15. $L(\beta) = \dfrac{1}{\sqrt{1 - \beta^2}}$

16. $f(x) = \dfrac{1}{\pi s}\left(1 + \left(\dfrac{x-l}{s}\right)^2\right)^{-1}$ (s, l are constants)

17. $f(x) = \cos(\cos x)$

18. $f(x) = \sqrt{1 + \sqrt{x}}$

B

19. In a certain type of electronic circuit[27] the *overall gain* A_v is given by

$$A_v = \frac{A_o}{1 - T}$$

where the *loop gain* T is a function of the *open-loop gain* A_o.

(a) Show that

$$\frac{dA_v}{dA_o} = \frac{1}{1 - T} - \frac{A_o}{(1 - T)^2}\,\frac{d(1-T)}{dA_o}\,.$$

(b) In the case where T is directly proportional to A_o, use part (a) to show that

$$\frac{dA_v}{dA_o} = \frac{1}{(1 - T)^2}\,.$$

(Hint: First show that $A_o \cdot \frac{d(1-T)}{dA_o} = -T$.*)*

20. Show that the Chain Rule can be extended to 3 functions: if u is a differentiable function of x, v is a differentiable function of u, and f is a differentiable function of v, then

$$\frac{df}{dx} = \frac{df}{dv} \cdot \frac{dv}{du} \cdot \frac{du}{dx}$$

so that f is a differentiable function of x. Notice that the 3 derivatives are linked together as in a chain (hence the name of the rule). The Chain Rule can be extended to any finite number of functions by the above technique.

---

[27] This is an example of a *current-differencing negative feedback amplifier*. See pp.473-479 in Schilling, D.L. and C. Belove, *Electronic Circuits: Discrete and Integrated*, 2nd ed., New York: McGraw-Hill, Inc., 1979.

21. In an internal combustion engine, as a piston moves downward the connecting rod rotates the crank in the clockwise direction, as shown in Figure 1.5.1 below.[28]

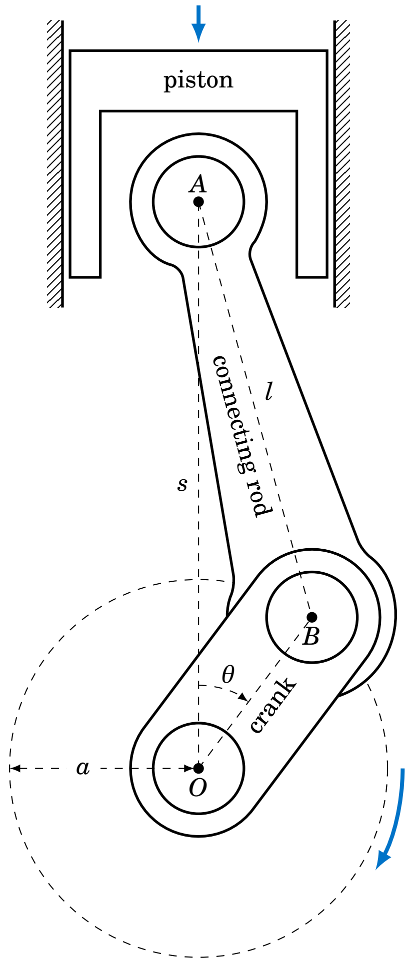


Figure 1.5.1

The point A can only move vertically, causing the point B to move around a circle of radius a centered at the point O, which is directly below the point A and does not move. As the crank rotates it makes an angle θ with the line $\overline{OA}$. Let $l = AB$ and $s = OA$ as in the picture. Assume that all lengths are measured in centimeters, and let the time variable t be measured in minutes.

(a) Show that $s = a\cos\theta + \left(l^2 - a^2\sin^2\theta\right)^{1/2}$ for $0 \le \theta \le \pi$.

(b) The *mean piston speed* is $\bar{S}_p = 2LN$, where $L = 2a$ is the *piston stroke*, and N is the *rotational velocity* of the crank, measured in revolutions per minute (rpm). The instantaneous piston velocity is $S_p = \frac{ds}{dt}$. Let $R = l/a$. Show that for $0 \le \theta \le \pi$,

$$\left|\frac{S_p}{\bar{S}_p}\right| = \frac{\pi}{2}\sin\theta\left[1 + \frac{\cos\theta}{\left(R^2 - \sin^2\theta\right)^{1/2}}\right] .$$

[28]See pp.43-45 in HEYWOOD, J.B., *Internal Combustion Engine Fundamentals*, New York: McGraw-Hill Inc., 1988.

1.6 Higher Order Derivatives

The derivative $f'(x)$ of a differentiable function $f(x)$ can be thought of as a function in its own right, and if it is differentiable then its derivative—denoted by $f''(x)$—is the **second derivative** of $f(x)$ (the **first derivative** being $f'(x)$). Likewise, the derivative of $f''(x)$ would be the **third derivative** of $f(x)$, written as $f'''(x)$. Continuing like this yields the fourth derivative, fifth derivative, and so on. In general the ***n*-th derivative** of $f(x)$ is obtained by differentiating $f(x)$ a total of n times. Derivatives beyond the first are called **higher order derivatives**.

Example 1.13

For $f(x) = 3x^4$ find $f''(x)$ and $f'''(x)$.

Solution: Since $f'(x) = 12x^3$ then the second derivative $f''(x)$ is the derivative of $12x^3$, namely:

$$f''(x) = 36x^2$$

The third derivative $f'''(x)$ is then the derivative of $36x^2$, namely:

$$f'''(x) = 72x$$

Since the prime notation for higher order derivatives can be cumbersome (e.g. writing 50 prime marks for the fiftieth derivative), other notations have been created:

Notation for the second derivative of $y = f(x)$: The following are all equivalent:

$$f''(x) \;,\quad f^{(2)}(x) \;,\quad \frac{d^2y}{dx^2} \;,\quad \frac{d^2}{dx^2}(f(x)) \;,\quad y'' \;,\quad y^{(2)} \;,\quad \ddot{y} \;,\quad \ddot{f}(x) \;,\quad \frac{d^2f}{dx^2} \;,\quad D^2f(x)$$

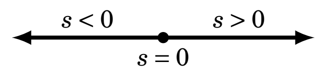
Notation for the *n*-th derivative of $y = f(x)$: The following are all equivalent:

$$f^{(n)}(x) \;,\quad \frac{d^ny}{dx^n} \;,\quad \frac{d^n}{dx^n}(f(x)) \;,\quad y^{(n)} \;,\quad \frac{d^nf}{dx^n} \;,\quad D^nf(x)$$

Notice that the parentheses around n in the notation $f^{(n)}(x)$ indicate that n is not an exponent—it is the number of derivatives to take. The n in the Leibniz notation $\frac{d^ny}{dx^n}$ indicates the same thing, and in general makes working with higher order derivatives easier:

$$\begin{aligned}
\frac{d^2y}{dx^2} &= \frac{d}{dx}\left(\frac{dy}{dx}\right) \\
\frac{d^3y}{dx^3} &= \frac{d}{dx}\left(\frac{d^2y}{dx^2}\right) = \frac{d^2}{dx^2}\left(\frac{dy}{dx}\right) \\
&\vdots \\
\frac{d^ny}{dx^n} &= \frac{d}{dx}\left(\frac{d^{n-1}y}{dx^{n-1}}\right) = \frac{d^{n-1}}{dx^{n-1}}\left(\frac{dy}{dx}\right)
\end{aligned}$$

A natural question to ask is: what do higher order derivatives represent? Recall that the first derivative $f'(x)$ represents the instantaneous rate of change of a function $f(x)$ at the value x. So the second derivative $f''(x)$ represents the instantaneous rate of change of the function $f'(x)$ at the value x. In other words, the second derivative is a rate of change of a rate of change.



The most famous example of this is for motion in a straight line: let $s(t)$ be the position of an object at time t as the object moves along the line. The motion can take two directions, e.g. forward/backward or up/down. Take one direction to represent positive position and the other to represent negative direction, as in the drawing on the right. The (instantaneous) velocity $v(t)$ of the object at time t is $s'(t)$, i.e. the first derivative of $s(t)$. The **acceleration** $a(t)$ of the object at time t is defined as $v'(t)$, the instantaneous rate of change of the velocity. Thus, $a(t) = s''(t)$, i.e. acceleration is the second derivative of position. To summarize:

$$\begin{aligned}
s(t) &= \text{position at time } t \\
v(t) &= \text{velocity at time } t \\
&= \frac{ds}{dt} = s'(t) = \dot{s}(t) \\
a(t) &= \text{acceleration at time } t \\
&= \frac{dv}{dt} = v'(t) = \dot{v}(t) \\
&= \frac{d}{dt}\left(\frac{ds}{dt}\right) = \frac{d^2s}{dt^2} = s''(t) = \ddot{s}(t)
\end{aligned}$$

Example 1.14

Ignoring wind and air resistance, the position s of a ball thrown straight up with an initial velocity of 34 m/s from a starting point 2 m off the ground is given by $s(t) = -4.9t^2 + 34t + 2$ at time t (measured in seconds) with s measured in meters. Find the velocity and acceleration of the ball at any time $t \ge 0$.

Solution: The ball moves in a straight vertical line, first straight up then straight down until it hits the ground. Its velocity $v(t)$ is

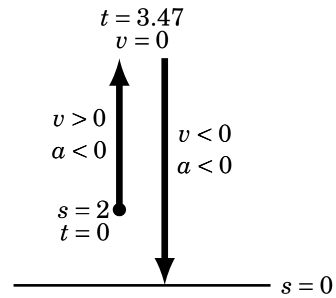
$$v(t) = \frac{ds}{dt} = -9.8t + 34 \text{ m/s}$$

while its acceleration $a(t)$ is

$$a(t) = \frac{d^2s}{dt^2} = \frac{d}{dt}\left(\frac{ds}{dt}\right) = \frac{d}{dt}(-9.8t + 34) = -9.8 \text{ m/s}^2,$$

which is the acceleration due to the force of gravity on Earth.

Note that time $t = 0$ is the time at which the ball was thrown, so that $v(0)$ is the initial velocity of the ball. Indeed, $v(0) = -9.8(0) + 34 = 34$ m/s, as expected.



Notice in Example 1.14 that the acceleration of the ball is not only constant but also negative. To see why this makes sense, first consider the case where the ball is moving upward. The ball has an initial upward velocity of 34 m/s then slows down to 0 m/s at the instant it reaches its maximum height above the ground. So the velocity is decreasing, i.e. its rate of change—the acceleration—is negative.

The ball reaches its maximum height above the ground when its velocity is zero, that is, when $v(t) = -9.8t + 34 = 0$, i.e. at time $t = 34/9.8 = 3.47$ seconds after being thrown (see the picture above). The ball then starts moving downward and its velocity is negative (e.g. at time $t = 4$ s the velocity is $v(4) = -9.8(4) + 34 = -5.2$ m/s). Recall that negative velocity indicates downward motion, while positive velocity means the motion is upward (away from the Earth's center). So in the case where the ball begins moving downward it goes from 0 m/s to a negative velocity, with the ball moving faster towards the ground, which it hits with a velocity of −33.43 m/s (why?). So again the velocity is decreasing, which again means that the acceleration is negative.

Common terminology involving motion might cause some confusion with the above discussion. For example, even though the ball's acceleration is negative as it falls to the ground, it is common to say that the ball is *accelerating* in that situation, not *decelerating* (as the ball is doing while moving upward). In general, **acceleration** is understood to mean that the **magnitude** (i.e. the absolute value) of the velocity is increasing. That magnitude is called the **speed** of the object. **Deceleration** means the speed is decreasing.

The first and second derivatives of an object's position with respect to time represent the object's velocity and acceleration, respectively. Do the third, fourth, and other higher order derivatives have any physical meanings? It turns out they do. The third derivative of position is called the **jerk** of the object. It represents the rate of change of acceleration, and is often used in fields such as vehicle dynamics (e.g. minimizing jerk to provide smoother braking). The fourth, fifth, and sixth derivatives of position are called **snap**, **crackle**, and **pop**, respectively.[29]

The **zero-th derivative** $f^{(0)}(x)$ of a function $f(x)$ is defined to be the function $f(x)$ itself: $f^{(0)}(x) = f(x)$. There is a way to define **fractional derivatives**, e.g. the *one-half derivative* $f^{(1/2)}(x)$, which will be discussed in Chapter 6.

An immediate consequence of the definition of higher order derivatives is:

$$\frac{d^{m+n}}{dx^{m+n}}(f(x)) = \frac{d^m}{dx^m}\left(\frac{d^n}{dx^n}(f(x))\right) \quad \text{for all integers } m \ge 0 \text{ and } n \ge 0.$$

Recall that the **factorial** $n!$ of an integer $n > 0$ is the product of the integers from 1 to n:

$$n! = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot n$$

[29] Yes, those really are their names, obviously inspired by a certain breakfast cereal. Snap has found some uses in flight dynamics, e.g. minimizing snap to optimize flight paths of drones.

For example:

$$1! = 1 \qquad 3! = 1\cdot 2\cdot 3 = 6$$
$$2! = 1\cdot 2 = 2 \qquad 4! = 1\cdot 2\cdot 3\cdot 4 = 24$$

By convention 0! is defined to be 1. The following statement can be proved using induction:

$$\frac{d^n}{dx^n}(x^n) = n! \quad \text{for all integers } n \ge 0$$

Thus,

$$\frac{d^{n+1}}{dx^{n+1}}(x^n) = \frac{d}{dx}\left(\frac{d^n}{dx^n}(x^n)\right) = \frac{d}{dx}(n!) = 0$$

for all integers $n \ge 0$, since $n!$ is a constant. Polynomials are linear combinations of nonnegative powers of a variable (e.g. x), so the above result combined with the Sum Rule and the Constant Multiple rule—which also hold for higher-order derivatives—yields this important fact:

> The $(n+1)$-st derivative ("n plus first derivative") of a polynomial of degree n is 0:
> For any polynomial $p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ of degree n, $\dfrac{d^{n+1}}{dx^{n+1}}(p(x)) = 0.$

This is the basis for the commonly-used statement that "any polynomial can be differentiated to 0" by taking a sufficient number of derivatives. For example, differentiating the polynomial $p(x) = 100x^{100} + 50x^{99}$ 101 times would yield 0 (as would differentiating more than 101 times).

Exercises

A

For Exercises 1-6 find the second derivative of the given function.

1. $f(x) = x^3 + x^2 + x + 1$ **2.** $f(x) = x^2 \sin x$ **3.** $f(x) = \cos 3x$

4. $f(x) = \dfrac{\sin x}{x}$ **5.** $f(x) = \dfrac{1}{x}$ **6.** $F(r) = \dfrac{Gm_1m_2}{r^2}$

7. Find the first five derivatives of $f(x) = \sin x$. Use those to find $f^{(100)}(x)$ and $f^{(2014)}(x)$.

8. Find the first five derivatives of $f(x) = \cos x$. Use those to find $f^{(100)}(x)$ and $f^{(2014)}(x)$.

9. If an object moves along a straight line such that its position $s(t)$ at time t is directly proportional to t for all t (written as $s \propto t$), then show that the object's acceleration is always 0.

B

10. Use induction to show that $\frac{d^n}{dx^n}(x^n) = n!$ for all integers $n \ge 1$.

11. Show that for all integers $n \ge m \ge 1$, $\frac{d^m}{dx^m}(x^n) = \frac{n!}{(n-m)!}x^{n-m}$.

12. Find the general expression for the n-th derivative of $f(x) = \frac{1}{ax+b}$ for all constants a and b ($a \neq 0$).

13. Show that the function $y = A\cos(\omega t + \phi) + B\sin(\omega t + \phi)$ satisfies the *differential equation*

$$\frac{d^2y}{dt^2} + \omega^2 y = 0$$

for all constants A, B, ω, and ϕ.

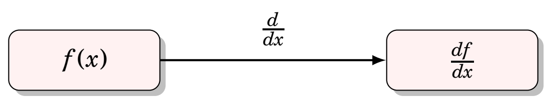
14. If $s(t)$ represents the position at time t of an object moving along a straight line, then show that:

$$s' \text{ and } s'' \text{ have the same sign} \Rightarrow \text{the object is accelerating}$$
$$s' \text{ and } s'' \text{ have opposite signs} \Rightarrow \text{the object is decelerating}$$

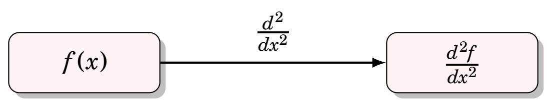
15. For all twice-differentiable functions f and g, show that $(f \cdot g)'' = f'' \cdot g + 2f' \cdot g' + f \cdot g''$.

C

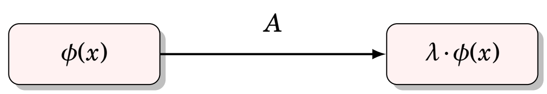
16. Recall that taking a derivative is a way of *operating* on a function. That is, think of $\frac{d}{dx}$ as the *differentiation operator* on the collection of differentiable functions, taking a function $f(x)$ to its derivative function $\frac{df}{dx}$:



Likewise, $\frac{d^2}{dx^2}$ is an operator on twice-differentiable functions, taking a function $f(x)$ to its second derivative function $\frac{d^2f}{dx^2}$:



In general, an *eigenfunction* of an operator A is a function $\phi(x)$ such that $A(\phi(x)) = \lambda \cdot \phi(x)$, that is,



for all x in the domain of ϕ, for some constant λ called the *eigenvalue* of the eigenfunction.

(a) Show for all constants k that $\phi(x) = \cos kx$ is an eigenfunction of the $\frac{d^2}{dx^2}$ operator, and find its eigenvalue. That is, show that $\frac{d^2}{dx^2}(\phi(x)) = \lambda \cdot \phi(x)$ for some constant λ (the eigenvalue).

(b) The *wave function* ψ for a particle of mass m moving in a one-dimensional box of length L, given by

$$\psi(x) = \sqrt{\frac{2}{L}}\,\sin\frac{\pi x}{L} \qquad \text{for } 0 \le x \le L,$$

is a solution (assuming zero potential energy) of the time-independent *Schrödinger equation*

$$-\frac{h^2}{8\pi^2 m}\frac{d^2\psi}{dx^2} = E\,\psi(x)$$

where h is *Planck's constant* and E is a constant that represents the total energy of the wave function. Find an expression for the constant E in terms of the other constants. Notice that this makes $\psi(x)$ an eigenfunction of the $\frac{d^2}{dx^2}$ operator.

CHAPTER 2

Derivatives of Common Functions

2.1 Inverse Functions

The derivatives calculated in the previous chapter were mostly for polynomials and a few trigonometric functions. This chapter will show how to find the derivatives of other types of functions, beginning in this section with inverse functions. The idea here is that if a function is differentiable and has an inverse then that inverse function is also differentiable.

Recall that a **function** is a rule that assigns a single object y from one set (the **range**) to each object x from another set (the **domain**). That rule can be written as $y = f(x)$, where f is the function (see Figure 2.1.1). There is a simple *vertical rule* for determining whether a rule $y = f(x)$ is a function: f is a function if and only if every vertical line intersects the graph of $y = f(x)$ in the xy-coordinate plane at most once (see Figure 2.1.2).

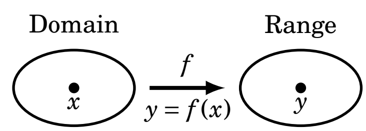
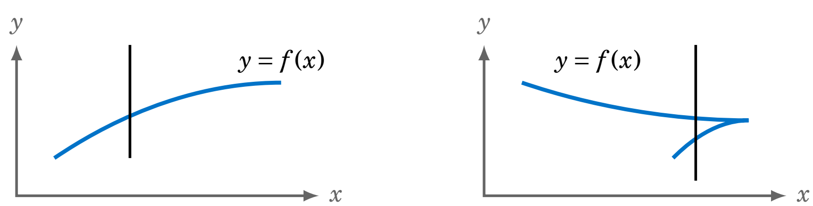


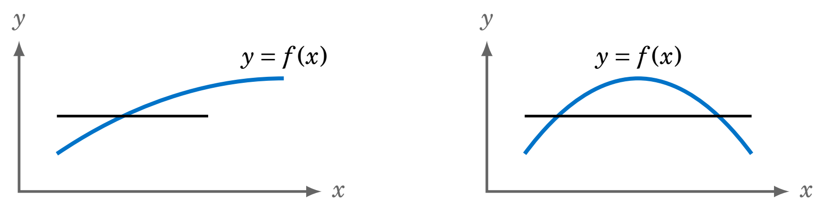
Figure 2.1.1



(a) f is a function **(b)** f is not a function

Figure 2.1.2 Vertical rule for functions

Recall that a function f is **one-to-one** (often written as $1-1$) if it assigns distinct values of y to distinct values of x. In other words, if $x_1 \neq x_2$ then $f(x_1) \neq f(x_2)$. Equivalently, f is one-to-one if $f(x_1) = f(x_2)$ implies $x_1 = x_2$. There is a simple *horizontal rule* for determining whether a function $y = f(x)$ is one-to-one: f is one-to-one if and only if every horizontal line intersects the graph of $y = f(x)$ in the xy-coordinate plane at most once (see Figure 2.1.3).



(a) f is one-to-one **(b)** f is not one-to-one

Figure 2.1.3 Horizontal rule for one-to-one functions

If a function f is one-to-one on its domain, then f has an **inverse function**, denoted by f^{-1}, such that $y = f(x)$ if and only if $f^{-1}(y) = x$. The domain of f^{-1} is the range of f.

The basic idea is that f^{-1} "undoes" what f does, and vice versa. In other words,

$$\begin{aligned} f^{-1}(f(x)) &= x \quad \text{for all } x \text{ in the domain of } f\text{, and} \\ f(f^{-1}(y)) &= y \quad \text{for all } y \text{ in the range of } f. \end{aligned}$$

Intuitively it is clear that a function is one-to-one (and hence invertible) when it is either strictly increasing or strictly decreasing; if the function, say, increases and then decreases (as in Figure 2.1.3(b)) then the horizontal rule would be violated around that "turning point." For differentiable functions a positive derivative means the function is increasing, while a negative derivative means the function is decreasing (this will be proved in Chapter 3).

However, a function can still be one-to-one even if its derivative is zero only at isolated points (i.e. not identically zero over an entire interval of points) and either positive everywhere else or negative everywhere else. For example, the function $f(x) = x^3$ has derivative $f'(x) = 3x^2$, which is zero only at the isolated point $x = 0$ and positive for all other values of x. Clearly f is one-to-one over the set of all real numbers (why?) and hence it has an inverse function $x = f^{-1}(y) = \sqrt[3]{y}$ defined for all real numbers y (i.e. the range of f).

Thus, having a derivative that is either always positive or always negative is *sufficient* for a function to be one-to-one but not *necessary*. Having a nonzero derivative is necessary, though, for the inverse function to be differentiable.

In algebra you learned that $\frac{a}{b} = \frac{1}{\frac{b}{a}}$ for all real numbers $a \neq 0$ and $b \neq 0$ (and hence $\frac{b}{a} \neq 0$). The same holds true for the infinitesimals dy and dx (nonzero by definition) since they can be treated like numbers, which immediately yields a formula for the derivative of an inverse function:

Derivative of an Inverse Function: If $y = f(x)$ is differentiable and has an inverse function $x = f^{-1}(y)$, then f^{-1} is differentiable and its derivative is

$$\frac{dx}{dy} = \frac{1}{\frac{dy}{dx}} \qquad \text{if } \frac{dy}{dx} \neq 0.$$

The inverse of a function would still exist at a point where $\frac{dy}{dx} = 0$ but it would not be differentiable there, since its derivative would be the undefined quantity $\frac{1}{0}$.

Since y is a function of x, $\frac{dy}{dx}$ will be in terms of x and hence $\frac{1}{\frac{dy}{dx}}$ will be in terms of x. However, since (by invertibility) x is a function of y, $\frac{dx}{dy}$ would normally be in terms of y, not x, so that the two sides of the equation $\frac{dx}{dy} = \frac{1}{\frac{dy}{dx}}$ are not in the same terms! One way to handle this discrepancy is to use the formula $y = f(x)$ to solve for x in terms of y then substitute that expression into $\frac{dy}{dx}$, so that $\frac{dx}{dy} = \frac{1}{\frac{dy}{dx}}$ is now in terms of y. That might not always be possible, however (e.g. try solving for x in the formula $y = x \sin x$).

Example 2.1

Find the inverse f^{-1} of the function $f(x) = x^3$ then find the derivative of f^{-1}.

Solution: The function $y = f(x) = x^3$ is one-to-one over the set of all real numbers (why?) so it has an inverse function $x = f^{-1}(y)$ defined for all real numbers, namely $x = f^{-1}(y) = \sqrt[3]{y}$.

The derivative of f^{-1} is

$$\frac{dx}{dy} = \frac{1}{\frac{dy}{dx}} = \frac{1}{3x^2} \quad \text{, which is in terms of } x\text{, so putting it in terms of } y \text{ yields}$$

$$= \frac{1}{3\left(\sqrt[3]{y}\right)^2} = \frac{1}{3y^{2/3}}$$

which agrees with the derivative obtained by differentiating $x = \sqrt[3]{y}$ directly. Note that this derivative is defined for all y except $y = 0$, which occurs when $x = \sqrt[3]{0} = 0$, i.e. at the point $(x, y) = (0, 0)$.

Functions are often expressed in terms of x, so it is common to see an inverse function also expressed in terms of x: writing the inverse of $f(x) = x^3$ as $f^{-1}(x) = \sqrt[3]{x}$ (not as $f^{-1}(y) = \sqrt[3]{y}$)), as confusing as that might be. In that case, the idea is to switch the roles of x and y in the original function $y = f(x)$, making it $x = f(y)$, and then write $y = f^{-1}(x)$ and use

$$\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}}$$

to put the derivative of f^{-1} in terms of x, following the same procedure mentioned earlier.

Example 2.2

Find the inverse f^{-1} of the function $f(x) = x^3$ then find the derivative of f^{-1}.

Solution: Rewrite $y = f(x) = x^3$ as $x = f(y) = y^3$, so that its inverse function $y = f^{-1}(x) = \sqrt[3]{x}$ has derivative

$$\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} = \frac{1}{3y^2} \quad \text{, which is in terms of } y\text{, so putting it in terms of } x \text{ yields}$$

$$= \frac{1}{3\left(\sqrt[3]{x}\right)^2} = \frac{1}{3x^{2/3}}$$

which agrees with the derivative obtained by differentiating $y = \sqrt[3]{x}$ directly.

To obtain a formula in prime notation for the derivative of an inverse function, notice that for all x in the domain of an invertible differentiable function f,

$$f^{-1}(f(x)) = x \quad \Rightarrow \quad \frac{d}{dx}\left(f^{-1}(f(x))\right) = \frac{d}{dx}(x) \quad \Rightarrow \quad \left(f^{-1}\right)'(f(x)) \cdot f'(x) = 1$$

by the Chain Rule, and hence:

$$\left(f^{-1}\right)'(f(x)) = \frac{1}{f'(x)} \qquad \text{if } f'(x) \neq 0$$

Two equivalent ways to write this are:

$$\left(f^{-1}\right)'(c) = \frac{1}{f'(a)} \qquad \text{where } c = f(a) \text{ and } f'(a) \neq 0$$

and

$$\left(f^{-1}\right)'(x) = \frac{1}{f'(f^{-1}(x))} \qquad \text{if } f'(f^{-1}(x)) \neq 0$$

Exercises

A

For Exercises 1-8, show that the given function $y = f(x)$ is one-to-one over the given interval, then find the formulas for the inverse function f^{-1} and its derivative. Use Example 2.2 as a guide, including putting f^{-1} and its derivative in terms of x.

1. $f(x) = x$, for all x

2. $f(x) = 3x$, for all x

3. $f(x) = x^2$, for all $x \geq 0$

4. $f(x) = \sqrt{x}$, for all $x \geq 0$

5. $f(x) = \frac{1}{x}$, for all $x > 0$

6. $f(x) = \frac{1}{x}$, for all $x < 0$

7. $f(x) = \frac{1}{x^2}$, for all $x > 0$

8. $f(x) = x^5$, for all x

9. The unit circle $x^2 + y^2 = 1$ does not define y as a single function of x, since $y = \pm\sqrt{1-x^2}$ defines two separate functions. But the part of the unit circle in the first quadrant, i.e. for $0 \leq x \leq 1$ and $0 \leq y \leq 1$, does define $y = f(x) = \sqrt{1-x^2}$ as a single function of x that is one-to-one on the interval $[0,1]$. Find the formulas for its inverse function f^{-1} and its derivative.

B

10. Show that if f is differentiable and invertible, and if f^{-1} is twice-differentiable, then

$$\left(f^{-1}\right)''(x) = -\frac{f''(f^{-1}(x))}{\left(f'(f^{-1}(x))\right)^3} .$$

2.2 Trigonometric Functions and Their Inverses

The graphs of the six trigonometric functions are shown in Figure 2.2.1:

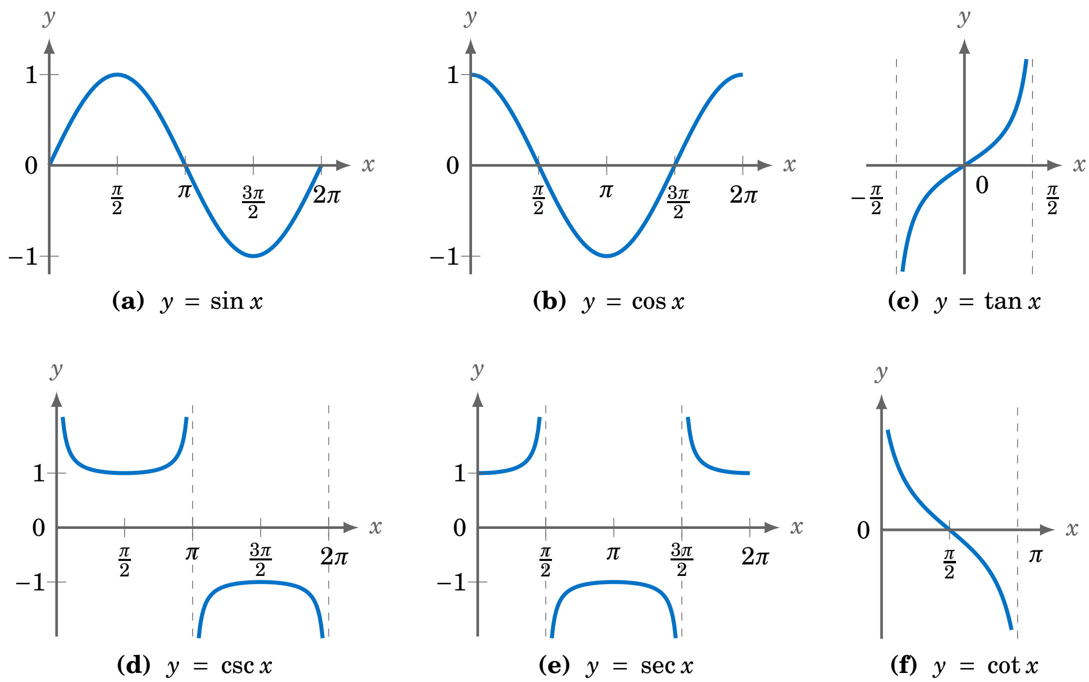


Figure 2.2.1 Graphs of the six trigonometric functions

Recall that $\sin x$, $\cos x$, $\csc x$, and $\sec x$ have a period of 2π (i.e. repeat the same values every 2π radians), while $\tan x$ and $\cot x$ have a period of π.

The derivatives of the six trigonometric functions—given in Section 1.4—are:

$$\frac{d}{dx}(\sin x) = \cos x \qquad \frac{d}{dx}(\csc x) = -\csc x \cot x$$

$$\frac{d}{dx}(\cos x) = -\sin x \qquad \frac{d}{dx}(\sec x) = \sec x \tan x$$

$$\frac{d}{dx}(\tan x) = \sec^2 x \qquad \frac{d}{dx}(\cot x) = -\csc^2 x$$

The six trigonometric functions are not one-to-one over their entire domains, but recall from trigonometry that they are one-to-one when restricted to smaller domains, and hence have inverse functions, called the **inverse trigonometric functions**.

For example, $y = \sin x$ is one-to-one over the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, as shown in Figure 2.2.2 below:

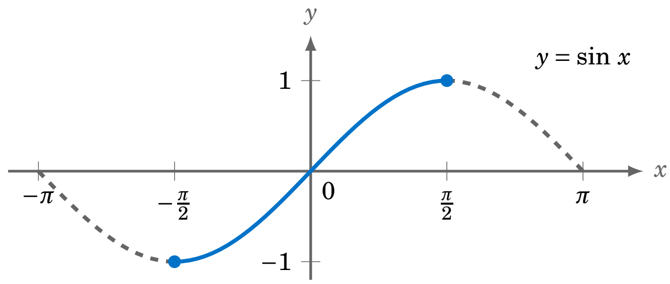
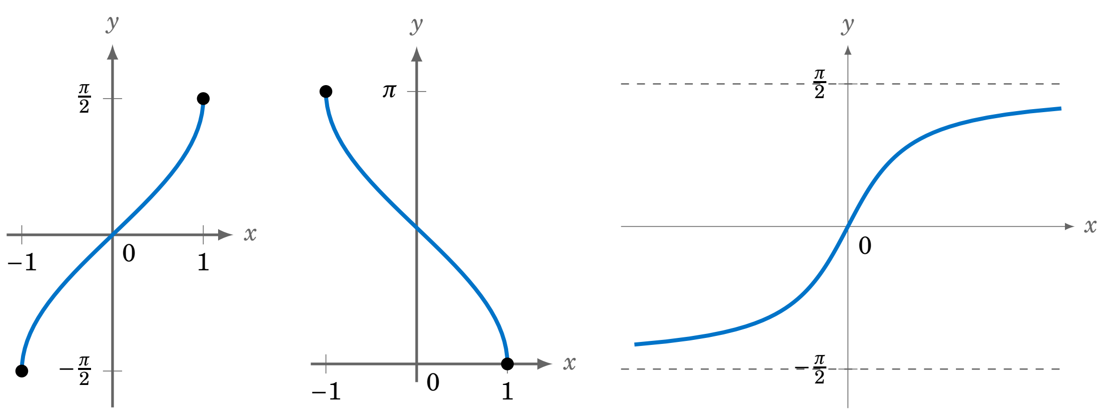


Figure 2.2.2 $y = \sin x$ is one-to-one with x restricted to $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$

Similarly, recall that $\cos x$ is one-to-one over $[0, \pi]$, $\tan x$ is one-to-one over $(-\pi/2, \pi/2)$, $\csc x$ is one-to-one over $(-\pi/2, 0) \cup (0, \pi/2)$, $\sec x$ is one-to-one over $(0, \pi/2) \cup (\pi/2, \pi)$, and $\cot x$ is one-to-one over $(0, \pi)$. Hence, the inverse trigonometric functions $\sin^{-1} x$, $\cos^{-1} x$, $\tan^{-1} x$, $\csc^{-1} x$, $\sec^{-1} x$ and $\cot^{-1} x$ are defined,[1] with the following domains and ranges:

function	$\sin^{-1} x$	$\cos^{-1} x$	$\tan^{-1} x$	$\csc^{-1} x$	$\sec^{-1} x$	$\cot^{-1} x$
domain	$[-1, 1]$	$[-1, 1]$	$(-\infty, \infty)$	$\lvert x \rvert \ge 1$	$\lvert x \rvert \ge 1$	$(-\infty, \infty)$
range	$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$	$[0, \pi]$	$\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$	$\left(-\frac{\pi}{2}, 0\right) \cup \left(0, \frac{\pi}{2}\right)$	$\left(0, \frac{\pi}{2}\right) \cup \left(\frac{\pi}{2}, \pi\right)$	$(0, \pi)$

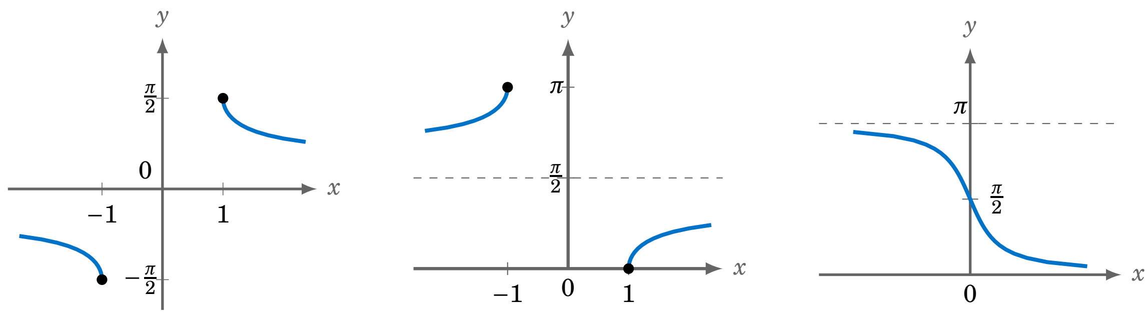
The graphs of all six inverse trigonometric functions are shown in Figures 2.2.3 and 2.2.4 below:



(a) $y = \sin^{-1} x$ **(b)** $y = \cos^{-1} x$ **(c)** $y = \tan^{-1} x$

Figure 2.2.3 Graphs of $\sin^{-1} x$, $\cos^{-1} x$, $\tan^{-1} x$

[1]The *arc* notation $\arcsin x$, $\arccos x$, $\arctan x$, $\text{arccsc}\, x$, $\text{arcsec}\, x$, $\text{arccot}\, x$ is often used in place of $\sin^{-1} x$, $\cos^{-1} x$, $\tan^{-1} x$, $\csc^{-1} x$, $\sec^{-1} x$, $\cot^{-1} x$, respectively.



(a) $y = \csc^{-1} x$ **(b)** $y = \sec^{-1} x$ **(c)** $y = \cot^{-1} x$

Figure 2.2.4 Graphs of $\csc^{-1} x$, $\sec^{-1} x$, $\cot^{-1} x$

The derivatives of the six inverse trigonometric functions are:

$$\frac{d}{dx}(\sin^{-1} x) = \frac{1}{\sqrt{1-x^2}} \quad (\text{for } |x| < 1) \qquad \frac{d}{dx}(\csc^{-1} x) = -\frac{1}{|x|\sqrt{x^2-1}} \quad (\text{for } |x| > 1)$$

$$\frac{d}{dx}(\cos^{-1} x) = -\frac{1}{\sqrt{1-x^2}} \quad (\text{for } |x| < 1) \qquad \frac{d}{dx}(\sec^{-1} x) = \frac{1}{|x|\sqrt{x^2-1}} \quad (\text{for } |x| > 1)$$

$$\frac{d}{dx}(\tan^{-1} x) = \frac{1}{1+x^2} \qquad \frac{d}{dx}(\cot^{-1} x) = -\frac{1}{1+x^2}$$

For the derivative of $\cos^{-1} x$, recall that $y = \cos^{-1} x$ is an *angle* between 0 and π radians, defined for $-1 \le x \le 1$. Since $\cos y = x$ by the definition of y, then $\frac{dx}{dy} = -\sin y$ and

$$\sin^2 y = 1 - \cos^2 y = 1 - x^2 \quad \Rightarrow \quad \sin y = \pm\sqrt{1-x^2} = \sqrt{1-x^2}$$

since $0 \le y \le \pi$ (which means $\sin y$ must be nonnegative). Thus:

$$\frac{d}{dx}(\cos^{-1} x) = \frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} = \frac{1}{-\sin y} = -\frac{1}{\sqrt{1-x^2}} \quad \checkmark$$

For the derivative of $\sec^{-1} x$, since $y = \sec^{-1} x$ is defined for $|x| \ge 1$, then $0 \le y < \pi/2$ for $x \ge 1$ and $\pi/2 < y \le \pi$ for $x \le -1$. Recall also that $\sec y$ and $\tan y$ are both positive when $0 < y < \pi/2$ and are both negative when $\pi/2 < y < \pi$. So in both cases the product $\sec y \tan y$ is nonnegative, i.e. $\sec y \tan y = |\sec y \tan y|$. Thus, since $\sec y = x$ and

$$1 + \tan^2 y = \sec^2 y \quad \Rightarrow \quad \tan^2 y = \sec^2 y - 1 \quad \Rightarrow \quad \tan y = \pm\sqrt{\sec^2 y - 1} = \pm\sqrt{x^2-1}$$

then for $|x| > 1$:

$$\frac{d}{dx}(\sec^{-1} x) = \frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} = \frac{1}{\sec y \tan y} = \frac{1}{|\sec y \tan y|} = \frac{1}{\left|x\sqrt{x^2-1}\right|} = \frac{1}{|x|\sqrt{x^2-1}} \quad \checkmark$$

The proofs of the derivative formulas for the remaining inverse trigonometric functions are similar, and are left as exercises.

Example 2.3

Find the derivative of the function $y = 3\tan(\pi - 2x)$.

Solution: By the Chain Rule with $u = \pi - 2x$, the derivative of $y = 3\tan(\pi - 2x) = 3\tan u$ is:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} = \left(3\sec^2 u\right)(-2) = -6\sec^2(\pi - 2x)$$

Example 2.4

Find the derivative of the function $y = \sin^{-1}(x/4)$.

Solution: By the Chain Rule with $u = x/4$, the derivative of $y = \sin^{-1}(x/4) = \sin^{-1} u$ is:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} = \frac{1}{\sqrt{1-u^2}} \cdot \frac{1}{4} = \frac{1}{4\sqrt{1-(x^2/16)}} = \frac{1}{\sqrt{16-x^2}}$$

Exercises

A

For Exercises 1-16, find the derivative of the given function $y = f(x)$.

1. $y = \sec^2 3x$ **2.** $y = \csc(x^2+1)$ **3.** $y = \cot 3x$ **4.** $y = \cos(\tan x)$

5. $y = \tan^{-1}(x/3)$ **6.** $y = \sec^{-1}(x^2+1)$ **7.** $y = \cot^{-1} 3x$ **8.** $y = \cos^{-1}(\sin x)$

9. $y = \cot^{-1}(1/x)$ **10.** $y = \tan^{-1}\sqrt{x}$ **11.** $y = \left(\sin^{-1} 3x\right)^2$ **12.** $y = \tan^{-1}\frac{1}{x} + \tan^{-1} x$

13. $y = \tan^{-1}\frac{x-1}{x+1}$ **14.** $y = x\sin^{-1}(2x+1)$ **15.** $y = x\cot^{-1} x$ **16.** $y = \tan^{-1}\frac{1}{x} + \cot^{-1} x$

17. Find the derivative of $y = \sin^{-1} x + \cos^{-1} x$. Explain why no derivative formulas were needed.

B

For Exercises 18-21 prove the given derivative formula.

18. $\frac{d}{dx}(\sin^{-1} x) = \frac{1}{\sqrt{1-x^2}}$ **19.** $\frac{d}{dx}(\tan^{-1} x) = \frac{1}{1+x^2}$

20. $\frac{d}{dx}(\cot^{-1} x) = -\frac{1}{1+x^2}$ **21.** $\frac{d}{dx}(\csc^{-1} x) = -\frac{1}{|x|\sqrt{x^2-1}}$

22. The *Chebyshev polynomials* $T_n(x) = \cos(n\cos^{-1} x)$ are defined for all $|x| \le 1$ and $n = 0, 1, 2, \ldots$.

(a) Show that the Chebyshev polynomials $T_n(x)$ satisfy the differential equation

$$(1-x^2)\,T_n''(x) - x\,T_n'(x) + n^2\,T_n(x) = 0$$

(b) Find polynomial expressions for $T_0(x)$, $T_1(x)$ and $T_2(x)$.

(c) Show that $T_{n+1}(x) + T_{n-1}(x) = 2x\,T_n(x)$ for all $n \ge 1$. *(Hint: Write $\theta = \cos^{-1} x$ so that $\cos\theta = x$)*

2.3 The Exponential and Natural Logarithm Functions

Functions of the form a^x, where the exponent x varies, are called *exponential functions*. Unless otherwise noted, assume that $a > 0$ (0^x is just 0, and $(-1)^{1/2}$ is not a real number). You already know how a^x is defined when the exponent x is a rational number (i.e. $x = m/n$ where m and n are integers, $n \neq 0$). But what if x were irrational, such as $\sqrt{2}$? What would $3^{\sqrt{2}}$ mean?

The idea is that $\sqrt{2} = 1.414213562\ldots$ can be approximated by rational numbers $14/10 = 1.4$, $141/100 = 1.41$, $1414/1000 = 1.414$, $14142/10000 = 1.4142$, and so on, taking larger and larger numerators and denominators in the rational approximations to get more and more decimal places of $\sqrt{2}$. Then $3^{\sqrt{2}}$ would be the number that 3 raised to those rational approximations approaches:

$$3^{\sqrt{2}} = \lim_{\substack{m/n \to \sqrt{2} \\ m/n \text{ rational}}} 3^{m/n}$$

Each quantity $3^{m/n}$ is defined inside the above limit, and as the rational numbers m/n get closer to the value of $\sqrt{2}$ it can be shown the limit of the values of $3^{m/n}$ will exist:[2]

$$\begin{aligned}
3^{1.4} &= 3^{\frac{14}{10}} &&= 4.65553672174608 \\
3^{1.41} &= 3^{\frac{141}{100}} &&= 4.70696500171657 \\
3^{1.414} &= 3^{\frac{1414}{1000}} &&= 4.72769503526854 \\
3^{1.4142} &= 3^{\frac{14142}{10000}} &&= 4.72873393017119 \\
3^{1.41421} &= 3^{\frac{141421}{100000}} &&= 4.72878588090861 \\
3^{1.414213} &= 3^{\frac{1414213}{1000000}} &&= 4.72880146624114 \\
&\vdots && \vdots \\
3^{1.414213562\ldots} &= 3^{\sqrt{2}} &&= 4.72880438783742
\end{aligned}$$

Of course you would never do all this by hand—you would simply use a computer or calculator, which use much more efficient algorithms for calculating powers in general.[3]

All the usual rules of exponents that you learned in algebra apply to a^x when defined in the manner described above, with $a > 0$ and x varying over all real numbers. Of all the possible values for the base a, the one that appears the most in mathematics, the sciences and engineering is the base e, defined as:

$$e = \lim_{x\to\infty} \left(1 + \frac{1}{x}\right)^x \tag{2.1}$$

The approximate value of is $e = 2.71828182845905\ldots$ (often called the *Euler number*).

[2]For the general case see pp.61-63 in FRANKLIN, P., *A Treatise on Advanced Calculus*, New York: Dover Publications, Inc., 1964.

[3]For example, to see how square roots and cube roots are calculated see Chapter 2 in FIKE, C.T., *Computer Evaluation of Mathematical Functions*, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1968.

The limit in this definition means that as x becomes larger—approaching infinity (∞)—the values of $\left(1+\frac{1}{x}\right)^x$ approach a number, denoted by e. More decimal places for e can be obtained by making x sufficiently large.[4] For example, when $x = 5\times 10^6$ the value is 2.718281555200129. For extremely large values of x, that is, when $x \gg 1$ (the symbol $\gg$ means "much larger than"),

$$e \approx \left(1+\frac{1}{x}\right)^x \quad\Rightarrow\quad e^{1/x} \approx \left(\left(1+\frac{1}{x}\right)^x\right)^{1/x} = 1+\frac{1}{x} \quad\Rightarrow\quad \left(e^{1/x}-1\right)x \approx \left(\frac{1}{x}\right)x = 1\,,$$

so letting $h = 1/x$, and noting that $h = 1/x \to 0$ if and only if $x\to\infty$, yields the useful limit:[5]

$$\lim_{h\to 0} \frac{e^h - 1}{h} = 1 \tag{2.2}$$

Using the above limit, the derivative of $y = e^x$ can be found:

$$\frac{d}{dx}\left(e^x\right) = e^x$$

Proof: Using the limit definition of the derivative for $f(x) = e^x$,

$$\begin{aligned}\frac{d}{dx}\left(e^x\right) &= \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = \lim_{h\to 0}\frac{e^{x+h}-e^x}{h}\\ &= \lim_{h\to 0}\frac{e^x\left(e^h-1\right)}{h} = e^x\cdot\lim_{h\to 0}\frac{e^h-1}{h} \quad \text{(since } e^x \text{ does not depend on } h\text{)}\\ &= e^x\cdot 1 = e^x\end{aligned}$$

QED

In general, for a differentiable function $u = u(x)$ as the exponent the Chain Rule yields:

$$\frac{d}{dx}\left(e^u\right) = e^u\cdot\frac{du}{dx}$$

Example 2.5

Find the derivative of $y = 4e^{-x^2}$.

Solution: $\frac{dy}{dx} = 4e^{-x^2}\cdot\frac{d}{dx}(-x^2) = 4e^{-x^2}\cdot(-2x) = -8xe^{-x^2}$

---

[4]It will be shown in Chapter 9 that this limit does in fact exist.

[5]This is admittedly a "hand waving" argument. In Chapter 3 a more exact method will be discussed for proving limits like this one.

The function e^x is often referred to simply as **the exponential function**, even though there are obviously many exponential functions. What makes the base e so special? Take $y = Ae^{kt}$ to represent the amount of some physical quantity at time t, for some constants A and k. Then

$$\frac{dy}{dt} = \frac{d}{dt}\left(Ae^{kt}\right) = k \cdot Ae^{kt} = ky \, ,$$

which says that *the instantaneous rate of change of the quantity is directly proportional to the amount present at that instant*. It turns out that many physical quantities exhibit that behavior, some of which will be discussed shortly. Conversely, in Chapter 5 it will be shown that any solution to the differential equation $\frac{dy}{dt} = ky$ must be of the form $y = Ae^{kt}$ for some constant A. This is what gives the exponential function its special significance.

Let $f(x) = e^x$ be the exponential function. Then $f(x) > 0$ for all x and $f'(x) = f(x) = e^x > 0$ for all x, and so $f(x)$ is strictly increasing. The graph is shown in Figure 2.3.1.

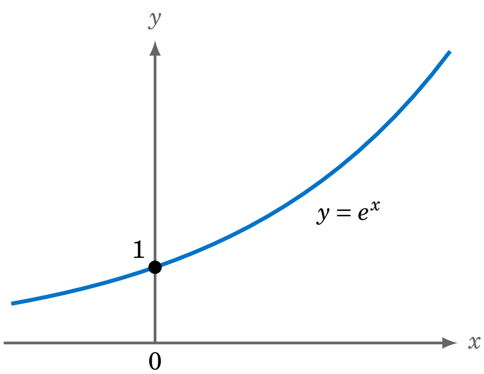


Figure 2.3.1 The exponential function

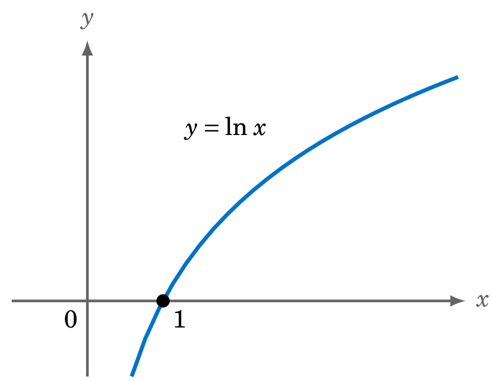


Figure 2.3.2 Natural logarithm function

Thus, the exponential function is one-to-one over the set of all real numbers and hence has an inverse function, called the **natural logarithm function**, denoted (as a function of x) as $f^{-1}(x) = \ln x$. The graph is shown in Figure 2.3.2. Below is a summary of the relationship between e^x and $\ln x$:

$$\begin{aligned} \text{domain of } \ln x &= \text{all } x > 0 = \text{range of } e^x \\ \text{range of } \ln x &= \text{all } x \quad\;\; = \text{domain of } e^x \end{aligned}$$

$$y = e^x \quad \text{if and only if} \quad x = \ln y \tag{2.3}$$

$$e^{\ln x} = x \quad \text{for all } x > 0 \tag{2.4}$$

$$\ln\left(e^x\right) = x \quad \text{for all } x \tag{2.5}$$

The reader should be aware that many—if not most—fields outside of mathematics use the notation $\log x$ instead of $\ln x$ for the natural logarithm function.[6]

[6] This text almost used $\log x$ as well, prevented only by the desire for compatibility with other mathematics texts.

From algebra you should be familiar with the following properties of the natural logarithm, along with their equivalent properties in terms of the exponential function:[7]

$$\ln(ab) = \ln a + \ln b \qquad e^a \cdot e^b = e^{a+b}$$

$$\ln\left(\frac{a}{b}\right) = \ln a - \ln b \qquad \frac{e^a}{e^b} = e^{a-b}$$

$$\ln a^b = b \ln a \qquad \left(e^a\right)^b = e^{ab}$$

$$\ln 1 = 0 \qquad e^0 = 1$$

To find the derivative of $y = \ln x$, use $x = e^y$:

$$\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} = \frac{1}{\frac{d}{dy}(e^y)} = \frac{1}{e^y} = \frac{1}{x}$$

Hence:

$$\frac{d}{dx}(\ln x) = \frac{1}{x}$$

In general, for a differentiable function $u = u(x)$, the Chain Rule yields:

$$\frac{d}{dx}(\ln u) = \frac{1}{u} \cdot \frac{du}{dx} = \frac{u'}{u}$$

Example 2.6

Find the derivative of $y = \ln\left(x^2 + 3x - 1\right)$.

Solution: $\dfrac{dy}{dx} = \dfrac{1}{x^2+3x-1} \cdot \dfrac{d}{dx}(x^2+3x-1) = \dfrac{2x+3}{x^2+3x-1}$

Recall that $|x| = -x$ for $x < 0$, in which case $\ln(-x)$ is defined and

$$\frac{d}{dx}(\ln|x|) = \frac{d}{dx}(\ln(-x)) = \frac{1}{-x} \cdot (-1) = \frac{1}{x}\,.$$

[7] Note that when using the formula $\ln\left(\frac{a}{b}\right) = \ln a - \ln b$ in numerical computations—especially on hand-held calculators—it is preferable to use the left side of the equation, i.e. $\ln\left(\frac{a}{b}\right)$, since the right side $\ln a - \ln b$ is vulnerable to the problem of *subtractive cancellation*, which can give an incorrect answer of 0 if a and b are nearly equal. For a discussion of subtractive cancellation see § 1.3 in HENRICI, P., *Essentials of Numerical Analysis, with Pocket Calculator Demonstrations*, New York: John Wiley & Sons, Inc., 1982.

Combine that result with the derivative $\frac{d}{dx}(\ln x) = \frac{1}{x}$ for $x > 0$ to get:

$$\frac{d}{dx}(\ln|x|) = \frac{1}{x}$$

Logarithmic Differentiation

For some functions it is easier to differentiate the natural logarithm of the function first and then solve for the derivative of the original function. This technique is called **logarithmic differentiation**, demonstrated in the following two examples.

Example 2.7

Find the derivative of $y = x^x$.

Solution: For this example assume $x > 0$ (since x is both the base and the exponent). Note that you cannot use the Power Rule for this function since the exponent x is a variable, not a fixed number. Instead, take the natural logarithm of both sides of the equation $y = x^x$ and then take the derivative of both sides and solve for y':

$$\begin{aligned}
\ln y &= \ln\left(x^x\right) = x \cdot \ln x \\
\frac{d}{dx}(\ln y) &= \frac{d}{dx}(x \cdot \ln x) \\
\frac{y'}{y} &= 1 \cdot \ln x + x \cdot \frac{1}{x} \\
y' &= y(\ln x + 1) = x^x(\ln x + 1)
\end{aligned}$$

Example 2.8

Find the derivative of $y = \frac{(2x+1)^7(3x^3-7x+6)^4}{(1+\sin x)^5}$.

Solution: Use logarithmic differentiation by taking the natural logarithm of y and then use properties of logarithms to simplify the differentiation before solving for y':

$$\begin{aligned}
\ln y &= \ln\left(\frac{(2x+1)^7(3x^3-7x+6)^4}{(1+\sin x)^5}\right) = \ln\left((2x+1)^7(3x^3-7x+6)^4\right) - \ln\left((1+\sin x)^5\right) \\
&= \ln\left((2x+1)^7\right) + \ln\left((3x^3-7x+6)^4\right) - \ln\left((1+\sin x)^5\right) \\
\frac{d}{dx}(\ln y) &= \frac{d}{dx}\left(7\ln(2x+1) + 4\ln(3x^3-7x+6) - 5\ln(1+\sin x)\right) \\
\frac{y'}{y} &= 7 \cdot \frac{2}{2x+1} + 4 \cdot \frac{9x^2-7}{3x^3-7x+6} - 5 \cdot \frac{\cos x}{1+\sin x} \\
y' &= y \cdot \left(\frac{14}{2x+1} + \frac{36x^2-28}{3x^3-7x+6} - \frac{5\cos x}{1+\sin x}\right) \\
&= \frac{(2x+1)^7(3x^3-7x+6)^4}{(1+\sin x)^5} \cdot \left(\frac{14}{2x+1} + \frac{36x^2-28}{3x^3-7x+6} - \frac{5\cos x}{1+\sin x}\right)
\end{aligned}$$

Radioactive Decay

A classic example of the differential equation $\frac{dy}{dt} = ky$ is the case of *exponential decay* of a radioactive substance, often referred to simply as *radioactive decay*. In this case the general solution $y = Ae^{kt}$ represents the amount of the substance at time $t \ge 0$, and the *decay constant* k is negative: $\frac{dy}{dt} < 0$ since the substance is decaying (i.e. the amount of substance is decreasing) while $y > 0$, so $\frac{dy}{dt} = ky$ implies that $k < 0$.

The constant A is the initial amount of the substance, i.e. the amount at time $t = 0$: $y(0) = Ae^{0t} = Ae^0 = A$. For this reason A is sometimes denoted by A_0. The constant k turns out to be related to the **half-life** of the substance, defined as the time t_H required for half the current amount of substance to decay (see Figure 2.3.3).

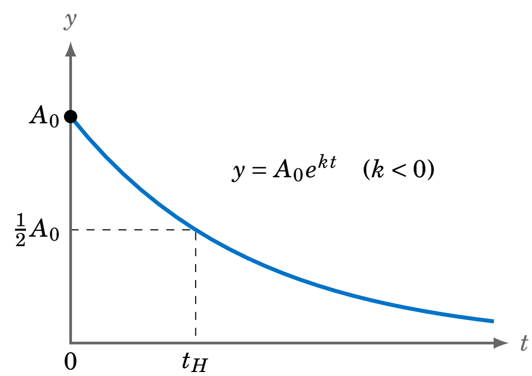


Figure 2.3.3

You might be tempted to think that the half-life is not a constant, that it might change depending on the amount of substance present. For example, perhaps it would take longer for 100 g of the substance to decay to 50 g than it would for 10 g to decay to 5 g. However, this is not so. To see why, pick any $t \ge 0$ as the current time, so that $y(t) = A_0e^{kt}$ is the current amount of the substance. By definition, that amount should be halved when the time t_H has passed, that is, $y(t + t_H) = \frac{1}{2}y(t)$. Then t_H does turn out to be independent of the initial amount A_0 and depends only on k, since

$$y(t+t_H) = \frac{1}{2}y(t) \quad\Rightarrow\quad A_0e^{k(t+t_H)} = \frac{1}{2}A_0e^{kt} \quad\Rightarrow\quad \cancel{A_0e^{kt}}\cdot e^{kt_H} = \frac{1}{2}\cancel{A_0e^{kt}}$$

$$\Rightarrow\quad e^{kt_H} = \frac{1}{2} \quad\Rightarrow\quad kt_H = \ln\left(\frac{1}{2}\right) = -\ln 2$$

and so:

$$t_H = -\frac{\ln 2}{k} \qquad \text{and} \qquad k = -\frac{\ln 2}{t_H} \tag{2.6}$$

Example 2.9

Suppose that 5 mg of a radioactive substance decays to 3 g in 6 hours. Find the half-life of the substance.

Solution: Consider $A_0 = 5$ mg as the initial amount, so that $y(t) = 5e^{kt}$ is the amount at time $t \ge 0$ hours. Use the given information that $y(6) = 3$ mg to find k, the decay constant of the substance:

$$3 = y(6) = 5e^{k6} \quad\Rightarrow\quad 6k = \ln\left(\frac{3}{5}\right) \quad\Rightarrow\quad k = \frac{1}{6}\ln 0.6$$

Then the half-life t_H is:

$$t_H = -\frac{\ln 2}{k} = -\frac{\ln 2}{\frac{1}{6}\ln 0.6} = 8.14 \text{ hours}$$

Note in the above example that the given time $t = 6$ was used for finding the constant k and then the half-life t_H. For the converse problem—given the half-life find the time required for a certain amount to decay—you would do the opposite: use the given t_H to find k and then solve for the required time t from the equation $y(t) = A_0 e^{kt}$.

Example 2.10

Another example of the differential equation $\frac{dy}{dt} = ky$ is *exponential growth* of cell bacteria, in which case $k > 0$ since the number of cells $y(t)$ at time t is increasing.

Example 2.11

Another example is for the current I in a simple series electric circuit with a constant direct current (DC) source of voltage V, a capacitor with capacitance C, a resistor with resistance R, and a switch, as in Figure 2.3.4. If the capacitor is initially uncharged when the switch is open, and if the switch is closed at time $t = 0$, then the current $I(t)$ through the circuit at time $t \ge 0$ satisfies (by Kirchoff's Second Law) the differential equation

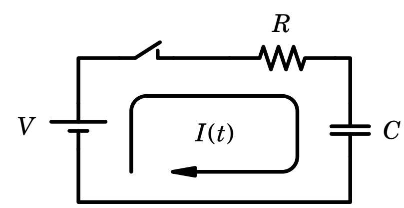


Figure 2.3.4

$$RC\frac{dI}{dt} + I = 0 \quad \Rightarrow \quad \frac{dI}{dt} = -\frac{I}{RC}$$

so that $I(t) = I_0 e^{-t/RC}$ where I_0 is the initial current at $t = 0$. *Ohm's Law* says that $V = I_0 R$, so

$$I(t) = \frac{V}{R} e^{-t/RC}$$

is the current at time $t \ge 0$, which decreases exponentially.

Example 2.12

In the previous examples the quantities that decayed or grew exponentially did so as functions of time. There are other possible variables besides time, though. For example, the atmospheric pressure p measured as a function of height h above the surface of the Earth satisfies—assuming constant temperature—the differential equation

$$\frac{dp}{dh} = -\frac{w_0}{p_0} p$$

where p_0 is the pressure at height $h = 0$ (i.e. ground level) and w_0 is the weight of a cubic foot of air at pressure p_0 (with air pressure measured in lbs per square foot and height measured in feet). Thus,

$$p(h) = p_0 \, e^{-\frac{w_0}{p_0} h} .$$

So the atmospheric pressure decreases exponentially as the height above the ground increases.

Exercises

A

For Exercises 1-12, find the derivative of the given function.

1. $y = e^{2x}$

2. $y = xe^{x^2}$

3. $y = e^{-x} - e^x$

4. $y = e^{\sin x}$

5. $y = \dfrac{1 + e^x}{1 - e^x}$

6. $y = \dfrac{1}{1 + e^{-2x}}$

7. $y = e^{e^x}$

8. $y = e^{2 \ln x}$

9. $y = \ln(3x)$

10. $y = \ln(x^2 + 2x + 1)^4$

11. $y = \left(\ln(\tan x^2)\right)^3$

12. $y = \ln(e^x + e^{2x})$

13. Show that $\dfrac{d}{dx}(\ln(kx)) = \dfrac{1}{x}$ for all constants $k > 0$.

14. Show that $\dfrac{d}{dx}(\ln(x^n)) = \dfrac{n}{x}$ for all integers $n \ge 1$.

For Exercises 15-18, use logarithmic differentiation to find $\frac{dy}{dx}$.

15. $y = x^{x^2}$

16. $y = x^{\ln x}$

17. $y = x^{\sin x}$

18. $y = \dfrac{(x+2)^8(3x-1)^7}{(1-5x)^4}$

B

19. Suppose it takes 8 hours for 30% of a radioactive substance to decay. Find the half-life of the substance.

20. The radioactive isotope radium-223 has a half-life of 11.43 days. How long would it take for 3 kg of radium-223 to decay to 1 kg?

21. If a certain cell population grows exponentially—i.e. is of the form A_0e^{kt} with $k > 0$—and if the population doubles in 6 hours, how long would it take for the population to quadruple?

For Exercises 22-25, use induction to prove the given formula for all $n \ge 0$.

22. $\dfrac{d^n}{dx^n}\left(e^{kx}\right) = k^n e^{kx}$ (any constant $k \neq 0$)

23. $\dfrac{d^n}{dx^n}(xe^x) = (x+n)e^x$

24. $\dfrac{d^n}{dx^n}(xe^{-x}) = (-1)^n(x-n)e^{-x}$

25. $\dfrac{d^{n+1}}{dx^{n+1}}(x^n \ln x) = \dfrac{n!}{x}$

26. Show that $f'(x) = f(x)(1-f(x))$ for the *sigmoid neuron function* $f(x) = \dfrac{1}{1+e^{-x}}$. This derivative relation is used in neural network learning algorithms.

27. If $y = Ce^{-\kappa t}\cos\left(\sqrt{n^2-\kappa^2}\,t + \gamma\right)$ then show that

$$\frac{d^2y}{dt^2} + 2\kappa\frac{dy}{dt} + n^2y = 0$$

for all constants C, n, κ, γ, with $0 \le \kappa \le n$.

C

28. Suppose that $e^y + e^x = e^{y+x}$. Show that $\frac{dy}{dx} = -e^{y-x}$.

29. For an infinitesimal dx show that $e^{dx} = 1 + dx$. *(Hint: Use $\frac{d}{dx}(e^x) = e^x$.)*

30. For an infinitesimal dx show that $\ln(1+dx) = dx$.

2.4 General Exponential and Logarithmic Functions

For a general exponential function $y = a^x$, with $a > 0$, use logarithmic differentiation to find its derivative:

$$\begin{aligned} \ln y &= \ln\left(a^x\right) = x \ln a \\ \frac{d}{dx}(\ln y) &= \frac{d}{dx}(x \ln a) = \ln a \\ \frac{y'}{y} &= \ln a \quad \Rightarrow \quad y' = y \cdot \ln a \end{aligned}$$

Thus, the derivative of $y = a^x$ is:

$$\frac{d}{dx}\left(a^x\right) = (\ln a)\, a^x$$

In general, for an exponent of the form $u = u(x)$:

$$\frac{d}{dx}\left(a^u\right) = (\ln a)\, a^u \cdot \frac{du}{dx}$$

Example 2.13

Find the derivative of $y = 2^{\cos x}$.

Solution: This is the case where $a = 2$, so:

$$\frac{dy}{dx} = (\ln 2)\, 2^{\cos x} \cdot \frac{d}{dx}(\cos x) = -(\ln 2)(\sin x)\, 2^{\cos x}$$

Note that any exponential function $y = a^x$ can be expressed in terms of the exponential function e^x. Since

$$a^x > 0 \quad \Rightarrow \quad e^{\ln(a^x)} = a^x \ ,$$

and since $\ln(a^x) = x \ln a$, then:

$$a^x = e^{x \ln a}$$

Computers and calculators often use the above formula to calculate a^x.

The function $y = a^x$ has an inverse for any $a > 0$, except for $a = 1$ (in that case $y = 1^x = 1$ is just a constant function). To see this, notice that since $a^x > 0$ for all x, and $\ln a < 0$ for $0 < a < 1$, while $\ln a > 0$ for $a > 1$, then $\frac{dy}{dx} = (\ln a)a^x$ is always negative if $0 < a < 1$ and always positive if $a > 1$. Thus, $y = a^x$ is a strictly decreasing function if $0 < a < 1$, and it is a strictly increasing function if $a > 1$. The graphs in each case are shown in Figure 2.4.1.

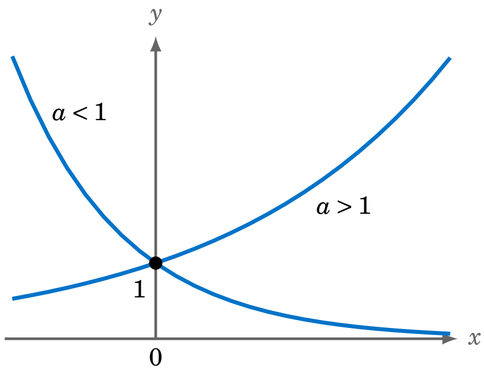


Figure 2.4.1 $y = a^x$

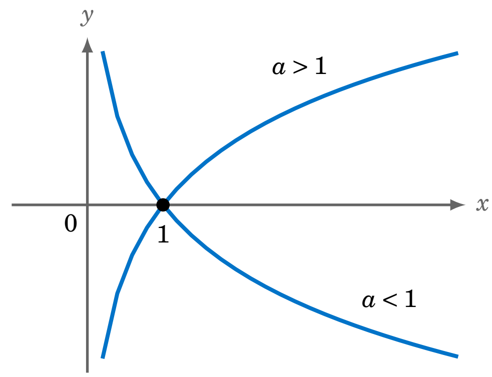


Figure 2.4.2 $y = \log_a x$

Hence, for any $a > 0$ with $a \neq 1$ the function $f(x) = a^x$ is one-to-one, so it has an inverse function, called the **base *a* logarithm** and denoted by $f^{-1}(x) = \log_a x$. It is often spoken as "log base a of x". The graphs for $a < 1$ and $a > 1$ are shown in Figure 2.4.2. Note that the natural logarithm is just the base a logarithm in the special case with $a = e$, i.e. $\ln x = \log_e x$. The base a logarithm has properties similar to those of the natural logarithm (and the corresponding properties of a^x):

$$\log_a(bc) = \log_a b + \log_a c \qquad a^b \cdot a^c = a^{b+c}$$

$$\log_a\left(\frac{b}{c}\right) = \log_a b - \log_a c \qquad \frac{a^b}{a^c} = a^{b-c}$$

$$\log_a b^c = c\log_a b \qquad \left(a^b\right)^c = a^{bc}$$

$$\log_a 1 = 0 \qquad a^0 = 1$$

Note that $\log_a x$ can be put in terms of the natural logarithm, since

$$x = a^{\log_a x} \quad \Rightarrow \quad \ln x = \ln\left(a^{\log_a x}\right) = (\log_a x)\cdot(\ln a)$$

so dividing the last expression by $\ln a$ gives:

$$\log_a x = \frac{\ln x}{\ln a}$$

The above formula is useful on calculators that do not have a $\log_a x$ key or function. Taking the derivative of both sides yields:

$$\frac{d}{dx}\left(\log_a x\right) = \frac{1}{x \ln a}$$

In general, when taking the logarithm of a function $u = u(x)$:

$$\frac{d}{dx}\left(\log_a u\right) = \frac{1}{u \ln a} \cdot \frac{du}{dx} = \frac{u'}{u \ln a}$$

Example 2.14

Find the derivative of $y = \log_2(\cos 4x)$.

Solution: This is the case where $a = 2$, so:

$$\frac{dy}{dx} = \frac{1}{(\cos 4x)(\ln 2)} \cdot \frac{d}{dx}(\cos 4x) = -\frac{4 \sin 4x}{(\ln 2)(\cos 4x)}$$

The number a is the **base** of both the logarithm function $\log_a x$ and the exponential function a^x. Base 2 and base 10 are the most commonly used bases other than base e. Base 10 is how numbers are normally expressed, as combinations of powers of 10 (e.g. $2014 = \mathbf{2}\cdot 10^3 + \mathbf{0}\cdot 10^2 + \mathbf{1}\cdot 10^1 + \mathbf{4}\cdot 10^0$). Base 2 is especially useful in computer science, since computers represent all numbers in *binary* format, i.e. as a sequence of zeros and ones, indicating how many successive powers of two to take and then sum up.[8] For example, the number 6 is represented in binary format as 110, since $\mathbf{1}\cdot 2^2 + \mathbf{1}\cdot 2^1 + \mathbf{0}\cdot 2^0 = 4+2+0 = 6$.

Exercises

A

For Exercises 1-9, find the derivative of the given function.

1. $y = \dfrac{3^x + 3^{-x}}{2}$

2. $y = 2^{\ln 3x}$

3. $y = 2^{2^x}$

4. $y = \tan^{-1} \pi^x$

5. $y = \log_2(x^2+1)$

6. $y = \log_{10} e^x$

7. $y = \sin\left(\log_2 \pi x\right)$

8. $y = \log_2 4^{2x}$

9. $y = 8^{\log_2 x}$

B

10. Show that for all constants k the function $y = Aa^{\frac{kx}{\ln a}}$ satisfies the differential equation $\frac{dy}{dx} = ky$. Does this contradict the statement made in Section 2.3 that the only solution to that differential equation is of the form $y = Ae^{kx}$? Explain your answer.

[8] Binary notation leads to the joke "There are 10 kinds of people in the world: those who understand binary and those who do not."

CHAPTER 3

Topics in Differential Calculus

3.1 Tangent Lines

Everyone knows that the Earth is not flat, but *locally*, e.g. in your immediate vicinity, isn't the Earth *effectively* flat? In other words, "flat" is a fairly good approximation of the Earth's surface "near" you, and it simplifies matters enough for you to do some useful things.

This idea of approximating curved shapes by straight shapes is a frequent theme in calculus. Recall from Chapter 1 that by the Microstraightness Property a differentiable curve $y = f(x)$ actually *is* a straight line over an infinitesimal interval, having slope $\frac{dy}{dx}$. The extension of that line to all values of x is called the *tangent line*:

For a curve $y = f(x)$ that is differentiable at $x = a$, the **tangent line** to the curve at the point $P = (a, f(a))$ is the unique line through P with slope $m = f'(a)$. P is called the **point of tangency**. The equation of the tangent line is thus given by:

$$y - f(a) = f'(a) \cdot (x - a) \tag{3.1}$$

Figure 3.1.1 on the right shows the tangent line to a curve $y = f(x)$ at a point P. If you were to look at the curve near P with a microscope, it would look almost identical to its tangent line through P. Why is this line—of all possible lines through P—such a good approximation of the curve near P? It is because at the point P the tangent line and the curve both have the *same rate of change*, namely, $f'(a)$. So the curve's values and the line's values change by roughly the same amount slightly away from P (where the line and curve have the same value), making their values nearly equal.

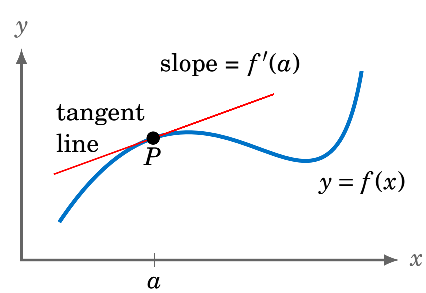


Figure 3.1.1 Tangent line

Example 3.1

Find the tangent line to the curve $y = x^2$ at $x = 1$.

Solution: By formula (3.1), the equation of the tangent line is

$$y - f(a) = f'(a) \cdot (x - a)$$

with $a = 1$ and $f(x) = x^2$. So $f(a) = f(1) = 1^2 = 1$. Both the curve $y = x^2$ and the tangent line pass through the point $(1, f(1)) = (1,1)$. The derivative of $f(x) = x^2$ is $f'(x) = 2x$, so $f'(a) = f'(1) = 2(1) = 2$, which is the slope of the tangent line at $(1,1)$. Hence, the equation of the tangent line is $y - 1 = 2(x - 1)$, or (in slope-intercept form) $y = 2x - 1$.

The curve and tangent line are shown in Figure 3.1.2. Near the point $(1,1)$ the curve and tangent line are close together, but the separation grows farther from that point, especially in the negative x direction.

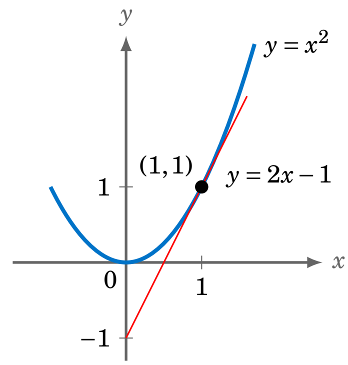
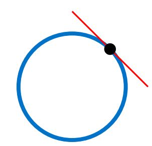


Figure 3.1.2

In trigonometry you probably learned about tangent lines to circles, where a tangent line is defined as the unique line that touches the circle at only one point, as in the figure on the right. In this case the tangent line is always on one side of the circle, namely, the exterior of the circle; it does not cut through the interior of the circle. In fact, that definition is a special case of the calculus definition. In general, though, the tangent line to any other type of curve will not necessarily be on only one side of the curve, as it was in Example 3.1.



Example 3.2

Find the tangent line to the curve $y = x^3$ at $x = 0$.

Solution: Use formula (3.1) with $a = 0$ and $f(x) = x^3$. Then $f(a) = f(0) = 0^3 = 0$. The derivative of $f(x) = x^3$ is $f'(x) = 3x^2$, so $f'(a) = f'(0) = 3(0)^2 = 0$. Hence, the equation of the tangent line is $y - 0 = 0(x - 0)$, which is $y = 0$. In other words, the tangent line is the x-axis itself.

As shown in Figure 3.1.3, the tangent line cuts through the curve. In general it is possible for a tangent line to intersect the curve at more than one point, depending on the function.

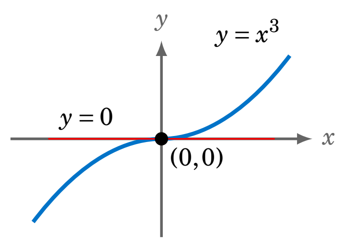


Figure 3.1.3

Example 3.3

Find the tangent line to the curve $y = \sin x$ at $x = 0$.

Solution: Use formula (3.1) with $a = 0$ and $f(x) = \sin x$. Then $f(a) = f(0) = \sin 0 = 0$. The derivative of $f(x) = \sin x$ is $f'(x) = \cos x$, so $f'(a) = f'(0) = \cos 0 = 1$. Hence, the equation of the tangent line is $y - 0 = 1(x - 0)$, which is $y = x$, as in Figure 3.1.4. Near $x = 0$, the tangent line $y = x$ is close to the line $y = \sin x$, which was shown in Section 1.3 (namely, $\sin dx = dx$, so that $\sin x \approx x$ for $x \ll 1$).

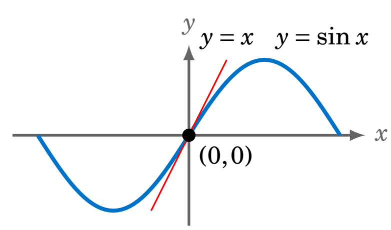


Figure 3.1.4

There are several important things to note about tangent lines:

- *The slope of a curve's tangent line is the slope of the curve.*
  Since the slope of a tangent line equals the derivative of the curve at the point of tangency, the **slope of a curve** at a particular point can be defined as the slope of its tangent line at that point. So curves can have varying slopes, depending on the point, unlike straight lines, which have a constant slope. An easy way to remember all this is to think "slope = derivative."

- *The tangent line to a straight line is the straight line itself.*
  This follows easily from the definition of a tangent line, but is also easy to see with the "slope = derivative" idea: a straight line's slope (i.e. derivative) never changes, so its tangent line—having the same slope—will be parallel and hence must coincide with the straight line (since they have the points of tangency in common). For example, the tangent line to the straight line $y = -3x+2$ is $y = -3x+2$ at every point on the straight line.

- *The tangent line can be thought of as a limit of secant lines.*
  A **secant line** to a curve is a line that passes through two points on the curve. Figure 3.1.5 shows a secant line L_{PQ} passing through the points $P = (x_0, f(x_0))$ and $Q = (w, f(w))$ on the curve $y = f(x)$,

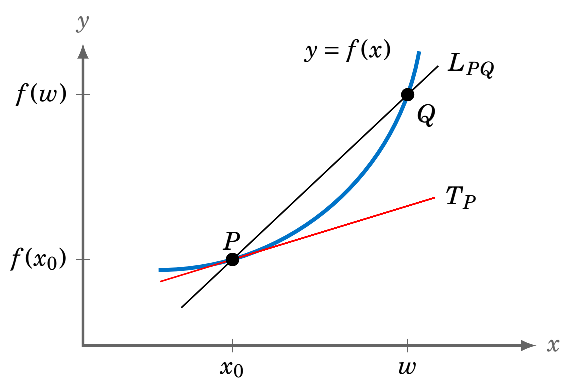


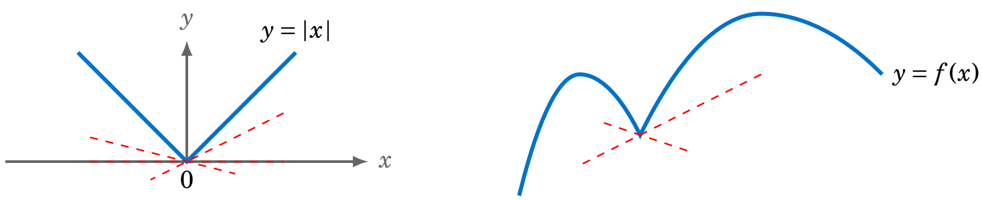
Figure 3.1.5 Secant line L_{PQ} approaching the tangent line T_P as $Q \to P$

As the point Q moves along the curve toward P, the line L_{PQ} approaches the tangent line T_P at the point P, provided the curve is smooth at P (i.e. $f'(x_0)$ exists). This is because the slope of L_{PQ} is $(f(w)-f(x_0))/(w-x_0)$, and so

$$\lim_{Q\to P} \left(\text{slope of } L_{PQ}\right) = \lim_{w\to x_0} \frac{f(w) - f(x_0)}{w - x_0} = f'(x_0) = \text{slope of } T_P$$

which means that as Q approaches P the "limit" of the secant line L_{PQ} has the same slope and goes through the same point P as the tangent line T_P.

- *Smooth curves have tangent lines, nonsmooth curves do not.*
  For example, think of the absolute value function $f(x) = |x|$. Its graph has a sharp edge at the point $(0,0)$, making it nonsmooth there, as shown in Figure 3.1.6(a) below. There is no real way to define a tangent line at $(0,0)$, because as mentioned in Section 1.2, the derivative of $f(x)$ does not exist at $x = 0$. The same holds true for curves with cusps, as in Figure 3.1.6(b).



(a) $f(x) = |x|$ **(b)** Curve with a cusp

Figure 3.1.6 Nonsmooth curves: no tangent line at the nonsmooth points

  Many lines go through the point of nonsmoothness, some of which are indicated by the dashed lines in the above figures, but none of them can be the tangent line. Sharp edges and cusps have to be "smoothed out" to have a tangent line.

As a point moves along a smooth curve, the corresponding tangent lines to the curve make varying angles with the positive x-axis—the angle is thus a function of x. Let $\phi = \phi(x)$ be the smallest angle that the tangent line L to a curve $y = f(x)$ makes with the positive x-axis, so that $-90° < \phi(x) < 90°$ for all x (see Figure 3.1.7).

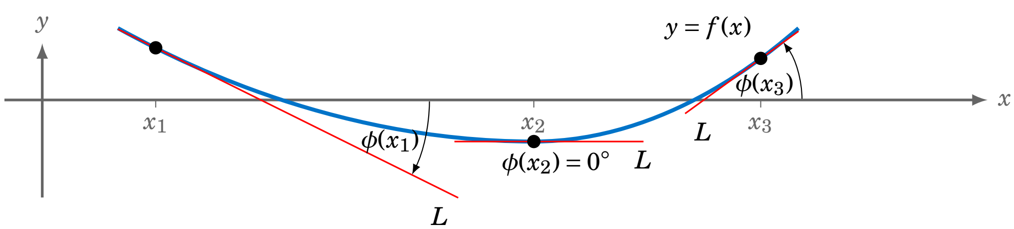
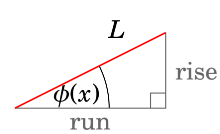


Figure 3.1.7 The angle $\phi(x)$ between the tangent line and positive x-axis

As Figure 3.1.7 shows, $-90° < \phi(x) < 0°$ when the tangent line L has negative slope, $0° < \phi(x) < 90°$ when L has positive slope, and $\phi(x) = 0°$ when L is horizontal (i.e. has zero slope). The slope of a line is usually defined as the rise divided by the run in a right triangle, as shown in the figure on the right. The figure shows as well that by definition of the tangent of an angle, $\tan\phi(x)$ also equals the rise (opposite) over run (adjacent). Thus, since the slope of L is $f'(x)$, this means that $\tan\phi(x) = f'(x)$. In other words:



The tangent line to a curve $y = f(x)$ makes an angle $\phi(x)$ with the positive x-axis, given by

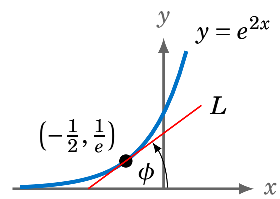
$$\phi(x) = \tan^{-1} f'(x)\,. \qquad (3.2)$$

Example 3.4

Find the angle ϕ that the tangent line to the curve $y = e^{2x}$ at $x = -\frac{1}{2}$ makes with the positive x-axis, such that $-90° < \phi < 90°$.

Solution: The angle is $\phi = \phi(-1/2) = \tan^{-1} f'(-1/2)$, where $f(x) = e^{2x}$. Since $f'(x) = 2e^{2x}$, then

$$\phi = \phi(-1/2) = \tan^{-1} f'(-1/2) = \tan^{-1} 2e^{-1} = \tan^{-1} 0.7358 = 36.3°\,.$$



The figure on the right shows the tangent line L to the curve at $x = -\frac{1}{2}$ and the angle ϕ.

You learned about perpendicular lines in elementary geometry. Figure 3.1.8 shows the natural way to define how a line N can be perpendicular to a *curve* $y = f(x)$ at a point P on the curve: the line is perpendicular to the tangent line of the curve at P. Call this line N the **normal line** to the curve at P. Since N and L are perpendicular, their slopes are negative reciprocals of each other (provided neither slope is 0). The equation of the normal line follows easily:

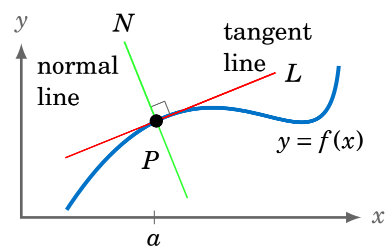


Figure 3.1.8 Normal line N

The equation of the normal line to a curve $y = f(x)$ at a point $P = (a, f(a))$ is

$$y - f(a) = -\frac{1}{f'(a)} \cdot (x - a) \quad \text{if } f'(a) \neq 0. \qquad (3.3)$$

If $f'(a) = 0$, then the normal line is vertical and is given by $x = a$.

Example 3.5

Find the normal line to the curve $y = x^2$ at $x = 1$. (Note: This is the curve from Example 3.1.)

Solution: The equation of the normal line is

$$y - f(a) = -\frac{1}{f'(a)} \cdot (x - a)$$

with $a = 1$, $f(x) = x^2$, and $f'(x) = 2x$. So $f(a) = 1$ and $f'(a) = 2$. Hence, the equation of the normal line is $y - 1 = -\frac{1}{2}(x - 1)$, or (in slope-intercept form) $y = -\frac{1}{2}x + \frac{3}{2}$.

Exercises

A

For Exercises 1-12, find the equation of the tangent line to the curve $y = f(x)$ at $x = a$.

1. $f(x) = x^2 + 1$; at $x = 2$

2. $f(x) = x^2 - 1$; at $x = 2$

3. $f(x) = -x^2 + 1$; at $x = 3$

4. $f(x) = 1$; at $x = -1$

5. $f(x) = 4x$; at $x = 1$

6. $f(x) = e^x$; at $x = 0$

7. $f(x) = x^2 - 3x + 7$; at $x = 2$

8. $f(x) = \dfrac{x+1}{x-1}$; at $x = 0$

9. $f(x) = (x^3 + 2x - 1)^3$; at $x = -1$

10. $f(x) = \tan x$; at $x = 0$

11. $f(x) = \sin 2x$; at $x = 0$

12. $f(x) = \sqrt{1 - x^2}$; at $x = 1/\sqrt{2}$

13. Find the equations of the tangent lines to the curve $y = x^3 - 2x^2 + 4x + 1$ which are parallel to the line $y = 3x - 5$.

14. Draw an example of a curve having a tangent line that intersects the curve at more than one point.

For Exercises 15-17, find the angle ϕ that the tangent line to the curve $y = f(x)$ at $x = a$ makes with the positive x-axis, such that $-90° < \phi < 90°$.

15. $f(x) = x^2$; at $x = 2$

16. $f(x) = \cos 2x$; at $x = \pi/6$

17. $f(x) = x^2 + 2x - 3$; at $x = -1$

18. Show that if $\phi(x)$ is the angle that the tangent line to a curve $y = f(x)$ makes with the positive x-axis such that $0° \le \phi(x) < 180°$, then

$$\phi(x) = \begin{cases} \cos^{-1}\left(\dfrac{1}{\sqrt{1 + (f'(x))^2}}\right) & \text{when } f'(x) \ge 0 \\ \cos^{-1}\left(\dfrac{-1}{\sqrt{1 + (f'(x))^2}}\right) & \text{when } f'(x) < 0. \end{cases}$$

(Hint: Draw a right triangle.)

For Exercises 19-21, find the angle ϕ that the tangent line to the curve $y = f(x)$ at $x = a$ makes with the positive x-axis, such that $0° \le \phi < 180°$.

19. $f(x) = x^2$; at $x = -1$

20. $f(x) = e^{-x}$; at $x = 1$

21. $f(x) = \ln 2x$; at $x = 10$

For Exercises 22-24, find the equation of the normal line to the curve $y = f(x)$ at $x = a$.

22. $f(x) = \sqrt{x}$; at $x = 4$

23. $f(x) = x^2 + 1$; at $x = 2$

24. $f(x) = x^2 - 7x + 4$; at $x = 3$

25. Find the equations of the normal lines to the curve $y = x^3 - 2x^2 - 11x + 3$ which have a slope of $-\frac{1}{4}$.

B

26. Show that the area of the triangle formed by the x-axis, the y-axis, and the tangent line to the curve $y = 1/x$ at any point P is constant (i.e. the area is the same for all P).

27. For a constant $a > 0$, let P be a point on the curve $y = ax^2$, and let Q be the point where the tangent line to the curve at P intersects the y-axis. Show that the x-axis bisects the line segment $\overline{PQ}$.

28. Let P be a point on the curve $y = 1/x$ in the first quadrant, and let Q be the point where the tangent line to the curve at P intersects the x-axis. Show that the triangle ΔPOQ is isosceles, where O is the origin.

3.2 Limits: Formal Definition

So far only the intuitive notion of a limit has been used, namely:

A real number L is the limit of $f(x)$ as x approaches a if the values of $f(x)$ can be made *arbitrarily* close to L by picking values of x *sufficiently* close to a.

That notion can be put in terms of a formal definition as follows:

Let L and a be real numbers. Then L is the **limit** of a function $f(x)$ as x approaches a, written as

$$\lim_{x \to a} f(x) = L ,$$

if for any given number $\epsilon > 0$, there exists a number $\delta > 0$, such that

$$|f(x) - L| < \epsilon \quad \text{whenever} \quad 0 < |x - a| < \delta .$$

A visual way of thinking of this definition is shown in Figure 3.2.1 below:

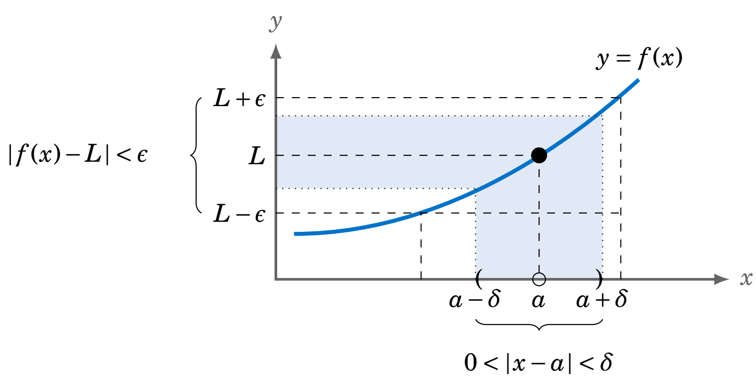


Figure 3.2.1 $\lim_{x \to a} f(x) = L$

Figure 3.2.1 says that for any interval around L on the y-axis, you will be able to find at least one small interval around $x = a$ (but excluding a) on the x-axis that the function $y = f(x)$ maps completely inside that interval on the y-axis. Choosing smaller intervals around L on the y-axis could force you to find smaller intervals around a on the x-axis.

In Figure 3.2.1, $f(x)$ is made arbitrarily close to L (within any distance $\epsilon > 0$) by picking x sufficiently close to a (within some distance $\delta > 0$). Since $0 < |x - a| < \delta$ means that $x = a$ itself is excluded, the solid dot at (a, L) could even be a hollow dot. That is, $f(a)$ does not have to equal L, or even be defined; $f(x)$ just needs to *approach* L as x approaches a. Thus–perhaps counter-intuitively—the existence of the limit does not actually depend on what happens *at* $x = a$ itself.

Example 3.6

Show that $\lim_{x\to a} x = a$ for any real number a.

Solution: Though the limit is obvious, the following "epsilon-delta" proof shows how to use the formal definition. The idea is to let $\epsilon > 0$ be given, then "work backward" from the inequality $|f(x) - L| < \epsilon$ to get an inequality of the form $|x - a| < \delta$, where $\delta > 0$ usually depends on ϵ. In this case the limit is $L = a$ and the function is $f(x) = x$, so since

$$|f(x) - a| < \epsilon \quad \Leftrightarrow \quad |x - a| < \epsilon \, ,$$

then choosing $\delta = \epsilon$ means that

$$0 < |x - a| < \delta \quad \Rightarrow \quad |x - a| < \epsilon \quad \Rightarrow \quad |f(x) - a| < \epsilon \, ,$$

which by definition means that $\lim_{x\to a} x = a$.

Calculating limits in this way might seem silly since—as in Example 3.6—it requires extra effort for a result that is obvious. The formal definition is used most often in proofs of general results and theorems. For example, the rules for limits—listed in Section 1.2—can be proved by using the formal definition.

Example 3.7

Suppose that $\lim_{x\to a} f(x)$ and $\lim_{x\to a} g(x)$ both exist. Show that

$$\lim_{x\to a} (f(x) + g(x)) = \left(\lim_{x\to a} f(x)\right) + \left(\lim_{x\to a} g(x)\right)$$

Solution: Let $\lim_{x\to a} f(x) = L_1$ and $\lim_{x\to a} g(x) = L_2$. The goal is to show that $\lim_{x\to a} (f(x) + g(x)) = L_1 + L_2$. So let $\epsilon > 0$. Then $\epsilon/2 > 0$, and so by definition there exist numbers $\delta_1 > 0$ and $\delta_2 > 0$ such that

$$\begin{aligned} 0 < |x - a| < \delta_1 \quad &\Rightarrow \quad |f(x) - L_1| < \epsilon/2 \text{ , and} \\ 0 < |x - a| < \delta_2 \quad &\Rightarrow \quad |g(x) - L_2| < \epsilon/2 \, . \end{aligned}$$

Now let $\delta = \min(\delta_1, \delta_2)$. Then $\delta > 0$ and

$$\begin{aligned} 0 < |x - a| < \delta \quad &\Rightarrow \quad 0 < |x - a| < \delta_1 \quad \text{and} \quad 0 < |x - a| < \delta_2 \\ &\Rightarrow \quad |f(x) - L_1| < \epsilon/2 \quad \text{and} \quad |g(x) - L_2| < \epsilon/2 \end{aligned}$$

Since $|A + B| \le |A| + |B|$ for all real numbers A and B, then

$$|f(x) + g(x) - (L_1 + L_2)| = |(f(x) - L_1) + (g(x) - L_2)| \le |f(x) - L_1| + |g(x) - L_2|$$

and thus

$$0 < |x - a| < \delta \quad \Rightarrow \quad |f(x) + g(x) - (L_1 + L_2)| < \epsilon/2 + \epsilon/2 = \epsilon$$

which by definition means that $\lim_{x\to a} (f(x) + g(x)) = L_1 + L_2$. ✓

The proofs of the other limit rules are similar.[1] In general, using the formal definition will not be necessary for evaluating limits of specific functions—in many cases a simple analysis of the function is all that is needed, often from its graph.

Example 3.8

Evaluate $\lim_{x \to 1} f(x)$ for the following function:

$$f(x) = \begin{cases} x & \text{if } x > 1 \\ 2 & \text{if } x = 1 \\ 1 & \text{if } x < 1 \end{cases}$$

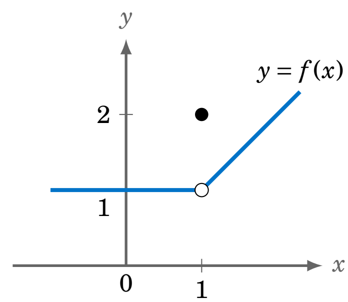
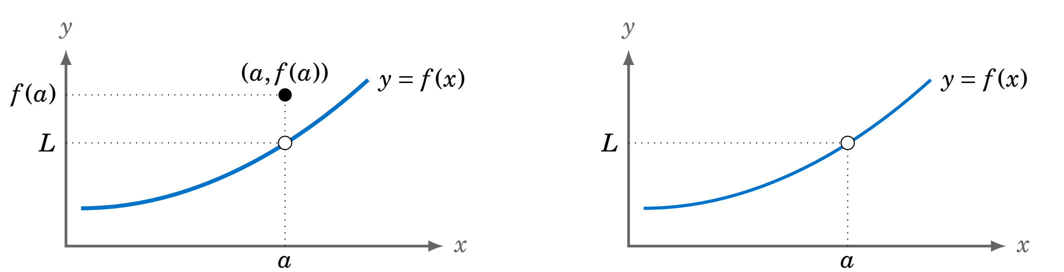


Figure 3.2.2

Solution: From the graph of $f(x)$ in Figure 3.2.2, it is clear that as x approaches 1 from the right (i.e. for $x > 1$) $f(x)$ approaches 1 along the line $y = x$, whereas as x approaches 1 from the left (i.e. for $x < 1$) $f(x)$ approaches 1 along the horizontal line $y = 1$. Thus, $\lim_{x \to 1} f(x) = 1$. Note that the limit did not depend on the value of $f(x)$ at $x = 1$.

As Example 3.8 shows, what matters for a limit is what happens to the value of $f(x)$ as x gets *near* a, not at $x = a$ itself. Figure 3.2.3(a) below shows how as x approaches a, $f(x)$ approaches a number different from $f(a)$. Figure 3.2.3(b) shows that $x = a$ does not even need to be in the domain of $f(x)$, i.e. $f(a)$ does not have to be defined. So it will not always be the case that $\lim_{x \to a} f(x) = f(a)$.



(a) $\lim_{x \to a} f(x) = L \neq f(a)$ **(b)** $\lim_{x \to a} f(x) = L$, $f(x)$ not defined at $x = a$

Figure 3.2.3 Excluding $x = a$ from $\lim_{x \to a} f(x)$

In Example 3.8 the direction in which x approached the number 1 did not affect the limit. But what if $f(x)$ had approached different values depending on how x approached 1? In that case the limit would not exist. The following definitions and notation for **one-sided limits** will make situations like that simpler to state.

[1]For example, see Section 2.2 in PROTTER, M.H. AND C.B. MORREY, *A First Course in Real Analysis*, New York: Springer-Verlag, 1977.

Call L the **right limit** of a function $f(x)$ as x approaches a, written as

$$\lim_{x\to a+} f(x) = L\,,$$

if $f(x)$ approaches L as x approaches a for values of x larger than a.
Call L the **left limit** of a function $f(x)$ as x approaches a, written as

$$\lim_{x\to a-} f(x) = L\,,$$

if $f(x)$ approaches L as x approaches a for values of x smaller than a.

The following statement follows immediately from the above definitions:

The limit of a function exists if and only if both its right limit and left limit exist and are equal:

$$\lim_{x\to a} f(x) = L \quad \Leftrightarrow \quad \lim_{x\to a-} f(x) = L = \lim_{x\to a+} f(x)$$

Example 3.9

Evaluate $\lim_{x\to 0-} f(x)$, $\lim_{x\to 0+} f(x)$, and $\lim_{x\to 0} f(x)$ for the following function:

$$f(x) = \begin{cases} x^2 & \text{if } x < 0 \\ 2-x & \text{if } x \ge 0 \end{cases}$$

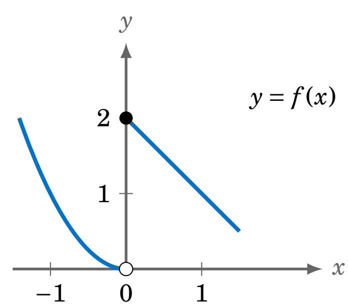


Figure 3.2.4

Solution: From the graph of $f(x)$ in Figure 3.2.4, it is clear that as x approaches 0 from the left (i.e. for $x<0$) $f(x)$ approaches 0 along the parabola $y = x^2$, whereas as x approaches 0 from the right (i.e. for $x>0$) $f(x)$ approaches 2 along the line $y = 2-x$. Hence, $\lim_{x\to 0-} f(x) = 0$ and $\lim_{x\to 0+} f(x) = 2$. Thus, $\lim_{x\to 0} f(x)$ does not exist since the left and right limits do not agree at $x = 0$.

Example 3.10

Evaluate $\lim_{x\to 0+} \sin\left(\frac{1}{x}\right)$.

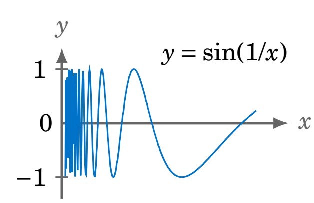


Figure 3.2.5

Solution: For $x>0$ the function $f(x) = \sin(1/x)$ is defined, and its graph is shown in Figure 3.2.5. As x approaches 0 from the right, $\sin(1/x)$ will be 1 for the numbers $x = 2/\pi$, $2/5\pi$, $2/9\pi$, $2/13\pi$, ... (which approach 0), and $\sin(1/x)$ will be -1 for the numbers $x = 2/3\pi$, $2/7\pi$, $2/11\pi$, $2/15\pi$, ... (which also approach 0). So as x approaches 0 from the right, $\sin(1/x)$ will oscillate between 1 and -1. Thus, $\lim_{x\to 0+} \sin\left(\frac{1}{x}\right)$ does not exist .

So far only *finite* limits have been considered, that is, $L = \lim_{x\to a} f(x)$ where L is a real (i.e. finite) number. Define an *infinite limit*, with $L = \infty$ or $-\infty$, as follows:

For a real number a, the limit of a function $f(x)$ **equals infinity** as x approaches a, written as

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

if $f(x)$ grows without bound as x approaches a, i.e. $f(x)$ can be made larger than any positive number by picking x sufficiently close to a:
For any given number $M > 0$, there exists a number $\delta > 0$, such that

$$f(x) > M \quad \text{whenever} \quad 0 < |x-a| < \delta .$$

For a real number a, the limit of a function $f(x)$ **equals negative infinity** as x approaches a, written as

$$\lim_{x\to a} f(x) = -\infty ,$$

if $f(x)$ grows negatively without bound as x approaches a, i.e. $f(x)$ can be made smaller than any negative number by picking x sufficiently close to a:
For any given number $M < 0$, there exists a number $\delta > 0$, such that

$$f(x) < M \quad \text{whenever} \quad 0 < |x-a| < \delta .$$

The above definitions can be modified accordingly for one-sided limits. If $\lim_{x\to a} f(x) = \infty$ or $\lim_{x\to a} f(x) = -\infty$, then the line $x = a$ is a **vertical asymptote** of $f(x)$, and $f(x)$ *approaches the line* $x = a$ *asymptotically*. The formal definitions are rarely needed.

Example 3.11

Evaluate $\lim_{x\to 0} \frac{1}{x}$.

Solution: For $x \neq 0$ the function $f(x) = \frac{1}{x}$ is defined, and its graph is shown in Figure 3.2.6. As x approaches 0 from the right, $1/x$ approaches ∞, that is,

$$\lim_{x\to 0+} \frac{1}{x} = \infty .$$

As x approaches 0 from the left, $1/x$ approaches $-\infty$, that is,

$$\lim_{x\to 0-} \frac{1}{x} = -\infty .$$

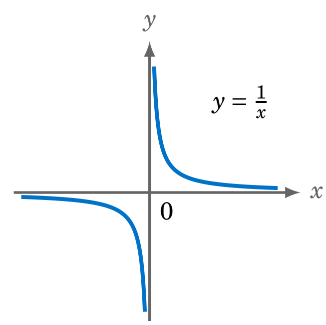


Figure 3.2.6

Since the right limit and the left limit are not equal, then $\lim_{x\to 0} \frac{1}{x}$ does not exist .
Note that the y-axis (i.e. the line $x = 0$) is a vertical asymptote for $f(x) = \frac{1}{x}$.

Example 3.12

Evaluate $\lim_{x\to 0} \frac{1}{x^2}$.

Solution: For $x \neq 0$ the function $f(x) = \frac{1}{x^2}$ is defined, and its graph is shown in Figure 3.2.7. As x approaches 0 from either the right or the left, $1/x^2$ approaches ∞, that is,

$$\lim_{x\to 0+} \frac{1}{x^2} = \infty = \lim_{x\to 0-} \frac{1}{x^2} .$$

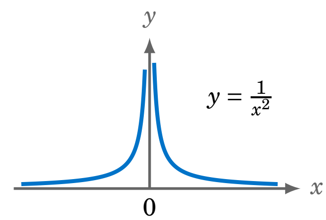


Figure 3.2.7

Since the right limit and the left limit both equal ∞, then $\lim_{x\to 0} \frac{1}{x^2} = \infty$.

Note that the y-axis (i.e. the line $x = 0$) is a vertical asymptote for $f(x) = \frac{1}{x^2}$.

In the limit $\lim_{x\to a} f(x)$ so far only real values of a have been considered. However, a could be either ∞ or $-\infty$:

For a real number L, the limit of a function $f(x)$ equals L as x approaches ∞, written as

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

if $f(x)$ can be made arbitrarily close to L for x sufficiently large and positive:
For any given number $\epsilon > 0$, there exists a number $N > 0$, such that

$$|f(x) - L| < \epsilon \quad \text{whenever} \quad x > N .$$

For a real number L, the limit of a function $f(x)$ equals L as x approaches $-\infty$, written as

$$\lim_{x\to -\infty} f(x) = L ,$$

if $f(x)$ can be made arbitrarily close to L for x sufficiently small and negative:
For any given number $\epsilon > 0$, there exists a number $N < 0$, such that

$$|f(x) - L| < \epsilon \quad \text{whenever} \quad x < N .$$

The above definitions can be modified accordingly for L replaced by either ∞ or $-\infty$.
One way to interpret the statement $\lim_{x\to\infty} f(x) = L$ is: *the long-term behavior of $f(x)$ is to approach a steady-state at L*. If $\lim_{x\to\infty} f(x) = L$ or $\lim_{x\to -\infty} f(x) = L$, then the line $y = L$ is a **horizontal asymptote** of $f(x)$, and $f(x)$ *approaches the line $y = L$ asymptotically*. Again, for most limits of specific functions, only the intuitive notions are needed.

Example 3.13

From Figures 3.2.6 and 3.2.7, it is clear that

$$\lim_{x\to\infty} \frac{1}{x} = 0 = \lim_{x\to-\infty} \frac{1}{x} \qquad \text{and} \qquad \lim_{x\to\infty} \frac{1}{x^2} = 0 = \lim_{x\to-\infty} \frac{1}{x^2}$$

Note that the x-axis (i.e. the line $y=0$) is a horizontal asymptote for $f(x)=\frac{1}{x}$ and $f(x)=\frac{1}{x^2}$.

Some limits are obvious, and you can use them to calculate other limits:

$$\lim_{x\to\infty} x^n = \begin{cases}\infty & \text{for any real } n>0\\ 0 & \text{for any real } n<0\end{cases} \qquad \lim_{x\to-\infty} x^n = \begin{cases}\infty & \text{for } n=2,4,6,8,\ldots\\ -\infty & \text{for } n=1,3,5,7,\ldots\end{cases}$$

$$\lim_{x\to\infty} e^x = \infty \qquad \lim_{x\to-\infty} e^x = 0 \qquad \lim_{x\to\infty} e^{-x} = 0 \qquad \lim_{x\to-\infty} e^{-x} = \infty$$

$$\lim_{x\to\infty} \ln x = \infty \qquad \lim_{x\to0+} \ln x = -\infty$$

A related notion is that of **Big O notation** (that is the capital letter O, not a zero):

Say that

$$f(x) = O(g(x)) \text{ as } x\to\infty,$$

spoken as "f is big O of g", if there exist positive numbers M and x_0 such that

$$|f(x)| \le M|g(x)| \text{ for all } x\ge x_0.$$

For example, obviously $2x^3 = O(x^3)$, by picking $M=2$, with x_0 any positive number. In general, $f(x)=O(g(x))$ means that f exhibits the same long-term behavior as g, up to a constant multiple. You can think of g as the more basic "type" of function that describes f, as far as long-term behavior.

Example 3.14

Show that $5x^4-2 = O(x^4)$.

Solution: First, recall from algebra that $|a+b| \le |a|+|b|$ for all real numbers a and b. Thus,

$$\left|5x^4-2\right| \le \left|5x^4\right| + |-2| = 5\left|x^4\right| + 2$$

for all x. So since $\left|x^4\right| = x^4 \ge 1$ for all $x\ge1$, then

$$\left|5x^4-2\right| \le 5\left|x^4\right| + 2 \le 5\left|x^4\right| + 2\left|x^4\right| \quad\Rightarrow\quad \left|5x^4-2\right| \le 7\left|x^4\right| \text{ for all } x\ge1,$$

which shows that $5x^4-2 = O(x^4)$, with $M=7$ and $x_0=1$.

Some limits need algebraic manipulation before they can be evaluated.

Example 3.15

Evaluate $\lim\limits_{x\to\infty} \left(\sqrt{x+1}-\sqrt{x}\right)$.

Solution: Note that both $\sqrt{x+1}$ and $\sqrt{x}$ approach ∞ as x goes to ∞, resulting in a limit of the form $\infty-\infty$. This is an example of an *indeterminate form*, which can equal anything (as will be discussed shortly); it does *not* have to equal 0 (i.e. the ∞'s do not necessarily "cancel out"). The trick here is to use the conjugate of $\sqrt{x+1}-\sqrt{x}$, so that

$$\lim_{x\to\infty} \left(\sqrt{x+1}-\sqrt{x}\right) = \lim_{x\to\infty} \left(\sqrt{x+1}-\sqrt{x}\right)\cdot\frac{\sqrt{x+1}+\sqrt{x}}{\sqrt{x+1}+\sqrt{x}} = \lim_{x\to\infty} \frac{(x+1)\,-\,x}{\sqrt{x+1}+\sqrt{x}} = \lim_{x\to\infty} \frac{1}{\sqrt{x+1}+\sqrt{x}} = 0$$

since the numerator is 1 and both terms in the sum in the denominator approach ∞ (i.e. $\frac{1}{\infty}=0$).

Some other indeterminate forms are ∞/∞, $0/0$ and $\infty\cdot 0$. How would you handle such limits? One way is to use **L'Hôpital's Rule**[2]; a simplified form is stated below:

L'Hôpital's Rule: If f and g are differentiable functions and

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\pm\infty}{\pm\infty} \quad\text{or}\quad \frac{0}{0}$$

then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}\,.$$

The number a can be real, ∞, or $-\infty$.

Example 3.16

Evaluate $\lim\limits_{x\to\infty} \dfrac{2x-1}{e^x}$.

Solution: This limit is of the form ∞/∞:

$$\begin{aligned}\lim_{x\to\infty} \frac{2x-1}{e^x} &\to \frac{\infty}{\infty}\\ &= \lim_{x\to\infty} \frac{2}{e^x} \quad \text{by L'Hôpital's Rule}\\ &= 0\end{aligned}$$

since the numerator is 2 and $e^x\to\infty$ as $x\to\infty$.

Note that one way of interpreting the limit being 0 is that e^x grows much faster than $2x-1$. In fact, using L'Hôpital's Rule it can be shown that e^x grows much faster than any polynomial, i.e. *exponential growth outstrips polynomial growth*.

[2]For a proof, see pp.89-91 in PROTTER, M.H. AND C.B. MORREY, *A First Course in Real Analysis*, New York: Springer-Verlag, 1977.

Example 3.17

Evaluate $\lim_{x\to\infty} \frac{x}{\ln x}$.

Solution: This limit is of the form ∞/∞:

$$\begin{aligned}\lim_{x\to\infty} \frac{x}{\ln x} &\rightarrow \frac{\infty}{\infty} \\ &= \lim_{x\to\infty} \frac{1}{\frac{1}{x}} \quad \text{by L'Hôpital's Rule} \\ &= \lim_{x\to\infty} x = \infty\end{aligned}$$

Note that one way of interpreting the limit being ∞ is that x grows much faster than $\ln x$. In fact, using L'Hôpital's Rule it can be shown that any polynomial grows much faster than $\ln x$, i.e. *polynomial growth outstrips logarithmic growth*.

Example 3.18

Evaluate $\lim_{x\to\infty} xe^{-2x}$.

Solution: This limit is of the form $\infty \cdot 0$, which can be converted to ∞/∞:

$$\begin{aligned}\lim_{x\to\infty} xe^{-2x} = \lim_{x\to\infty} \frac{x}{e^{2x}} &\rightarrow \frac{\infty}{\infty} \\ &= \lim_{x\to\infty} \frac{1}{2e^{2x}} \quad \text{by L'Hôpital's Rule} \\ &= 0\end{aligned}$$

Note that the limit is another consequence of exponential growth outstripping polynomial growth.

Example 3.19

Evaluate $\lim_{x\to\infty} \frac{2x^2 - 7x - 5}{3x^2 + 2x - 1}$.

Solution: This limit is of the form ∞/∞:

$$\begin{aligned}\lim_{x\to\infty} \frac{2x^2 - 7x - 5}{3x^2 + 2x - 1} &\rightarrow \frac{\infty}{\infty} \\ &= \lim_{x\to\infty} \frac{4x - 7}{6x + 2} \quad \text{by L'Hôpital's Rule} \\ &\rightarrow \frac{\infty}{\infty} \quad \text{, so use L'Hôpital's Rule again} \\ &= \frac{4}{6} = \frac{2}{3}\end{aligned}$$

Note that the limit ended up being the ratio of the leading coefficients of the polynomials in the numerator and denominator of the original limit. Note also that the lower-order terms (degree less than 2) ended up not mattering. In general you can always discard the lower-order terms when taking the limit of a ratio of polynomials (i.e. a *rational function*).

Example 3.20

Evaluate $\lim_{x\to 0} \dfrac{1-\cos x}{x}$.

Solution: This limit is of the form 0/0:

$$\begin{aligned} \lim_{x\to 0} \frac{1-\cos x}{x} &\rightarrow \frac{0}{0} \\ &= \lim_{x\to 0} \frac{\sin x}{1} \quad \text{by L'Hôpital's Rule} \\ &= \frac{\sin 0}{1} = 0 \end{aligned}$$

There is an intuitive justification for L'Hôpital's Rule: since the limit $\lim_{x\to a} \frac{f(x)}{g(x)}$ uses a ratio to compare how f changes relative to g as x approaches a, then it is really the rates of change of f and g—namely f' and g', respectively—that are being compared in that ratio, which is what L'Hôpital's Rule says.

The following result provides another way to calculate certain limits:

Squeeze Theorem: Suppose that for some functions f, g and h there is a number $x_0 \ge 0$ such that

$$g(x) \le f(x) \le h(x) \;\text{ for all } x > x_0$$

and that $\lim_{x\to\infty} g(x) = \lim_{x\to\infty} h(x) = L$. Then $\lim_{x\to\infty} f(x) = L$.

Similarly, if $g(x) \le f(x) \le h(x)$ for all $x \ne a$ in some interval I containing a, and if $\lim_{x\to a} g(x) = \lim_{x\to a} h(x) = L$, then $\lim_{x\to a} f(x) = L$.

Intuitively, the Squeeze Theorem says that if one function is "squeezed" between two functions approaching the same limit, then the function in the middle must also approach that limit. The theorem also applies to one-sided limits ($x \to a+$ or $x \to a-$).

Example 3.21

Evaluate $\lim_{x\to\infty} \dfrac{\sin x}{x}$.

Solution: Since $-1 \le \sin x \le 1$ for all x, then dividing all parts of those inequalities by $x > 0$ yields

$$-\frac{1}{x} \le \frac{\sin x}{x} \le \frac{1}{x} \;\text{ for all } x > 0 \quad\Rightarrow\quad \lim_{x\to\infty} \frac{\sin x}{x} = 0$$

by the Squeeze Theorem, since $\lim_{x\to\infty} \dfrac{-1}{x} = 0 = \lim_{x\to\infty} \dfrac{1}{x}$.

Exercises

A

For Exercises 1-18 evaluate the given limit.

1. $\lim_{x\to 2} \frac{x^2+3x-10}{x^2-x-2}$ **2.** $\lim_{x\to\infty} \frac{x^2+3x-10}{2x^2-x-2}$ **3.** $\lim_{x\to\infty} \frac{x^2+3x-10}{2x^3-x-2}$ **4.** $\lim_{x\to\infty} \frac{x^3+3x-10}{2x^2-x-2}$

5. $\lim_{x\to\pi/2} \frac{\cos x}{x-\pi/2}$ **6.** $\lim_{x\to\infty} \frac{x^2}{e^x}$ **7.** $\lim_{x\to-\infty} x^2e^x$ **8.** $\lim_{x\to 0+} \frac{\ln x}{e^{1/x}}$

9. $\lim_{x\to 0} \frac{\tan x - x}{x - \sin x}$ **10.** $\lim_{x\to 0} \frac{\sin 3x}{\sin 4x}$ **11.** $\lim_{x\to\infty} \frac{\cos x}{x}$ **12.** $\lim_{x\to 0} x \sin\left(\frac{1}{x}\right)$

13. $\lim_{x\to 0} \frac{\ln(1-x) - \sin^2 x}{1 - \cos^2 x}$ **14.** $\lim_{x\to 1} \left(\frac{1}{\ln x} - \frac{1}{x-1}\right)$ **15.** $\lim_{x\to\pi/2} (\sec x - \tan x)$

16. $\lim_{x\to 0+} \frac{e^{-1/x}}{x}$ **17.** $\lim_{x\to\infty} \left(\sqrt{x^2+4} - x\right)$ **18.** $\lim_{x\to 0} \frac{\cot x}{\csc x}$

B

19. The famous "twin paradox," a result of Einstein's special theory of relativity, says that if one of a pair of twins leaves the earth in a rocket traveling at a high speed, then he will be younger than his twin upon returning to earth.[3] This is due to the phenomenon of *time dilation*, which says that a clock moving with a speed v relative to a clock at rest in some inertial reference frame counts time slower relative to the clock at rest, by a factor of

$$\gamma = \frac{1}{\sqrt{1-\beta^2}} \,,$$

called the *Lorentz factor*, where $\beta = \frac{v}{c}$ is the fraction of the speed of light c at which the clock is moving ($c \approx 2.998 \times 10^8$ m/sec). Notice that $0 \le \beta < 1$ (why?). For example, a clock moving at half the speed of light, so that $\beta = 0.5$, would have $\gamma = 1.1547$, meaning that the clock runs about 15.47% slower than the clock on earth.

(a) Evaluate $\lim_{\beta\to 1-} \gamma$. What is the physical interpretation of this limit?

(b) Suppose an astronaut and his twin just turned 30 years old when the astronaut leaves earth on a high-speed journey through space. Upon returning to earth the astronaut is 35 and his twin is 70. At roughly what fraction of the speed of light must the astronaut have been traveling?

20. Show that $\lim_{x\to\infty} \frac{p(x)}{e^x} = 0$ for all polynomials $p(x)$ of degree $n \ge 1$ with a positive leading coefficient.

21. Show that $\lim_{x\to\infty} \frac{p(x)}{\ln x} = \infty$ for all polynomials $p(x)$ of degree $n \ge 1$ with a positive leading coefficient.

22. Show that $5x^3 + 6x^2 - 4x + 3 = O(x^3)$. **23.** Show that $\frac{2x^2+1}{x+1} = O(x)$. *(Hint: Consider $x \ge 1$)*

24. Call $h(x)$ an *infinitesimal function* as $x \to a$ if $\lim_{x\to a} h(x) = 0$. That is, an infinitesimal function approaches zero near some point. Prove the following result, where a and L are real numbers:

$$\lim_{x\to a} f(x) = L \quad \Leftrightarrow \quad f(x) = L + h(x) \text{ for all } x, \text{ where } h(x) \text{ is an infinitesimal function as } x \to a$$

[3]See p.154-159 in French, A.P., *Special Relativity*, Surrey, U.K.: Thomas Nelson & Sons Ltd., 1968.

3.3 Continuity

Recall from the previous section that a limit $\lim_{x\to a} f(x)$ can exist without being equal to $f(a)$, or with $f(a)$ not even being defined. Many functions encountered in applications, however, will meet those conditions, and they have a special name:

A function f is **continuous** at $x = a$ if

$$\lim_{x\to a} f(x) = f(a)\,. \tag{3.4}$$

A function is continuous on an interval I if it is continuous at every point in the interval. For a closed interval $I = [a,b]$, a function f is continuous on I if it is continuous on the open interval (a,b) and if $\lim_{x\to a+} f(x) = f(a)$ (i.e. f is **right continuous** at $x = a$) and $\lim_{x\to b-} f(x) = f(b)$ (i.e. f is **left continuous** at $x = b$). A function is **discontinuous** at a point if it is not continuous there. A continuous function is one that is continuous over its entire domain.

Equation (3.4) in the above definition implies that $f(a)$ is defined, i.e. $x = a$ is in the domain of f. Figure 3.3.1 below shows some examples of continuity and discontinuity:

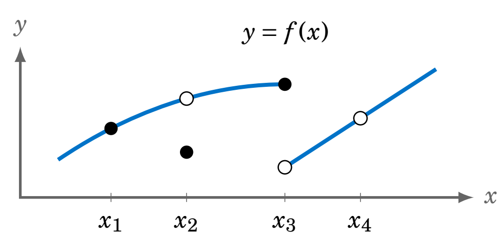


Figure 3.3.1 Continuous at x_1, discontinuous at x_2, x_3 and x_4

In the above figure, f is not continuous at $x = x_2$ because $\lim_{x\to x_2} f(x) \neq f(x_2)$; f is not continuous at $x = x_3$ because $\lim_{x\to x_3} f(x)$ does not exist (the right and left limits do not agree—f is said to have a **jump discontinuity** at $x = x_3$); and f is not continuous at $x = x_4$ because $f(x_4)$ is not defined. However, f is continuous at $x = x_1$.

A function is continuous if its graph is one unbroken piece over its entire domain. Polynomials, rational functions, trigonometric functions, exponential functions, and logarithmic functions are all continuous on their domains. For example, $\tan x$ is continuous over its domain, which is broken into disjoint intervals $(-\pi/2, \pi/2)$, $(\pi/2, 3\pi/2)$, $(3\pi/2, 5\pi/2)$, and so forth; the graph is unbroken on each of those intervals. However, $\tan x$ is not continuous over all of $\mathbb{R}$, since the function is not defined at all points in $\mathbb{R}$.

In the language of infinitesimals, a function f is continuous at $x = a$ if $f(a + dx) - f(a)$ is an infinitesimal for any infinitesimal dx. This definition is rarely used.

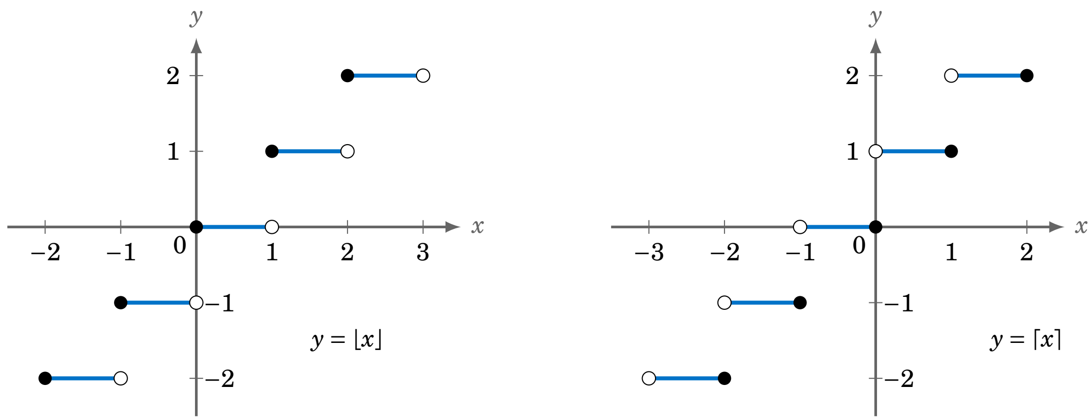
Physical examples of continuous functions are position, speed, velocity, acceleration, temperature, and pressure. Some discontinuous functions do arise in applications.

Example 3.22

The *floor function* $\lfloor x \rfloor$ is defined as

$$\lfloor x \rfloor \;=\; \text{the largest integer less than or equal to } x\;.$$

In other words, $\lfloor x \rfloor$ rounds a non-integer down to the previous integer, and integers stay the same. For example, $\lfloor 0.1 \rfloor = 0$, $\lfloor 0.9 \rfloor = 0$, $\lfloor 0 \rfloor = 0$, and $\lfloor -1.3 \rfloor = -2$. The graph of $\lfloor x \rfloor$ is shown in Figure 3.3.2(a).



(a) Floor function $\lfloor x \rfloor$ **(b)** Ceiling function $\lceil x \rceil$

Figure 3.3.2 Floor and ceiling functions

Similarly, the *ceiling function* $\lceil x \rceil$ is defined as

$$\lceil x \rceil \;=\; \text{the smallest integer greater than or equal to } x\;.$$

In other words, $\lceil x \rceil$ rounds a non-integer up to the next integer, and integers stay the same. For example, $\lceil 0.1 \rceil = 1$, $\lceil 0.9 \rceil = 1$, $\lceil 1 \rceil = 1$, and $\lceil -1.3 \rceil = -1$. The graph of $\lceil x \rceil$ is shown in Figure 3.3.2(b).

Clearly both $\lfloor x \rfloor$ and $\lceil x \rceil$ have jump discontinuities at the integers, but both are continuous at all non-integer values of x. Both functions are also examples of **step functions**, due to the staircase appearance of their graphs. Step functions are useful in situations when you want to model a quantity that takes only a discrete set of values. For example, in a car with a 4-gear transmission, $f(x)$ could be the gear the transmission has shifted to while the car travels at speed x. Up to a certain speed the car remains in first gear ($f = 1$) and then shifts to second gear ($f = 2$) after attaining that speed, then it remains in second gear until reaching another speed, upon which the car then shifts to third gear ($f = 3$), and so on. In general, discrete *changes in state* are often modeled with step functions.

Example 3.23

For an extreme case of discontinuity, consider the function

$$f(x) \;=\; \begin{cases} 0 & \text{if } x \text{ is rational} \\ 1 & \text{if } x \text{ is irrational} \end{cases}$$

This function is discontinuous at every value of x in $\mathbb{R}$, since within any positive distance δ of a real number x—no matter how small δ is—there will be an infinite number of both rational and irrational numbers. This is a property of $\mathbb{R}$. So the value of f will keep jumping between 0 and 1 no matter how close you get to x. In other words, for any number a in $\mathbb{R}$, $f(a)$ exists but it will never equal $\lim_{x\to a} f(x)$ because that limit will not exist.

By the various rules for limits, it is straightforward to show that sums, differences, constant multiples, products and quotients of continuous functions are continuous. Likewise a continuous function of a continuous function (i.e. a composition of continuous functions) is also continuous. In addition, a continuous function of a finite limit of a function can be passed inside the limit:

If f is a continuous function and $\lim\limits_{x\to a} g(x)$ exists and is finite, then:

$$f\left(\lim_{x\to a} g(x)\right) = \lim_{x\to a} f(g(x)) \tag{3.5}$$

The same relation holds for one-sided limits.

The above result is useful in evaluating the indeterminate forms 0^0, ∞^0, and 1^∞. The idea is to take the natural logarithm of the limit by passing the continuous function $\ln x$ inside the limit and evaluate the resulting limit.

Example 3.24

Evaluate $\lim\limits_{x\to 0+} x^x$.

Solution: This limit is of the form 0^0, so let $y = \lim\limits_{x\to 0+} x^x$ and then take the natural logarithm of y:

$$\begin{aligned}
\ln y &= \ln\left(\lim_{x\to 0+} x^x\right) \\
&= \lim_{x\to 0+} \ln x^x \quad \text{(pass the natural logarithm function inside the limit)} \\
&= \lim_{x\to 0+} x \ln x \quad \rightarrow \quad 0\cdot(-\infty) \\
&= \lim_{x\to 0+} \frac{\ln x}{1/x} \quad \rightarrow \quad \frac{-\infty}{\infty} \\
&= \lim_{x\to 0+} \frac{1/x}{-1/x^2} \quad \text{by L'Hôpital's Rule} \\
\ln y &= \lim_{x\to 0+} (-x) = 0
\end{aligned}$$

Thus, $\lim\limits_{x\to 0+} x^x = y = e^0 = 1$.

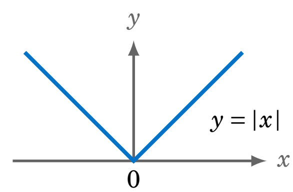
There is an important relationship between differentiability and continuity:

Every differentiable function is continuous.

Proof: If a function f is differentiable at $x = a$ then $f'(a) = \lim_{x \to a} \frac{f(x)-f(a)}{x-a}$ exists, so

$$\lim_{x \to a} (f(x) - f(a)) = \lim_{x \to a} (f(x) - f(a)) \cdot \frac{x-a}{x-a} = \lim_{x \to a} \frac{f(x)-f(a)}{x-a} \cdot \lim_{x \to a} (x-a) = f'(a) \cdot 0 = 0$$

which means that $\lim_{x \to a} f(x) = f(a)$, i.e. f is continuous at $x = a$. ✓



Note that the converse is not true. For example, the absolute value function $f(x) = |x|$ is continuous everywhere—its graph is unbroken, as shown in the picture on the right—but recall from Example 1.3 in Section 1.2 that it is not differentiable at $x = 0$. Continuous curves can have sharp edges and cusps, but differentiable curves cannot.

Two other important theorems[4] about continuous functions are:

Extreme Value Theorem: If f is a continuous function on a closed interval $[a, b]$ then f attains both a maximum value and a minimum value on that interval.

Intermediate Value Theorem: If f is a continuous function on a closed interval $[a, b]$ then f attains every value between $f(a)$ and $f(b)$.

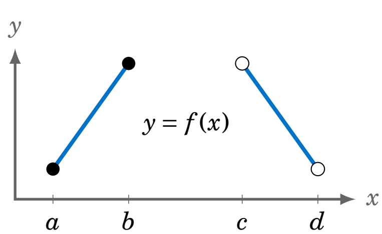


Figure 3.3.3 Extreme Value Theorem

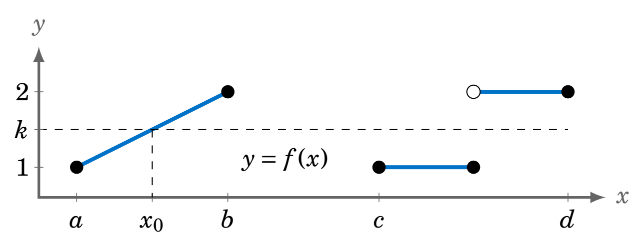


Figure 3.3.4 Intermediate Value Theorem

Figure 3.3.3 shows why a closed interval is required for the Extreme Value Theorem, as f attains neither a maximum nor minimum on the open interval (c, d). The Intermediate Value Theorem says that continuous functions cannot "skip over" intermediate values between two other function values. In Figure 3.3.4 the function f skips the value k between $f(c) = 1$ and $f(d) = 2$ because f is not continuous over all of $[c, d]$. On $[a, b]$ the value k is attained by f at $x = x_0$, i.e. $f(x_0) = k$, since f is continuous on $[a, b]$.

[4]The full proofs require some advanced results. See pp.97-98 and p.558 in TAYLOR, A.E. AND W.R. MANN, *Advanced Calculus*, 2nd ed., New York: John Wiley & Sons, Inc., 1972.

Example 3.25

Show that there is a solution to the equation $\cos x = x$.

Solution: Let $f(x) = \cos x - x$. Since f is continuous for all x, in particular it is continuous on $[0,1]$. So since $f(0) = 1 > 0$ and $f(1) = -0.459698 < 0$, then by the Intermediate Value Theorem there is a number c in the open interval $(0,1)$ such that $f(c) = 0$, since 0 is between the values $f(0)$ and $f(1)$. Hence, $\cos c - c = 0$, which means that $\cos c = c$. That is, $x = c$ is a solution of $\cos x = x$.

Note in the above example that the Intermediate Value Theorem does not tell you how to *find* the solution, just that the solution exists. To find the solution you can use the **bisection method**: divide the interval $[0,1]$ in half and apply the Intermediate Value Theorem to each half-interval to determine which one contains the solution; repeat this procedure on that half-interval, resulting in a smaller interval containing the solution, then repeat the procedure over and over, until you eventually obtain an interval so small that the midpoint of that interval can be taken as the solution. Listing 3.1 below shows one way of implementing the bisection method for Example 3.25 to find the root of $f(x) = \cos x - x$, using the Python programming language.

Listing 3.1 Bisection method in Python

```
1  import math
2
3  def f(x):
4      return math.cos(x) - x
5
6  def bisect(a, b):
7      midpt = (a+b)/2.0
8      tol = 1e-15
9      if b - a > tol:
10         val = f(midpt)
11         if val*f(a) < 0:
12             bisect(a, midpt)
13         elif val*f(b) < 0:
14             bisect(midpt, b)
15         else:
16             print("Root = %.13f" % (midpt))
17     else:
18         print("Root = %.13f" % (midpt))
19
20 bisect(0, 1)
```

Line 8 sets the *tolerance* to 10^{-15}: the program terminates upon reaching an interval whose length is smaller than that. The output is shown below:

```
Root = 0.7390851332152
```

This is the number obtained by taking the cosine of a number (in radians) repeatedly.

Exercises

A

For Exercises 1-18, indicate whether the given function $f(x)$ is continuous or discontinuous at the given value $x = a$ by comparing $f(a)$ with $\lim_{x \to a} f(x)$.

1. $f(x) = |x|$; at $x = 0$

2. $f(x) = |x-1|$; at $x = 0$

3. $f(x) = \lfloor x \rfloor$; at $x = 0$

4. $f(x) = \lfloor x \rfloor$; at $x = 0.3$

5. $f(x) = \lceil x \rceil$; at $x = 0$

6. $f(x) = \lceil x \rceil$; at $x = 0.5$

7. $f(x) = x - \lfloor x \rfloor$; at $x = 0$

8. $f(x) = x - \lfloor x \rfloor$; at $x = 1.1$

9. $f(x) = x - |x|$; at $x = 0$

10. $f(x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1 & \text{if } x > 0; \end{cases}$ at $x = 0$

11. $f(x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1 & \text{if } x > 0; \end{cases}$ at $x = 1$

12. $f(x) = \begin{cases} x^2 & \text{if } x \le 0, \\ 1 & \text{if } x > 0; \end{cases}$ at $x = 0$

13. $f(x) = \begin{cases} x+1 & \text{if } x \le 0, \\ 1 & \text{if } x > 0; \end{cases}$ at $x = 1$

14. $f(x) = \begin{cases} \sin(x^2) & \text{if } x \ne 0, \\ 0 & \text{if } x = 0; \end{cases}$ at $x = 0$

15. $f(x) = \begin{cases} \sin(1/x) & \text{if } x \ne 0, \\ 0 & \text{if } x = 0; \end{cases}$ at $x = 0$

16. $f(x) = \begin{cases} 0 & \text{if } x \text{ is rational}, \\ 1 & \text{if } x \text{ is irrational}; \end{cases}$ at $x = \sqrt{3}$

17. $f(x) = \begin{cases} 0 & \text{if } x \text{ is rational}, \\ x & \text{if } x \text{ is irrational}; \end{cases}$ at $x = 0$

18. $f(x) = \begin{cases} 0 & \text{if } x \text{ is rational}, \\ x & \text{if } x \text{ is irrational}; \end{cases}$ at $x = 1$

19. Evaluate $\lim_{x \to 0+} x^{x^2}$.

20. Evaluate $\lim_{x \to \infty} x^{1/x}$.

21. Evaluate $\lim_{x \to 0} (1-x)^{1/x}$.

22. If $f(x) = \dfrac{x^2 + x - 2}{x - 1}$ for $x \ne 1$, how should $f(1)$ be defined so that $f(x)$ is continuous at $x = 1$?

23. If $f(x) = 1/x$ for $x \ne 0$, is there a way to define $f(0)$ so that $f(x)$ is continuous for all x?

B

24. Can a function that is not continuous over a closed interval attain a maximum value and a minimum value in that interval? If so, then give an example; if not then explain why.

25. Show that there is a number x such that $x^5 - x = 3$.

26. Prove that $f(x) = x^8 + 3x^4 - 1$ has at least two distinct real roots.

27. Suppose that a function f is continuous on the interval $[0,3]$, f has no roots in $[0,3]$, and $f(1) = 1$. Prove that $f(x) > 0$ for all x in $[0,3]$.

28. Show that an object whose average speed is v_{avg} over the time interval $a \le t \le b$ will move with speed v_{avg} at some time t in $[a,b]$.

29. Let $f(x) = 1/(x-1)$. Then $f(0) = -1 < 0$ and $f(2) = 1 > 0$. Can you conclude by the Intermediate Value Theorem that $f(x)$ must be 0 for some x in $[0,2]$? Explain.

30. Show that if f' and f'' exist and are continuous at x then

$$f''(x) = \lim_{h \to 0} \frac{f(x) - 2f(x-h) + f(x-2h)}{h^2}.$$

31. Show that if f', f'' and f''' exist and are continuous at x then

$$f'''(x) = \lim_{h \to 0} \frac{f(x) - 3f(x-h) + 3f(x-2h) - f(x-3h)}{h^3}.$$

3.4 Implicit Differentiation

A function $y = f(x)$ is usually given by an explicit formula, such as $y = x^2$. It is then straightforward to find $\frac{dy}{dx}$ using the differentiation rules you have learned so far. But suppose instead that you were given merely an equation involving x and y, such as

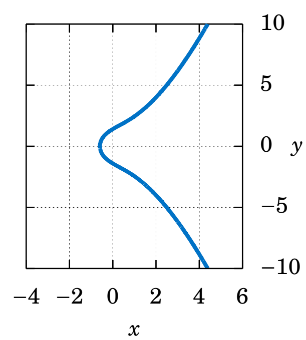
$$x^3 y^2 e^{\sin(xy)} = x^2 + xy + y^3 .$$

The set of points (x, y) satisfying this equation describes some sort of curve in the xy-plane, but it might not be possible to solve for y in terms of x—that is, there might not be an explicit formula for y as a function of the variable x. So in this case does the derivative $\frac{dy}{dx}$ even have any meaning, and if so then how would you find it?

It turns out that $\frac{dy}{dx}$ does make sense in such a case, because an equation involving x and y such as the one above *implicitly* defines y in terms of x in the following sense: as x varies so does y. Hence it should be possible to find the rate of change of y with respect to the variable x (i.e. $\frac{dy}{dx}$). To do so, take $\frac{d}{dx}$ of both sides of the equation, then assume that y really is a function of x so that you can use the Chain Rule to solve for $\frac{dy}{dx}$. The example below illustrates this procedure, called **implicit differentiation**.

Example 3.26

Find $\frac{dy}{dx}$ given the equation $x^3 + 3x + 2 = y^2$.



Solution: The above equation implicitly defines an *elliptic curve*, and its graph is shown on the right. This curve is not a function $y = f(x)$, since it violates the vertical line test, but y still varies with x. To find $\frac{dy}{dx}$ take $\frac{d}{dx}$ of both sides of the equation then solve for $\frac{dy}{dx}$:

$$\frac{d}{dx}(x^3 + 3x + 2) = \frac{d}{dx}(y^2)$$

$$3x^2 + 3 = 2y \cdot \frac{dy}{dx} \quad \text{by the Chain Rule, so}$$

$$\frac{dy}{dx} = \frac{3x^2 + 3}{2y}$$

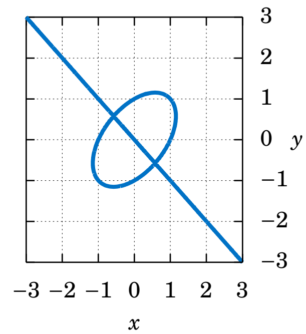
At first this might seem unsatisfying—or confusing—since $\frac{dy}{dx}$ is given in terms of both x and y. However, the derivative can still be evaluated at specific points (x, y) on the curve, i.e. any (x, y) satisfying the original equation. For example, it is easy to check that $(x, y) = (1, \sqrt{6})$ satisfies the equation $x^3 + 3x + 2 = y^2$, so $\frac{dy}{dx}(1, \sqrt{6}) = \frac{3(1)^2+3}{2\sqrt{6}} = \frac{\sqrt{6}}{2}$. Note that $\frac{dy}{dx}$ is not defined when $y = 0$.

Notice that taking the square root of both sides of the original equation does not result in an explicit formula for y, since $y = \pm\sqrt{x^3 + 3x + 2}$ defines *two* functions, not just one. The beauty of implicit differentiation is that the derivative $\frac{dy}{dx} = \frac{3x^2+3}{2y}$ calculated above gives you a single expression for the derivative of *both* those functions.

An *algebraic curve* is defined as the set of all points (x, y) satisfying a polynomial equation in the variables x and y, such as $x^2 - 3xy^4 + 1 = x^5 - y^2$. An *elliptic curve* is a special case of an algebraic curve, where the polynomial has the specific form $x^3 + ax + b = y^2$, such as the equation $x^3 + 3x + 2 = y^2$ from Example 3.26. Elliptic curves have certain properties that have found applications in cryptography.[5]

Example 3.27

Find $\frac{dy}{dx}$ given the equation $x + y = x^3 + y^3$.



Solution: The above equation implicitly defines an algebraic curve and its graph is shown on the right. To find $\frac{dy}{dx}$ take $\frac{d}{dx}$ of both sides of the equation then solve for $\frac{dy}{dx}$:

$$\frac{d}{dx}(x + y) = \frac{d}{dx}(x^3 + y^3)$$

$$1 + \frac{dy}{dx} = 3x^2 + 3y^2 \cdot \frac{dy}{dx} \quad \text{by the Chain Rule, so}$$

$$\frac{dy}{dx} = \frac{3x^2 - 1}{1 - 3y^2}$$

Notice that the curve consists of an oval shape (an ellipse, actually) with a line through it. In fact, that line is $y = -x$, as can be verified by replacing each instance of y in the equation $x + y = x^3 + y^3$ by $-x$ (resulting in the equation $0 = 0$). You might be wondering how $\frac{dy}{dx}$ is defined at the points where that line intersects the ellipse: is it the slope of the line $y = -x$ (i.e. -1), or is it the slope of the tangent line to the ellipse at those points (which would not equal -1)? This is discussed in the exercises.

The graph was created with the free open-source graphing program Gnuplot[6] using the following Gnuplot commands (which give an idea of how to plot implicit functions in general):

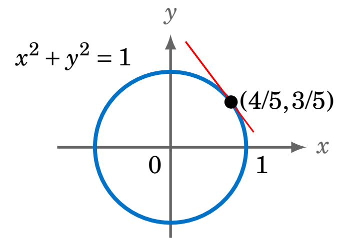
```
set size square
set view 0,0
set isosamples 500,500
set contour base
set cntrparam levels discrete 0
unset surface
set grid
unset key
unset ztics
set xlabel 'x'
set ylabel 'y'
f(x,y) = x + y - x**3 - y**3
splot [-3:3][-3:3] f(x,y) lw 3
```

[5] For example, see Section 12.2 in Buchmann, J.A., *Introduction to Cryptography*, New York: Springer-Verlag, 2001.

[6] See the documentation at `http://www.gnuplot.info/documentation.html`

Example 3.28

Find the tangent line to the curve $x^2 + y^2 = 1$ at the point $(4/5, 3/5)$.



Solution: This curve is the unit circle, shown in the picture on the right. First use implicit differentiation to find $\frac{dy}{dx}$:

$$\frac{d}{dx}(x^2 + y^2) = \frac{d}{dx}(1) \quad \Rightarrow \quad 2x + 2y \cdot \frac{dy}{dx} = 0 \quad \text{by the Chain Rule}$$

$$\Rightarrow \quad \frac{dy}{dx} = -\frac{x}{y}$$

The slope m of the tangent line to the curve at $(4/5, 3/5)$ is then $m = \frac{dy}{dx}(4/5, 3/5) = -\frac{4/5}{3/5} = -4/3$. Thus, the equation of the tangent line is $y - \frac{3}{5} = -\frac{4}{3}\left(x - \frac{4}{5}\right)$.

Exercises

A

For Exercises 1-9, use implicit differentiation to find $\frac{dy}{dx}$.

1. $x^3 y - 4xy^2 = y + x^2$

2. $xy = (x + y)^3$

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

4. $x^{2/3} + y^{2/3} = a^{2/3}$

5. $(x^2 - y^2)^2 = 2x^2 + y^2$

6. $\dfrac{x + y}{x - y} = x^2 + y^2$

7. $\cos(xy) = \sin(x^2 y^2)$

8. $x^3 - x = y^2$

9. $x^3 y^2 e^{\sin(xy)} = x^2 + xy + y^3$

10. In Example 3.28 is it possible to solve the equation $x^2 + y^2 = 1$ explicitly for y in terms of x? Explain.

11. In Example 3.28 what happens to the tangent line at the point $(1, 0)$? Why does this make sense geometrically?

12. Find the equation of the tangent line to the curve $x^3 + 3x^2 y + y^3 = 8$ at the point $(2, 0)$.

B

13. Find $\frac{d^2y}{dx^2}$ for the curve $x^2 + y^2 = 1$. You may use the results from Example 3.28.

14. Show that at every point (x_0, y_0) on the curve $y^2 = 4ax$, the equation of the tangent line to the curve is $yy_0 = 2a(x + x_0)$.

15. Show that at every point (x_0, y_0) on the ellipse $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$, the equation of the tangent line to the ellipse is $\dfrac{xx_0}{a^2} + \dfrac{yy_0}{b^2} = 1$.

16. Show that at every point (x_0, y_0) on the hyperbola $\dfrac{x^2}{a^2} - \dfrac{y^2}{b^2} = 1$, the equation of the tangent line to the hyperbola is $\dfrac{xx_0}{a^2} - \dfrac{yy_0}{b^2} = 1$.

17. Show that $\frac{dy}{dx}$ is not defined at the points of intersection of the line and ellipse described by the curve $x + y = x^3 + y^3$ from Example 3.27. *(Hint: Factor the equation $x + y = x^3 + y^3$.)*

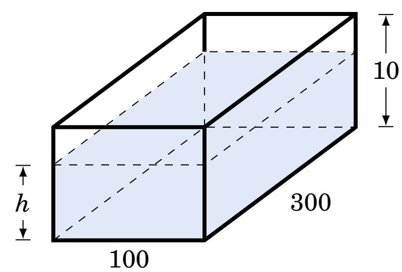
18. Show that the points $P = (2, 4)$ and $Q = (-31/64, -337/512)$ are on the elliptic curve $x^3 + 3x + 2 = y^2$ from Example 3.26, and that the tangent line to the curve at P also goes through Q.

3.5 Related Rates

If several quantities are related by an equation, then differentiating both sides of that equation with respect to a variable (usually t, representing time) produces a relation between the rates of change of those quantities. The known rates of change are then used in that relation to determine an unknown related rate.

Example 3.29

Suppose that water is being pumped into a rectangular pool at a rate of 60,000 cubic feet per minute. If the pool is 300 ft long, 100 ft wide, and 10 ft deep, how fast is the height of the water inside the pool changing?



Solution: Let V be the volume of the water in the pool. Since the volume of a rectangular solid is the product of the length, width, and height of the solid, then

$$V = (300)(100)h = 30000h \text{ ft}^3$$

where h is the height of the water, as in the picture on the right. Both V and h are functions of time t (measured in minutes), and $\frac{dV}{dt} = 60000 \text{ ft}^3/\text{min}$ was given. The goal is to find $\frac{dh}{dt}$. Since

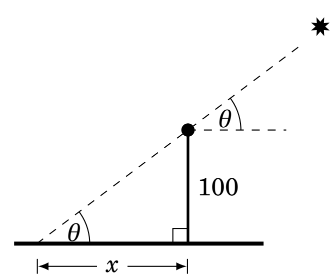
$$\frac{dV}{dt} = \frac{d}{dt}(30000h) = 30000\,\frac{dh}{dt}$$

then

$$\frac{dh}{dt} = \frac{1}{30000}\,\frac{dV}{dt} = \frac{1}{30000}\cdot 60000 = 2 \text{ ft/min.}$$

Example 3.30

Suppose that the angle of inclination from the top of a 100 ft pole to the sun is decreasing at a rate of 0.05 radians per minute. How fast is the length of the pole's shadow on the ground increasing when the angle of inclination is $\pi/6$ radians? You may assume that the pole is perpendicular to the ground.



Solution: Let θ be the angle of inclination and let x be the length of the shadow, as in the picture on the right. Both θ and x are functions of time t (measured in minutes), and $\frac{d\theta}{dt} = -0.05$ rad/min was given (the derivative is negative since θ is decreasing). The goal is to find $\frac{dx}{dt}$ when $\theta = \pi/6$, denoted by $\left.\frac{dx}{dt}\right|_{\theta=\pi/6}$ (the vertical bar means "evaluated at" the value of the subscript to the right of the bar). Since

$$x = 100\cot\theta \quad\Rightarrow\quad \frac{dx}{dt} = -100\csc^2\theta\cdot\frac{d\theta}{dt} = -100\csc^2\theta\cdot(-0.05) = 5\csc^2\theta$$

then

$$\left.\frac{dx}{dt}\right|_{\theta=\pi/6} = 5\csc^2(\pi/6) = 5\,(2)^2 = 20 \text{ ft/min.}$$

Example 3.31

The radius of a right circular cylinder is decreasing at the rate of 3 cm/min, while the height is increasing at the rate of 2 cm/min. Find the rate of change of the volume of the cylinder when the radius is 8 cm and the height is 6 cm.

Solution: Let r, h, and V be the radius, height, and volume, respectively, of the cylinder. Then $V = \pi r^2 h$. Since $\frac{dr}{dt} = -3$ cm/min and $\frac{dh}{dt} = 2$ cm/min, then by the Product Rule:

$$\frac{dV}{dt} = \frac{d}{dt}(\pi r^2 h) = \left(2\pi r \cdot \frac{dr}{dt}\right) h + \pi r^2 \cdot \frac{dh}{dt} \quad \Rightarrow \quad \left.\frac{dV}{dt}\right|_{\substack{r=8\\h=6}} = 2\pi(8)(-3)(6) + \pi(8^2)(2) = -160\pi \ \frac{\text{cm}^3}{\text{min}}$$

Exercises

A

1. A stone is dropped into still water. If the radius of the circular outer ripple increases at the rate of 4 ft/s, how fast is the area of the circle of disturbed water increasing when the radius is 10 ft?

2. The radius of a sphere decreases at a rate of 3 mm/hr. Determine how fast the volume and surface area of the sphere are changing when the radius is 5 mm.

3. A kite 80 ft above level ground moves horizontally at a rate of 4 ft/s away from the person flying it. How fast is the string being released at the instant when 100 ft of string have been released?

4. A 10-ft ladder is leaning against a wall on level ground. If the bottom of the ladder is dragged away from the wall at the rate of 5 ft/s, how fast will the top of the ladder descend at the instant when it is 8 ft from the ground?

5. A person 6 ft tall is walking at a rate of 6 ft/s away from a light which is 15 ft above the ground. At what rate is the end of the person's shadow moving along the ground away from the light?

6. An object moves along the curve $y = x^3$ in the xy-plane. At what points on the curve are the x and y coordinates of the object changing at the same rate?

7. The radius of a right circular cone is decreasing at the rate of 4 cm/min, while the height is increasing at the rate of 3 cm/min. Find the rate of change of the volume of the cone when the radius is 6 cm and the height is 7 cm.

8. Two boats leave the same dock at the same time, one goes north at 25 mph and the other goes east at 30 mph. How fast is the distance between the boats changing when they are 100 miles apart?

9. Repeat Exercise 8 with the angle between the boats being 110°.

B

10. An angle θ changes with time. For what values of θ do $\sin\theta$ and $\tan\theta$ change at the same rate?

11. Repeat Example 3.30 but with the ground making a 100° angle with the pole to the left of the pole.

C

12. An upright cylindrical tank full of water is tipped over at a constant angular speed. Assume that the height of the tank is at least twice its radius. Show that at the instant the tank has been tipped 45°, water is leaving the tank twice as fast as it did at the instant the tank was first tipped. *(Hint: Think of how the water looks inside the tank as it is being tipped.)*

3.6 Differentials

An *ideal gas* satisfies the equation $PV = RT$, where R is a constant and P, V, and T are the pressure, volume per mole, and temperature, respectively, of the gas. It will be proved that

$$\frac{dP}{P} + \frac{dV}{V} = \frac{dT}{T} \,. \tag{3.6}$$

Recall that dP, dV, and dT represent infinitesimal changes in the quantities P, V, and T, respectively. Notice that none of the quotients in Equation (3.6) have an infinitesimal in the denominator. For example, dP is divided not by dx or dt, as it would be in a derivative such as $\frac{dP}{dx}$ or $\frac{dP}{dt}$. Instead it is divided by P, which is not an infinitesimal. So Equation (3.6) is an equation that relates infinitesimals themselves, i.e. infinitesimal changes, not infinitesimal *rates* of change. This is, in fact, how many physical laws are stated, for reasons that will be discussed shortly.

Though infinitesimals have been used throughout this text, many calculus textbooks[7] do not even mention them, instead preferring to call them *differentials*.[8] For compatibility, the definition is given here:

For a differentiable function $f(x)$, the **differential** of $f(x)$ is

$$df = f'(x)\,dx \tag{3.7}$$

where dx is an infinitesimal change in x.

Note that this is identical to Equation (1.9) in Section 1.3.

Example 3.32

Find the differential df of $f(x) = x^3$.

Solution: By definition,

$$df = f'(x)\,dx = 3x^2\,dx$$

Equivalently, this can be written as

$$d(x^3) = 3x^2\,dx \,,$$

which is often the way it would appear in textbooks in the sciences.

All the rules for derivatives (e.g. sum rule, product rule) apply to differentials, and can be proved simply by multiplying the corresponding derivative rule by dx on both sides of the equation:

[7]More accurately, many *current* calculus textbooks never mention them. Calculus texts up through the 1930s or so not only mentioned infinitesimals but used them extensively, even to the point of the texts themselves having titles such as *Introduction to Infinitesimal Calculus*.

[8]Though often in an unclear and sometimes confusing and misleading manner, as will be seen later in this section.

Let f and g be differentiable functions, and let c be a constant. Then:

(a) $d(c) = 0$

(b) $d(cf) = c\,df$ (Constant Multiple Rule)

(c) $d(f+g) = df + dg$ (Sum Rule)

(d) $d(f-g) = df - dg$ (Difference Rule)

(e) $d(fg) = f\,dg + g\,df$ (Product Rule)

(f) $d\left(\dfrac{f}{g}\right) = \dfrac{g\,df - f\,dg}{g^2}$ (Quotient Rule)

(g) $d(f^n) = nf^{n-1}\,df$ (Power Rule)

(h) $d(f(g)) = \dfrac{df}{dg}\,dg$ (Chain Rule)

For example, to prove (e), multiply both sides of the usual Product Rule by dx so that

$$\frac{d(fg)}{dx} = f\,\frac{dg}{dx} + g\,\frac{df}{dx} \quad\Rightarrow\quad d(fg) = \cancel{dx}\left(f\,\frac{dg}{\cancel{dx}} + g\,\frac{df}{\cancel{dx}}\right)$$
$$\Rightarrow\quad d(fg) = f\,dg + g\,df \quad \checkmark$$

since the dx terms all cancel. The proofs of the other rules are similar.

The differential version of the ideal gas law in Equation (3.6)

$$\frac{dP}{P} + \frac{dV}{V} = \frac{dT}{T}$$

can now be proved by taking the differential of both sides of the equation $PV = RT$:

$$\begin{aligned}
d(PV) &= d(RT) = R\cdot dT && \text{by the Constant Multiple Rule}\\
V\,dP + P\,dV &= \frac{PV}{T}\,dT && \text{by the Product Rule and since } R = \frac{PV}{T}\\
\frac{V\,dP}{PV} + \frac{P\,dV}{PV} &= \frac{dT}{T} && \text{after dividing both sides by } PV\\
\frac{dP}{P} + \frac{dV}{V} &= \frac{dT}{T} && \checkmark
\end{aligned}$$

Notice that $\frac{dP}{P}$, $\frac{dV}{V}$ and $\frac{dT}{T}$ represent the *relative* infinitesimal changes in P, V, and T, respectively. The differential formulation is useful for finding one relative infinitesimal change when the other two are known.

Example 3.33

Suppose that M is the total mass of a rocket and its unburnt fuel at any time t (so M is a function of t). Over an infinitesimal time dt a mass dm of fuel is burnt and the gas byproducts are expelled out the rear of the rocket at a velocity v_E relative to the rocket. Using the law of conservation of momentum over the interval dt, show that

$$v_E\,dm \;=\; M\,dv$$

where m and v are the mass of burnt fuel and the velocity of the rocket, respectively, at the beginning of the time dt.

Solution: Momentum is defined as mass times velocity. The momentum of the rocket at the beginning of the time dt is thus Mv. At the end of the time dt, the momentum of the rocket consists of two parts, namely the momentum of the rocket and its remaining unburnt fuel, which is

$$((\text{mass before } dt) - (\text{increase in burnt fuel})) \times ((\text{velocity before } dt) + (\text{increase in velocity})) \tag{3.8}$$

$$(M - dm)(v + dv) \tag{3.9}$$

and the momentum of the fuel that was burnt and expelled out the rear, which is

$$(v - v_E)\,dm\;.$$

So by conservation of momentum,

$$\begin{aligned} Mv \;&=\; (M - dm)(v + dv) \;+\; (v - v_E)\,dm \\ Mv \;&=\; Mv \;-\; v\,dm \;+\; M\,dv \;-\; (dm)(dv) \;+\; v\,dm \;-\; v_E\,dm, \quad \text{so} \\ v_E\,dm \;&=\; M\,dv \;-\; (dm)(dv) \;=\; M\,dv \end{aligned}$$

since $(dm)(dv) = (m'(t)\,dt)(v'(t)\,dt) \;=\; m'(t)v'(t)(dt)^2 \;=\; m'(t)v'(t)\cdot 0 = 0$.

Dividing both sides of $v_E\,dm = M\,dv$ by dt yields the equation

$$M\,\dot{v} \;=\; \dot{m}\,v_E$$

using the dot notation—mentioned in Section 1.3—for the derivative with respect to the time variable t, which is still popular with physicists. Since $\dot{v}$ is just acceleration a, this formulation is the classic equation for the acceleration of a rocket.[9]

Letting f be the natural logarithm function and letting $g = u$ in the differential version of the Chain Rule yields the following useful result:

$$d(\ln u) \;=\; \frac{du}{u}$$

This is often used in a differential version of the technique of logarithmic differentiation discussed in Section 2.3.

[9] For other formulations see Chapter 1 in ROSSER, J.B., R.R. NEWTON AND G.L. GROSS, *Mathematical Theory of Rocket Flight*, New York: McGraw-Hill Book Company, Inc., 1947.

Example 3.34

Prove the relation $\frac{dP}{P} + \frac{dV}{V} = \frac{dT}{T}$ using logarithmic differentiation.

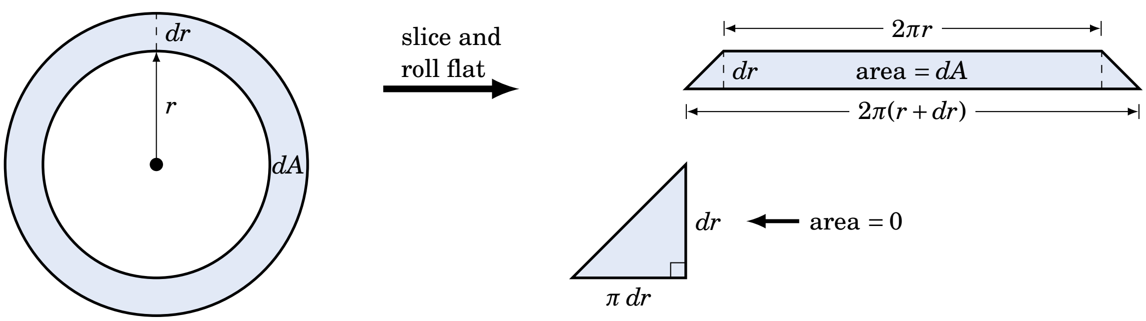
Solution: Take the natural logarithm and then the differential of both sides of the equation $PV = RT$:

$$\begin{aligned}\ln(PV) = \ln(RT) \quad &\Rightarrow \quad \ln P + \ln V = \ln R + \ln T \\ &\Rightarrow \quad d(\ln P + \ln V) = d(\ln R + \ln T) \\ &\Rightarrow \quad \frac{dP}{P} + \frac{dV}{V} = 0 + \frac{dT}{T} = \frac{dT}{T} \quad \text{(since } \ln R \text{ is a constant)}\end{aligned}$$

Example 3.35

The derivative of the area πr^2 of a circle of radius r, as a function of r, equals its circumference $2\pi r$. Use the notion of a differential as an infinitesimal change to explain why this makes sense geometrically.

Solution: Let $A = \pi r^2$ be the area of a circle of varying radius r. Then $A'(r) = 2\pi r$, which is equivalent to saying $dA = 2\pi r\,dr$. To see why this makes sense geometrically, imagine increasing the radius by dr, as in the picture below on the left. This increases the area A of the circle to $A + dA$, with dA the infinitesimal area of the shaded ring in the picture.



Slice that ring along the dashed line then roll it flat, yielding a trapezoid with height dr, top length $2\pi r$ (from the circumference of the inner circle of the ring), and bottom length $2\pi(r + dr)$ (from the circumference of the outer circle of the ring), as shown in the picture above on the right. The triangular edges of the trapezoid contribute nothing to the area of the trapezoid, since (by the Microstraightness Property) the hypotenuse of each is indeed a straight line, so each is a right triangle with height dr and (by symmetry) base πdr, thus having area $\frac{1}{2}\pi(dr)^2 = 0$. Hence the entire area dA of the trapezoid comes from the rectangular portion of height dr and base $2\pi r$, which means $dA = 2\pi r\,dr$, as expected.

The above example answers the question of whether it is a happy coincidence that the derivative of a circle's area turns out to be the circle's circumference—no, it is not! Some other such cases (e.g. the derivative of a sphere's volume is its surface area) are left to the exercises. Note that a similar "coincidence" does *not* occur for a square: if x is the length of each side then the area is x^2, but the derivative of x^2 is $2x$, which is *not* the perimeter of the square (i.e. $4x$). Why does this not follow the same pattern as the circle? Think about a key difference in the shape of a square in comparison to a circle, keeping differentiability in mind.

There are many benefits to using differentials—i.e. infinitesimals—in calculus.[10] For example, recall Example 3.31 in Section 3.5 on related rates, where the volume V of a right circular cylinder with radius r and height h changes with time t as

$$\frac{dV}{dt} = \left(2\pi r \cdot \frac{dr}{dt}\right) h + \pi r^2 \cdot \frac{dh}{dt}\,.$$

The above equation forces you to consider only the derivative with respect to the time variable t. What if you wanted to see the rates of change with respect to another variable, such as r, h, or some other quantity? In that case using the differential version of the above equation, namely

$$dV = 2\pi r h\, dr + \pi r^2\, dh$$

provides more flexibility—you are free to divide both sides by any differential, not just by dt. Many related rates problems would likely benefit from this approach.

Present-day calculus textbooks confuse the notion of a differential (infinitesimal) dx with the idea of a small but real value Δx. The two are *not* the same. An infinitesimal is *not* a real number and *cannot* be assigned a real value, no matter how small; Δx *can* be assigned real values. Using dx and Δx interchangeably is a source of much confusion for students (likewise for dy and Δy). This confusion rears its head in exercises involving the linear approximation of a curve by its tangent line near a point x_0, namely $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$ when $x - x_0$ is "small" (e.g. $\sqrt{63} \approx 7.9375$, by using $f(x) = \sqrt{x}$, $x = 63$, $x_0 = 64$, and $x - x_o = \Delta x = -1$). Such exercises have nothing to do with differentials, not to mention having dubious value nowadays. They are remnants of a bygone era, before the advent of modern computing obviated the need for such (generally) poor approximations.

Exercises

A

1. Find the differential df of $f(x) = x^2 - 2x + 5$.

2. Find the differential df of $f(x) = \sin^2(x^2)$.

3. Show that $d\left(\tan^{-1}(y/x)\right) = \dfrac{x\,dy - y\,dx}{x^2 + y^2}$

4. Given $y^2 - xy + 2x^2 = 3$, find dy.

5. The *elasticity* of a function $y = f(x)$ is $E(y) = \dfrac{x}{y} \cdot \dfrac{dy}{dx}$. Show that $E(y) = \dfrac{d(\ln y)}{d(\ln x)}$.

6. Prove the differential version of the Quotient Rule:

$$d\left(\frac{f}{g}\right) = \frac{g\,df - f\,dg}{g^2}$$

[10]For an excellent overview on this subject, see DRAY, T. AND C.A. MANOGUE, Putting Differentials Back into Calculus, *College Math. J.* **41** (2010), 90-100. Some of the material in this section is indebted to that paper, which is available at `http://www.math.oregonstate.edu/bridge/papers/differentials.pdf`

7. Let $y = cu^n$, where c and n are constants. Show that

$$\frac{dy}{y} = n\,\frac{du}{u}\;.$$

8. Obviously the derivative of the constant π^2 is not 2π. But is $d(\pi^2) = 2\pi\,d(\pi)$ true? Explain.

B

9. The *continuity relation* for an ideal gas is

$$\frac{PM}{\sqrt{T}} = \text{constant}$$

where P and T are the pressure and temperature, respectively, of the gas, and M is the *Mach number*. Show that

$$\frac{dP}{P} + \frac{dM}{M} = \frac{dT}{2T}\;.$$

10. For an ideal gas, satisfying the equation $PV = RT$ as before, the *Gibbs energy* G is defined as $G = H - TS$, where H and S are the *enthalpy* and *entropy*, respectively, of the gas.

(a) Show that

$$d\left(\frac{G}{RT}\right) = \frac{1}{RT}\,dG - \frac{G}{RT^2}\,dT\;.$$

(b) One of the *fundamental property relations* for an ideal gas (which you do not need to prove) is

$$dG = V\,dP - S\,dT\;.$$

Use this and part (a) to show that

$$d\left(\frac{G}{RT}\right) = \frac{V}{RT}\,dP - \frac{H}{RT^2}\,dT\;.$$

11. The derivative of the volume $\pi r^2 h$ of a right circular cylinder of radius r and height h, as a function of r, equals its lateral surface area $2\pi rh$. Use the notion of a differential as an infinitesimal change to explain why this makes sense geometrically.

12. The derivative of the volume $\frac{4\pi}{3}r^3$ of a sphere of radius r, as a function of r, equals its surface area $4\pi r^2$. Use the notion of a differential as an infinitesimal change to explain why this makes sense geometrically.

13. In *quantum calculus* the *q-differential* of a function $f(x)$ is

$$d_q f(x) = f(qx) - f(x)\,,$$

and the *q-derivative* of $f(x)$ is

$$D_q f(x) = \frac{d_q f(x)}{d_q x} = \frac{f(qx) - f(x)}{qx - x} = \frac{f(qx) - f(x)}{(q-1)x}\;.$$

(a) Show that for all positive integers n,

$$D_q\left(x^n\right) = [n]x^{n-1}\,,$$

where $[n] = 1 + q + q^2 + \cdots + q^{n-1}$.

(b) Use part (a) to show that for all positive integers n,

$$\lim_{q\to 1} D_q\left(x^n\right) = \frac{d}{dx}\left(x^n\right)\;.$$

CHAPTER 4

Applications of Derivatives

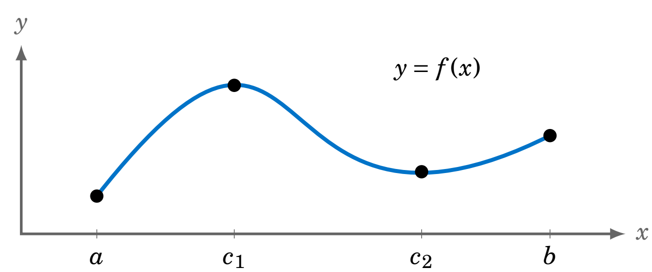
4.1 Optimization

Many physical problems involve *optimization*: finding either a maximum or minimum value of some quantity. Optimization problems often have a *constraint* involving two variables which allows you to rewrite the *objective function*—the function to optimize—as a function of a single variable: use the constraint to solve for one variable in terms of another, then substitute that expression into the objective function.

First, the intuitive notions of maximum and minimum need clarifying.

A function f has a **global maximum** at $x = c$ if $f(c) \ge f(x)$ for all x in the domain of f. Similarly, f has a **global minimum** at $x = c$ if $f(c) \le f(x)$ for all x in the domain of f. Say that f has a **local maximum** at $x = c$ if $f(c) \ge f(x)$ for all x "near" c, i.e. for all x such that $|x - c| < \delta$ for some number $\delta > 0$. Likewise, f has a **local minimum** at $x = c$ if $f(c) \le f(x)$ for all x such that $|x - c| < \delta$ for some number $\delta > 0$.

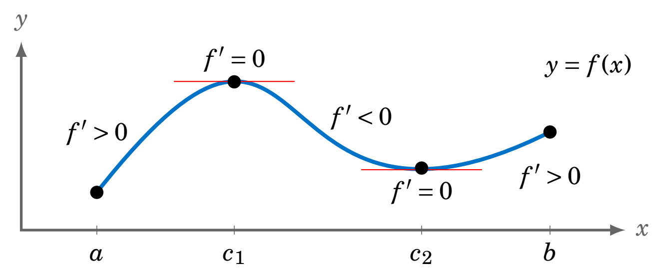
In other words, a global maximum is the largest value everywhere ("globally"), whereas a local maximum is only the largest value "locally." Likewise for a global vs local minimum. The picture below illustrates the differences.



In the picture, on the interval $[a, b]$ the function f has a global minimum at $x = a$, a global maximum at $x = c_1$, a local minimum at $x = c_2$, and a local maximum at $x = b$.

Every global maximum [minimum] is a local maximum [minimum], but not vice versa. In physical applications global maxima or minima[1] are the primary interest. The Extreme Value Theorem in Section 3.3 guarantees the existence of at least one global maximum and at least one global minimum for continuous functions defined on *closed* intervals (i.e. intervals of the form $[a,b]$). All the functions under consideration here will be differentiable, and hence continuous. So the only issues will be how to find the global maxima or minima, and how to handle intervals that are not closed.

Consider again the picture from the previous page, this time looking at how the derivative f' changes over $[a,b]$. Intuitively it is obvious that near an internal maximum (i.e. in the *open* interval (a,b)) such as at $x = c_1$, the function should increase before that point and then decrease after that point. That means that $f'(x) > 0$ before $x = c_1$ and $f'(x) < 0$ after the "turning point" $x = c_1$, as shown below.



Assuming that f' is continuous (which will be the case for all the functions in this section), then this means that $f' = 0$ at $x = c_1$, that is, $f'(c_1) = 0$. Similarly, near the internal minimum at $x = c_2$, $f'(x) < 0$ before $x = c_2$ and $f'(x) > 0$ after $x = c_2$, so that $f'(c_2) = 0$. Points at which the derivative is zero are called **critical points** (or **stationary points**) of the function. So $x = c_1$ and $x = c_2$ are critical points of f.

Note in the picture that f' goes from positive to zero to negative around $x = c_1$, so that f' is decreasing around $x = c_1$, i.e. $f'' = (f')' < 0$. Similarly, f' is increasing around $x = c_2$, i.e. $f'' > 0$. This leads to the following test for local maxima and minima:[2]

Second Derivative Test: Let $x = c$ be a critical point of f (i.e $f'(c) = 0$). Then:

(a) If $f''(c) > 0$ then f has a local minimum at $x = c$.

(b) If $f''(c) < 0$ then f has a local maximum at $x = c$.

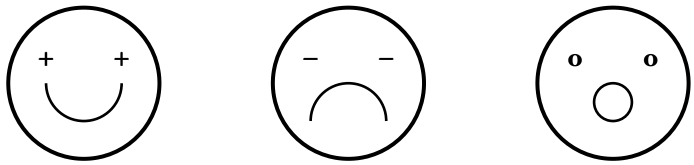
(c) If $f''(c) = 0$ then the test fails.

To see why the test fails when $f''(c) = 0$, consider $f(x) = x^3$: $f'(0) = 0$ and $f''(0) = 0$, yet $x = 0$ is neither a local minimum nor maximum in any open interval containing $x = 0$. Section 4.2 will present an alternative for when the Second Derivative Test fails.

[1] The words "maxima" and "minima" are the traditional plural forms of maximum and minimum, respectively.

[2] A formal proof requires the Mean Value Theorem, which will be presented in Section 4.4.

There is a simple visual mnemonic device for remembering the Second Derivative Test, due to a generic minimum or maximum resembling a smile or frown, respectively:



$f'' > 0$ local min. $f'' < 0$ local max. $f'' = 0$ test fails

The "eyes" in the faces represent the sign of f'' at a critical point, while the "mouths" indicate the nature of that point (when $f'' = 0$ nothing is known). The procedure for finding a global maximum or minimum can now be stated:

How to find a global maximum or minimum
Suppose that f is defined on an interval I. There are two cases:

1. **The interval I is closed**: The global maximum of f will occur either at an interior local maximum or at one of the endpoints of I whichever of these points provides the largest value of f will be where the global maximum occurs.
   Similarly, the global minimum of f will occur either at an interior local minimum or at one of the endpoints of I; whichever of these points provides the smallest value of f will be where the global minimum occurs.

2. **The interval I is not closed and has only one critical point**: If the only critical point is a local maximum then it is a global maximum. If the only critical point is a local minimum then it is a global minimum.

In each case of the above procedure try to use the Second Derivative Test to verify that a critical point is a local minimum or maximum, unless it is obvious from the nature of the problem that there can be only a minimum or only a maximum.

Example 4.1

Show that the rectangle with the largest area for a fixed perimeter is a square.

Solution: Let L be the perimeter of a rectangle with sides x and y. The idea is that L is a fixed constant, but x and y can vary. Figure 4.1.1 shows that there are many possible shapes for the rectangle, but in all cases $L = 2x + 2y$. Let A be the area of such a rectangle. Then $A = xy$, which is a function of two variables. But

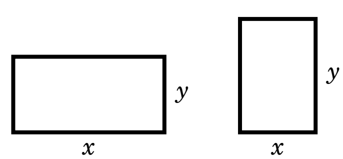


Figure 4.1.1

$$L = 2x + 2y \quad \Rightarrow \quad y = \frac{L}{2} - x\,,$$

and hence

$$A = x\left(\frac{L}{2} - x\right) = \frac{Lx}{2} - x^2$$

is now a function of x alone, on the open interval $(0, L/2)$ (since the length x is positive). Now find the critical points of A:

$$A'(x) = 0 \quad\Rightarrow\quad \frac{L}{2} - 2x = 0$$
$$\Rightarrow\quad x = \frac{L}{4} \text{ is the only critical point}$$

This problem is thus the case of a function defined on an open interval having only one critical point. Use the Second Derivative Test to verify that the sole critical point $x = L/4$ is a local maximum for A:

$$A''(x) = -2 \quad\Rightarrow\quad A''(L/4) = -2 < 0 \quad\Rightarrow\quad A \text{ has a local maximum at } x = L/4$$

Thus, A has a *global* maximum at $x = L/4$. Also, $y = L/2 - x = L/2 - L/4 = L/4$, which means that $x = y$, i.e. the rectangle is a square.
Note: The constraint in this example was $L = 2x + 2y$ and the objective function was $A = xy$.

Example 4.2

Suppose a right circular cylindrical can with top and bottom lids will be assembled to have a fixed volume. Find the radius and height of the can that minimizes the total surface area of the can.

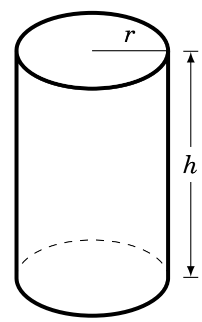


Figure 4.1.2

Solution: Let V be the fixed volume of the can with radius r and height h, as in Figure 4.1.2. The volume V is a constant, with $V = \pi r^2 h$. Let S be the total surface area of the can, including the lids. Then

$$S = 2\pi r^2 + 2\pi rh$$

where the first term in the sum on the right side of the equation is the combined area of the two circular lids and the second term is the lateral surface area of the can. So S is a function of r and h, but h can be eliminated since

$$V = \pi r^2 h \quad\Rightarrow\quad h = \frac{V}{\pi r^2}$$

and so

$$S = 2\pi r^2 + 2\pi r \cdot \frac{V}{\pi r^2} = 2\pi r^2 + \frac{2V}{r}$$

making S a function of r alone. Now find the critical points of S (i.e. solve $S'(r) = 0$):

$$S'(r) = 0 \quad\Rightarrow\quad 4\pi r - \frac{2V}{r^2} = 0$$
$$\Rightarrow\quad r^3 = \frac{V}{2\pi}$$
$$\Rightarrow\quad r = \sqrt[3]{\frac{V}{2\pi}} \text{ is the only critical point}$$

Since both r and h are lengths and have to be positive, then $0 < r < \infty$. So this is another case of a function defined on an open interval having only one critical point. Use the Second Derivative Test to verify that this critical point $r = \sqrt[3]{\frac{V}{2\pi}}$ is a local minimum for S:

$$S''(r) = 4\pi + \frac{4V}{r^3} \quad\Rightarrow\quad S''\left(\sqrt[3]{\frac{V}{2\pi}}\right) = 4\pi + \frac{4V}{\frac{V}{2\pi}} = 12\pi > 0 \quad\Rightarrow\quad S \text{ has a local minimum at } r = \sqrt[3]{\frac{V}{2\pi}}$$

Thus, S has a global minimum at $r = \sqrt[3]{\frac{V}{2\pi}}$, and

$$r \cdot r^2 = r^3 = \frac{V}{2\pi} \quad\Rightarrow\quad 2r = \frac{V}{\pi r^2} = h\,.$$

Hence, $r = \sqrt[3]{\frac{V}{2\pi}}$ and $h = 2\sqrt[3]{\frac{V}{2\pi}}$ will minimize the total surface area, i.e. the height should equal the diameter.

Note that this result can be applied to soda cans, where the volume is $V = 12$ fluid ounces $\approx$ 21.6 cubic inches: both a diameter and height of about 3.8 inches will minimize the amount (and hence the cost) of the aluminum used for the can. Yet soda cans are not that wide and short—they are usually thinner and taller. So why is a non-optimal size used in practice? Other factors—e.g. packing requirements, the need for small children to hold the can in one hand—might override the desire to minimize the cost of the aluminum. The lesson is that an optimal solution for one factor (material cost) might not always be truly optimal when all factors are considered; compromise is often necessary.

Example 4.3

Suppose that a projectile is launched from the ground with a fixed initial velocity v_0 at an angle θ with the ground. What value of θ would maximize the horizontal distance traveled by the projectile, assuming the ground is flat and not sloped (i.e. horizontal)?

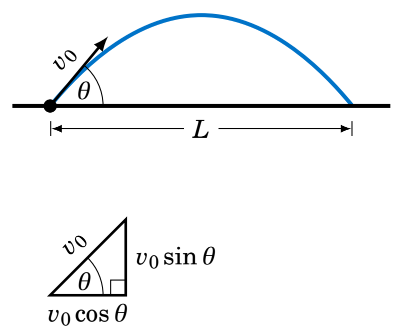


Figure 4.1.3

Solution: Let x and y represent the horizontal position and vertical position, respectively, of the projectile at time $t \ge 0$. From the triangle at the bottom of Figure 4.1.3, the horizontal and vertical components of the initial velocity are $v_0 \cos\theta$ and $v_0 \sin\theta$, respectively. Since distance is the product of velocity and time, then the horizontal and vertical distances traveled by the projectile by time t *due to the initial velocity* are $(v_0 \cos\theta)t$ and $(v_0 \sin\theta)t$, respectively. Ignoring wind and air resistance, the only other force on the projectile will be the downward force g due to gravity, so that the equations of motion for the projectile are:

$$x = (v_0 \cos\theta)t$$
$$y = -\frac{1}{2}gt^2 + (v_0 \sin\theta)t$$

The goal is to find θ that maximizes the length L shown in Figure 4.1.3. First write y as a function of x:

$$x = (v_0 \cos\theta)t \quad\Rightarrow\quad t = \frac{x}{v_0 \cos\theta} \quad\Rightarrow\quad y = -\frac{1}{2}g\left(\frac{x}{v_0 \cos\theta}\right)^2 + (v_0 \sin\theta)\cdot\frac{x}{v_0 \cos\theta}$$

$$\Rightarrow\quad y = -\frac{gx^2}{2v_0^2 \cos^2\theta} + x\tan\theta$$

Then L is the value of $x > 0$ that makes $y = 0$:

$$0 = -\frac{gL^2}{2v_0^2\cos^2\theta} + L\tan\theta \quad\Rightarrow\quad L = \frac{2v_0^2\sin\theta\,\cos\theta}{g} = \frac{v_0^2\sin 2\theta}{g}$$

So L is now a function of θ, with $0 < \theta < \pi/2$ (why?). So if there is a single local maximum then it must be the global maximum. Now get the critical points of L:

$$L'(\theta) = 0 \quad\Rightarrow\quad \frac{2v_0^2\cos 2\theta}{g} = 0$$

$$\Rightarrow\quad \cos 2\theta = 0$$

$$\Rightarrow\quad 2\theta = \frac{\pi}{2} \quad\Rightarrow\quad \theta = \frac{\pi}{4} \text{ is the only critical point}$$

Use the Second Derivative Test to verify that L has a local maximum at $\theta = \pi/4$:

$$L''(\theta) = -\frac{4v_0^2\sin 2\theta}{g} \quad\Rightarrow\quad L''(\pi/4) = -\frac{4v_0^2}{g} < 0$$

$$\Rightarrow\quad L \text{ has a local maximum at } \theta = \frac{\pi}{4}$$

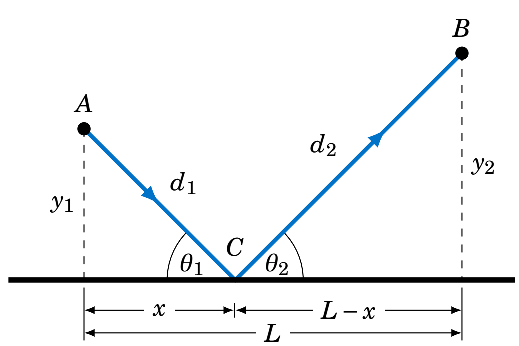
Thus, L has a global maximum at $\theta = \frac{\pi}{4}$, i.e. the projectile travels the farthest horizontally when launched at a 45° angle with the ground (with $L\left(\frac{\pi}{4}\right) = \frac{v_0^2}{g}$ being the maximum horizontal distance).

Note that once the formula for L as a function of θ was found to be $L = \frac{v_0^2\sin 2\theta}{g}$, calculus was not actually needed to solve this problem. Why? Since v_0^2 and g are positive constants (recall $g = 9.8\text{m/s}^2$), L would have its largest value when $\sin 2\theta$ has its largest value 1, which occurs when $\theta = \pi/4$.

Example 4.4

Fermat's Principle states that light always travels along the path that takes the least amount of time. So suppose that a ray of light is shone from a point A onto a flat horizontal reflective surface at an angle θ_1 with the surface and then reflects off the surface at an angle θ_2 to a point B. Show that Fermat's Principle implies that $\theta_1 = \theta_2$.

Solution: Let L be the horizontal distance between A and B, let d_1 be the distance the light travels from A to the point of contact C with the surface a horizontal distance x from A, let d_2 be the distance from C to B, and let y_1 and y_2 be the vertical distances from A and B, respectively, to the surface, as in the picture below.



Since time is distance divided by speed, and since the speed of light is constant, then minimizing the total time elapsed is equivalent to minimizing the total distance traveled, namely $D = d_1 + d_2$. The basic idea here is that Fermat's Principle implies that for the light to go from A to B in the shortest time, the unknown point C—and hence the unknown distance x—will have to be at a point that makes $\theta_1 = \theta_2$. The distances L, y_1 and y_2 are constants, so the goal is to write the total distance D as a function of x, find the x that minimizes D, then show that that value of x makes $\theta_1 = \theta_2$.

First, note that C has to be between A and B as in the picture, otherwise the total distance D would be larger than if C were directly below either A or B. This ensures that θ_1 and θ_2 are between 0 and $\pi/2$, and that $0 \le x \le L$.

Next, by the Pythagorean Theorem and the above picture,

$$d_1 = \sqrt{x^2 + y_1^2} \quad \text{and} \quad d_2 = \sqrt{(L-x)^2 + y_2^2}$$

and so the total distance $D = d_1 + d_2$ traveled by the light is a function of x:

$$D(x) = \sqrt{x^2 + y_1^2} + \sqrt{(L-x)^2 + y_2^2}$$

To find the critical points of D, solve the equation $D'(x) = 0$:

$$D'(x) = \frac{x}{\sqrt{x^2 + y_1^2}} - \frac{L-x}{\sqrt{(L-x)^2 + y_2^2}} = 0 \quad \Rightarrow \quad \frac{x}{d_1} = \frac{L-x}{d_2} \quad \Rightarrow \quad \sin\theta_1 = \sin\theta_2 \quad \Rightarrow \quad \theta_1 = \theta_2$$

since the sine function is one-to-one over the interval $[0, \frac{\pi}{2}]$.

This seems to prove the result, except for one remaining issue to resolve: verifying that the minimum for D really does occur at the x between 0 and L where $D'(x) = 0$, not at the endpoints $x = 0$ or $x = L$ of the closed interval $[0, L]$. Note that using the Second Derivative Test in this case does not matter, since you would have to check the value of D at the endpoints anyway and compare those values to the values of D at the critical points. To find expressions for the critical points, note that

$$D'(x) = 0 \quad \Rightarrow \quad \frac{x}{\sqrt{x^2 + y_1^2}} = \frac{L-x}{\sqrt{(L-x)^2 + y_2^2}} \quad \Rightarrow \quad \frac{x^2}{x^2 + y_1^2} = \frac{(L-x)^2}{(L-x)^2 + y_2^2}$$

$$\Rightarrow \quad \cancel{(L-x)^2 x^2} + x^2 y_2^2 = \cancel{(L-x)^2 x^2} + (L-x)^2 y_1^2$$

$$\Rightarrow \quad x y_2 = (L-x) y_1 \quad \Rightarrow \quad x = \frac{L y_1}{y_1 + y_2} \quad \text{is the only critical point,}$$

and x is between 0 and L. Now compare the values of $D^2(x)$ at $x = 0$, $x = L$, and $x = \frac{L y_1}{y_1 + y_2}$:

$$D^2(0) = L^2 + y_1^2 + y_2^2 + 2y_1\sqrt{L^2 + y_2^2}$$

$$D^2(L) = L^2 + y_1^2 + y_2^2 + 2y_2\sqrt{L^2 + y_1^2}$$

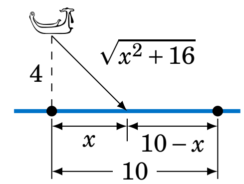
$$D^2\left(\tfrac{L y_1}{y_1 + y_2}\right) = L^2 + y_1^2 + y_2^2 + 2y_1 y_2$$

Since $y_2 < \sqrt{L^2 + y_2^2}$ and $y_1 < \sqrt{L^2 + y_1^2}$, then $D^2\left(\frac{L y_1}{y_1 + y_2}\right)$ is the smallest of the three values above, so that $D^2(x)$ has its minimum value at $x = \frac{L y_1}{y_1 + y_2}$, which means $D(x)$ has its minimum value there. ✓

Example 4.5

A man is in a boat 4 miles off a straight coast. He wants to reach a point 10 miles down the coast in the minimum possible time. If he can row 4 mi/hr and run 5 mi/hr, where should he land the boat?

Solution: Let T be the total time traveled. The goal is to minimize T. From the picture on the right, since time is distance divided by speed, break the total time into two parts: the time rowing in the water and the time running on the coast, so that



$$T = \text{time}_{\text{row}} + \text{time}_{\text{run}} = \frac{\text{dist}_{\text{row}}}{\text{speed}_{\text{row}}} + \frac{\text{dist}_{\text{run}}}{\text{speed}_{\text{run}}} = \frac{\sqrt{x^2+16}}{4} + \frac{10-x}{5} ,$$

where $0 \le x \le 10$ is the distance along the coast where the boat lands. Then

$$T'(x) = \frac{x}{4\sqrt{x^2+16}} - \frac{1}{5} = 0 \quad\Rightarrow\quad 5x = 4\sqrt{x^2+16} \quad\Rightarrow\quad 25x^2 = 16(x^2+16) \quad\Rightarrow\quad x = \frac{16}{3} ,$$

so $x = \frac{16}{3}$ is the only critical point. Thus, the (global) minimum of T will occur at $x = \frac{16}{3}$, $x = 0$, or $x = 10$. Since

$$T(0) = 3 \quad , \quad T(10) = \frac{\sqrt{29}}{2} \approx 2.693 \quad , \quad T\left(\frac{16}{3}\right) = \frac{13}{5} = 2.6$$

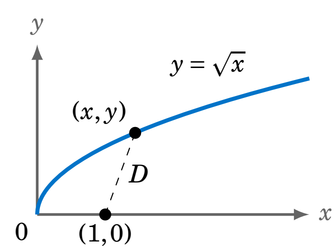
then $T(\frac{16}{3}) < T(10) < T(0)$. Hence, the global minimum occurs when landing the boat $x = \frac{16}{3} \approx 5.33$ miles down the coast.

Note that this example shows the importance of checking the endpoints. It was quite close between landing about 5.33 miles down the coast (2.6 hours) or simply rowing all the way to the destination (about 2.693 hours)—the difference is only about 5.6 minutes. With just a slight change in a few of the numbers, the minimum could have occurred at an endpoint. Moral: **always** check the endpoints![3]

Example 4.6

Find the point (x,y) on the graph of the curve $y = \sqrt{x}$ that is closest to the point $(1,0)$.

Solution: Let (x,y) be a point on the curve $y = \sqrt{x}$. Then $(x,y) = (x, \sqrt{x})$, so by the distance formula the distance D between (x,y) and $(1,0)$ is given by



$$D^2 = (x-1)^2 + (y-0)^2 = (x-1)^2 + (\sqrt{x})^2 = (x-1)^2 + x ,$$

which is a function of $x \ge 0$ alone. Note that minimizing D is equivalent to minimizing D^2. Since

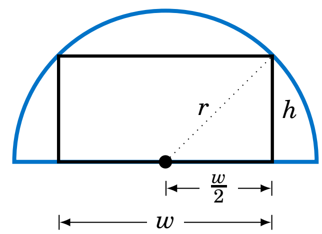
$$\frac{d(D^2)}{dx} = 2(x-1) + 1 = 2x - 1 = 0 \quad\Rightarrow\quad x = \tfrac{1}{2} \text{ is the only critical point,}$$

and since $\frac{d^2(D^2)}{dx^2} = 2 > 0$ for all x, then by the Second Derivative Test $x = 1/2$ is a local minimum. Hence, the global minimum for D^2 must occur at the endpoint $x = 0$ or at $x = 1/2$. But $D^2(0) = 1 > D^2(1/2) = 3/4$, so the global minimum occurs at $x = 1/2$. Hence, the closest point is $(x,y) = (1/2, \sqrt{1/2})$.

[3]Another possible lesson is that optimal in the mathematical sense might, again, not mean optimal in a practical sense. After all, presumably after the man is finished with whatever he had to do at the destination 10 miles down the coast, he then has the inconvenience of going back about 4.67 miles to retrieve his boat. At his running speed of 10 mph this would take 28 minutes, wiping out the 5.6 minutes he gained with his "optimal" landing spot!

Example 4.7

Find the width and height of the rectangle with the largest possible perimeter inscribed in a semicircle of radius r.



Solution: Let w be the width of the rectangle and let h be the height, as in the picture. Then the perimeter is $P = 2w+2h$. By symmetry and the Pythagorean Theorem,

$$h^2 = r^2 - \left(\frac{w}{2}\right)^2 \quad \Rightarrow \quad h = \frac{1}{2}\sqrt{4r^2-w^2}$$

and so $P = 2w+\sqrt{4r^2-w^2}$ for $0<w<2r$. Find the critical points of P:

$$P'(w) = 2 - \frac{w}{\sqrt{4r^2-w^2}} = 0 \quad \Rightarrow \quad w = 2\sqrt{4r^2-w^2}$$

$$\Rightarrow \quad w^2 = 16r^2 - 4w^2$$

$$\Rightarrow \quad w = \frac{4r}{\sqrt{5}} \text{ is the only critical point,}$$

and since

$$P''(w) = -\frac{4r^2}{(4r^2-w^2)^{3/2}} \quad \Rightarrow \quad P''\left(\tfrac{4r}{\sqrt{5}}\right) = -\frac{5^{3/2}}{2r} < 0$$

then P has a local maximum at $w = \frac{4r}{\sqrt{5}}$, by the Second Derivative Test. Since $P(w)$ is defined for w in the open interval $(0,2r)$, the local maximum is a global maximum. For the width $w = \frac{4r}{\sqrt{5}}$ the height is $h = \frac{r}{\sqrt{5}}$, which gives the dimensions for the maximum perimeter.

Note: If w were extended to include the cases of "degenerate" rectangles of zero width or height, i.e. $w=0$ or $w=2r$, then the maximum perimeter would still occur at $w = \frac{4r}{\sqrt{5}}$, since $P\left(\frac{4r}{\sqrt{5}}\right) = \frac{10r}{\sqrt{5}} \approx 4.472r$ is larger than $P(0) = 2r$ and $P(2r) = 4r$.

Exercises

A

1. Find the point on the curve $y = x^2$ that is closest to the point $(4, -1/2)$.

2. Prove that for $0 \le p \le 1$, $p(1-p) \le \frac{1}{4}$.

3. A farmer wishes to fence a field bordering a straight stream with 1000 yd of fencing material. It is not necessary to fence the side bordering the stream. What is the maximum area of a rectangular field that can be fenced in this way?

4. The power output P of a battery is given by $P = VI - RI^2$, where I, V, and R are the current, voltage, and resistance, respectively, of the battery. If V and R are constant, find the current I that maximizes P.

5. An impulse turbine consists of a high speed jet of water striking circularly mounted blades. The power P generated by the turbine is $P = VU(V-U)$, where V is the speed of the jet and U is the speed of the turbine. If the jet speed V is constant, find the turbine speed U that maximizes P.

6. A man is in a boat 5 miles off a straight coast. He wants to reach a point 15 miles down the coast in the minimum possible time. If he can row 6 mi/hr and run 10 mi/hr, where should he land the boat?

7. The current I in a voltaic cell is

$$I = \frac{E}{R+r},$$

where E is the electromotive force and R and r are the external and internal resistance, respectively. Both E and r are internal characteristics of the cell, and hence can be treated as constants. The power P developed in the cell is $P = RI^2$. For which value of R is the power P maximized?

8. Find the point(s) on the ellipse $\frac{x^2}{25} + \frac{y^2}{16} = 1$ closest to $(-1,0)$.

9. Find the maximum area of a rectangle that can be inscribed in an ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, where $a > 0$ and $b > 0$ are arbitrary constants. Your answer should be in terms of a and b.

10. Find the radius and angle of the circular sector with the maximum area and a fixed perimeter P.

11. In Example 4.3 show that the maximum height reached by the projectile when launched with an initial velocity v_0 at an angle $0 < \theta < \frac{\pi}{2}$ to the ground is $\frac{v_0^2 \sin^2\theta}{2g}$.

12. The *phase velocity* v of a capillary wave with surface tension T and water density p is

$$v = \sqrt{\frac{2\pi T}{\lambda p} + \frac{\lambda g}{2\pi}}$$

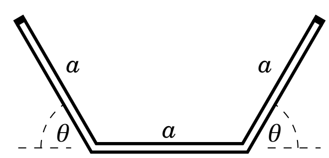
where λ is the wavelength. Find the value of λ that minimizes v.

13. For an inventory model with a constant order quantity $Q > 0$ and a constant linear inventory depletion rate D, the *total unit cost* TC to maintain an average inventory of $Q/2$ units is

$$TC = C + \frac{P}{Q} + \frac{(I+W)Q}{2D}$$

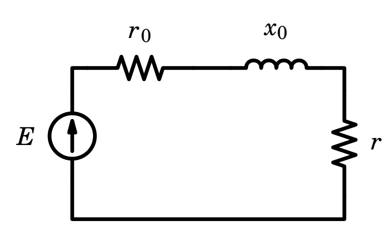
where C is the capital investment cost, P is the cost per order, I is the per unit interest charge per unit time, and W is the overall inventory holding cost. Find the value of Q that minimizes TC.

14. The opening of a rain gutter—shown in the figure on the right—has a bottom and two sides each with length a. The sides make an angle θ with the bottom. Find the value of θ that maximizes the amount of rain the gutter can hold.



B

15. In an electric circuit with a supplied voltage (emf) E, a resistor with resistance r_0, and an inductor with reactance x_0, suppose you want to add a second resistor. If r represents the resistance of this second resistor then the power P delivered to that resistor is given by



$$P = \frac{E^2 r}{(r+r_0)^2 + x_0^2}$$

with E, r_0, and x_0 treated as constants. For which value of r is the power P maximized?

16. The *stress* τ in the xy-plane along a varying angle ϕ is given by

$$\tau = \tau(\phi) = \frac{\sigma_x - \sigma_y}{2} \sin 2\phi + \tau_{xy} \cos 2\phi ,$$

where σ_x, σ_y, and τ_{xy} are *stress components* that can be treated as constants. Show that the maximum stress is

$$\tau = \frac{\sqrt{(\sigma_x - \sigma_y)^2 + 4\tau_{xy}^2}}{2} .$$

(Hint: Draw a right triangle with angle 2ϕ after finding the critical point(s).)

17. A certain jogger can run 0.16 km/min, and walks at half that speed. If he runs along a circular trail with circumference 50 km and then—before completing one full circle—walks back straight across to his starting point, what is the maximum time he can spend on the run/walk?

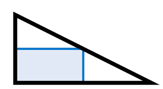
18. A rectangular poster is to contain 50 square inches of printed material, with 4-inch top and bottom margins and 2-inch side margins. What dimensions for the poster would use the least paper?

19. Find the maximum volume of a right circular cylinder that can be inscribed in a sphere of radius 3.

20. A figure consists of a rectangle whose top side coincides with the diameter of a semicircle atop it. If the perimeter of the figure is 20 m, find the radius and height of the semicircle and rectangle, respectively, that maximizes the area inside the figure.

21. A thin steel pipe 25 ft long is carried down a narrow corridor 5.4 ft wide. At the end of the corridor is a right-angle turn into a wider corridor. How wide must this corridor be in order to get the pipe around the corner? You may assume that the width of the pipe can be ignored.

22. A rectangle is inscribed in a right triangle, with one corner of the rectangle at the right angle of the triangle. Show that the maximum area of the rectangle occurs when a corner of the rectangle is at the midpoint of the hypotenuse of the triangle.



23. Find the relation between the radius and height of a cylindrical can with an open top that maximizes the volume of the can, given that the surface area of the can is always the same fixed amount.

24. An isosceles triangle is circumscribed about a circle of radius r. Find the height of the triangle that minimizes the perimeter of the triangle.

25. Suppose N voltaic cells are arranged in N/x rows in parallel, with each row consisting of x cells in series, creating a current I through an external resistance R. Each cell has internal resistance r and EMF (voltage) e. Find the x that maximizes the current I, which—due to Ohm's Law—is given by

$$I = \frac{xe}{(x^2r/N) + R} .$$

26. A single-degree-of-freedom harmonically forced vibration system with damping factor ζ has *magnification factor* $MF = ((1-r^2)^2 + (2\zeta r)^2)^{-1/2}$, where r is the frequency ratio. Find the value of r that maximizes MF.

27. At a distance $x \geq 0$ from the center of a uniform ring with charge q and radius a, the magnitude E of the electric field for points on the axis of the ring is

$$E = \frac{qx}{4\pi\epsilon_0 (a^2 + x^2)^{3/2}}$$

where ϵ_0 is the permittivity of free space. Find the distance x that maximizes E.

28. Find the equation of the tangent line to the ellipse $\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$ in the first quadrant that forms with the coordinate axes the right triangle with minimal area.

29. A "cold" star that has exhausted its nuclear fuel—called a *white dwarf*—has total energy E, given by

$$E = \frac{\hbar^2(3\pi^2Nq)^{5/3}}{10\pi^2 m}\left(\frac{4}{3}\pi R^3\right)^{-2/3} - \frac{3GM^2N^2}{5R}$$

where $\hbar$ is the reduced Planck constant, N is the number of *nucleons*—protons and neutrons—in the star, q is the charge of an electron, m is the mass of an electron, M is the mass of a nucleon, G is the gravitational constant, and R is the radius of the star. Show that the radius R that minimizes E is

$$R = \left(\frac{9\pi}{4}\right)^{2/3}\frac{\hbar^2 q^{5/3}}{GmM^2N^{1/3}}\,.$$

C

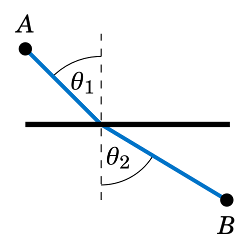
30. An object of mass m has orbital angular momentum l around a black hole with *Schwarzchild radius* r_S and mass M. The *effective potential* Φ of the object is

$$\Phi = -\frac{GM}{r} + \frac{l^2}{2m^2r^2} - \frac{r_S l^2}{2m^2r^3}$$

where G is the gravitational constant and r is the object's distance from the black hole. Show that Φ has a local maximum and minimum at $r=r_1$ and $r=r_2$, respectively, where

$$r_1 = \frac{l^2}{2GMm^2}\left(1-\sqrt{1-\frac{6GMm^2r_S}{l^2}}\right) \quad\text{and}\quad r_2 = \frac{l^2}{2GMm^2}\left(1+\sqrt{1-\frac{6GMm^2r_S}{l^2}}\right).$$

31. Recall Fermat's Principle from Example 4.4, which states that light travels along the path that takes the least amount of time. The speed of light in a vacuum is approximately $c = 2.998\times10^8$ m/s, but in some other medium (e.g. water) light is slower. Suppose that a ray of light goes from a point A in one medium where it moves at a speed v_1 and ends up at a point B in another medium where it moves at a speed v_2. Use Fermat's Principle to prove *Snell's Law*, which says that the light is refracted through the boundary between the two media such that



$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2}$$

where θ_1 and θ_2 are the angles that the light makes with the normal line perpendicular to the boundary of the media in the first and second medium, respectively, as in the picture above.

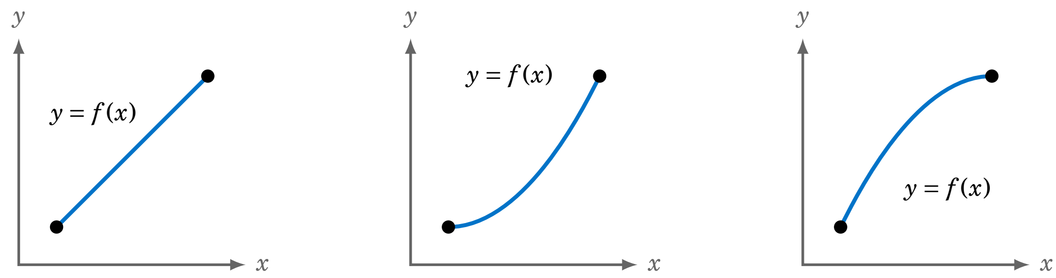
32. A sphere of radius a is inscribed in a right circular cone, with the sphere touching the base of the cone. Find the radius and height of the cone if its volume is a minimum.

33. Find the length of the shortest line segment from the positive x-axis to the positive y-axis going through a point (a,b) in the first quadrant.

34. Find the radius r of a circle c whose center is on a fixed circle C of radius R such that the arc length of the part of c within C is a maximum.

4.2 Curve Sketching

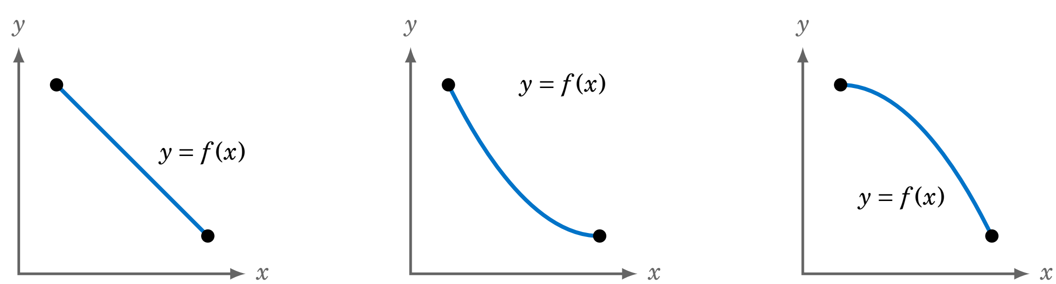
A function can increase between two points in different ways, as shown in Figure 4.2.1.



(a) $f'' = 0$: straight **(b)** $f'' > 0$: concave up **(c)** $f'' < 0$: concave down

Figure 4.2.1 Increasing function f: $f' > 0$, different signs for f''

In each case in the above figure the function is increasing, so that $f'(x) > 0$, but the manner in which the function increases is determined by its **concavity**, that is, by the sign of the second derivative $f''(x)$. The function in the graph on the far left is linear, i.e. of the form $f(x) = ax + b$ for some constants a and b, so that $f''(x) = 0$ for all x. But the functions in the other two graphs are nonlinear. In the middle graph the derivative f' is increasing, so that $f'' > 0$; in this case the function is called **concave up**. In the graph on the far right the derivative f' is decreasing, so that $f'' < 0$; in this case the function is called **concave down**. The same definitions would hold if the function were decreasing, as shown in Figure 4.2.2 below:



(a) $f'' = 0$ **(b)** $f'' > 0$: concave up **(c)** $f'' < 0$: concave down

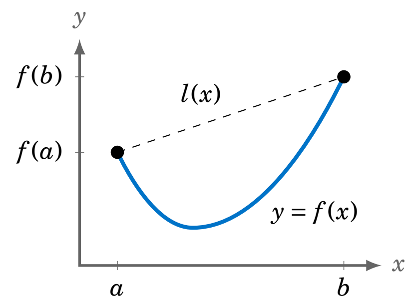
Figure 4.2.2 Decreasing function f: $f' < 0$, different signs for f''

In Figures 4.2.1(b) and 4.2.2(b) the function is below the line joining the points at each end, while in Figure 4.2.1(c) and 4.2.2(c) the function is above that line. This turns out to be true in general, as a result of the following theorem:

Concavity Theorem: Suppose that f is a twice-differentiable function on $[a,b]$. Then:

(a) If $f''(x) > 0$ on (a,b) then $f(x)$ is below the line $l(x)$ joining the points $(a, f(a))$ and $(b, f(b))$ for all x in (a,b).

(b) If $f''(x) < 0$ on (a,b) then $f(x)$ is above the line $l(x)$ joining the points $(a, f(a))$ and $(b, f(b))$ for all x in (a,b).



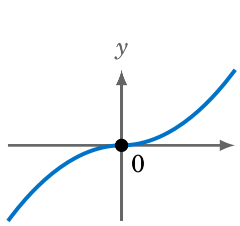
Proof: Only part (a) will be proved; the proof of part (b) is similar and left as an exercise. So assume that $f''(x) > 0$ on (a,b), and $l(x)$ be the line joining $(a, f(a))$ and $(b, f(b))$, as in the drawing on the right. The drawing suggests that $f(x) < l(x)$ over (a,b), but this is what must be proved.

The goal is to show that $g(x) = f(x) - l(x) < 0$ on (a,b), since this will show that $f(x) < l(x)$ on (a,b). Since f and l are both continuous on $[a,b]$ then so is g. Hence g has a global maximum somewhere in $[a,b]$, by the Extreme Value Theorem. Suppose the global maximum occurs at an interior point $x = c$, i.e. for some c in the open interval (a,b). Then $g'(c) = 0$ and $g''(c) = f''(c) - l''(c) = f''(c) > 0$, since $l(x)$ is a line and hence has a second derivative of 0 for all x. Then by the Second Derivative Test g has a local minimum at $x = c$, which contradicts g having a global maximum at $x = c$. Thus, the global maximum of g cannot occur at an interior point, so it must occur at one of the end points $x = a$ or $x = b$. In other words, either $g(x) < g(a)$ or $g(x) < g(b)$ for all x in (a,b). But $f(a) = l(a)$ and $f(b) = l(b)$, so $g(a) = 0 = g(b)$. Hence, $g(x) < 0$ for all x in (a,b), i.e. $f(x) < l(x)$ for all x in (a,b). ✓

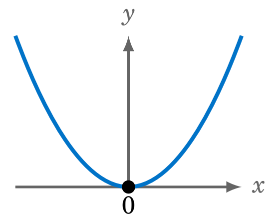
Points where the concavity of a function changes have a special name:

A function f has an **inflection point** at $x = c$ if the concavity of f changes around $x = c$. That is, the function goes from concave up to concave down, or vice versa.

Note that to be an inflection point it does not suffice for the second derivative to be 0 at that point; the *second derivative must change sign* around that point, either from positive to negative or from negative to positive. For example, $f(x) = x^3$ has an inflection point at $x = 0$, since $f''(x) = 6x < 0$ for $x < 0$ and $f''(x) = 6x > 0$ for $x > 0$, i.e. $f''(x)$ changes sign around $x = 0$ (and of course $f''(0) = 0$). But for $f(x) = x^4$, $x = 0$ is *not* an inflection point even though $f''(0) = 0$, since $f''(x) = 12x^2 \geq 0$ is always nonnegative. That is, $f(x) = x^4$ is always concave up. Figure 4.2.3 below shows the difference:



(a) $f(x) = x^3$: inflection point at $x = 0$



(b) $f(x) = x^4$: non-inflection point at $x = 0$

Figure 4.2.3 Inflection vs non-inflection point at $x = 0$ with $f''(0) = 0$

Figure 4.2.3(b) shows that a point where the second derivative is 0 is a *possible* inflection point, but you still must check that the second derivative changes sign around that point. Using local minima and maxima, concavity and inflection points, and where a function increases or decreases, you can sketch the graph of a function.

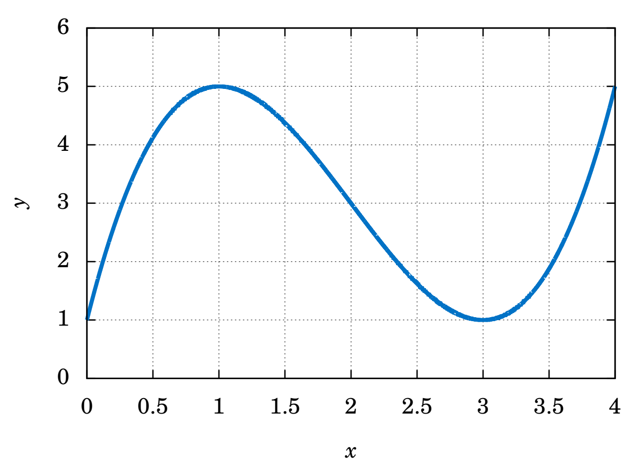
Example 4.8

Sketch the graph of $f(x) = x^3 - 6x^2 + 9x + 1$. Find all local maxima and minima, inflection points, where the function is increasing or decreasing, and where the function is concave up or concave down.

Solution: Since $f'(x) = 3x^2 - 12x + 9 = 3(x-1)(x-3)$ then $x = 1$ and $x = 3$ are the only critical points. And since $f''(x) = 6x - 12$ then $f''(1) = -6 < 0$ and $f''(3) = 6 > 0$. So by the Second Derivative Test, f has a local maximum at $x = 1$ and a local minimum at $x = 3$. Since $f''(x) = 6x - 12 < 0$ for $x < 2$ and $f''(x) = 6x - 12 > 0$ for $x > 2$, then $x = 2$ is an inflection point, and f is concave down for $x < 2$ and concave up for $x > 2$. The table below shows where f is increasing and decreasing, based on the sign of f':

x values	$3(x-1)$	$(x-3)$	$f'(x)$	direction
$x < 1$	$-$	$-$	$+$	f is increasing
$1 < x < 3$	$+$	$-$	$-$	f is decreasing
$x > 3$	$+$	$+$	$+$	f is increasing

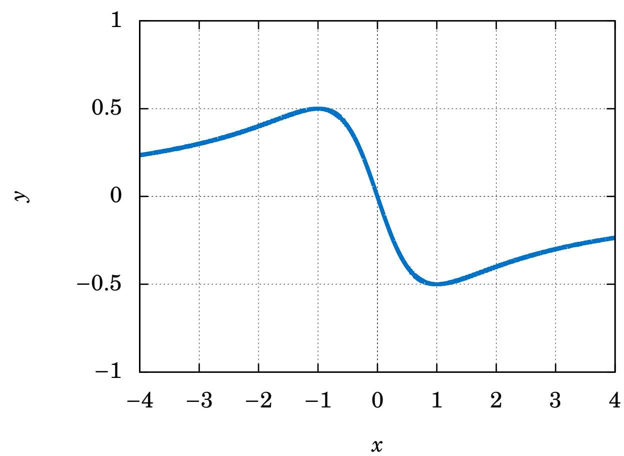
The graph is shown below:



Example 4.9

Sketch the graph of $f(x) = \frac{-x}{1+x^2}$. Find all local maxima and minima, inflection points, where the function is increasing or decreasing, and where the function is concave up or concave down. Also indicate any asymptotes.

Solution: Since $f'(x) = \frac{x^2-1}{(1+x^2)^2}$ then $x = 1$ and $x = -1$ are the only critical points. And since $f''(x) = \frac{2x(3-x^2)}{(1+x^2)^3}$ then $f''(1) = \frac{1}{2} > 0$ and $f''(-1) = -\frac{1}{2} < 0$. So by the Second Derivative Test, f has a local minimum at $x = 1$ and a local maximum at $x = -1$. Since $f''(x) > 0$ for $x < -\sqrt{3}$, $f''(x) < 0$ for $-\sqrt{3} < x < 0$, $f''(x) > 0$ for $0 < x < \sqrt{3}$, and $f''(x) < 0$ for $x > \sqrt{3}$, then $x = 0, \pm\sqrt{3}$ are inflection points, f is concave up for $x < -\sqrt{3}$ and for $0 < x < \sqrt{3}$, and f is concave down for $-\sqrt{3} < x < 0$ and for $x > \sqrt{3}$. Since $f'(x) > 0$ for $x < -1$ and $x > 1$ then f is increasing for $|x| > 1$. And $f'(x) < 0$ for $-1 < x < 1$ means f is decreasing for $|x| < 1$. Finally, since $\lim_{x\to\infty} f(x) = 0$ and $\lim_{x\to-\infty} f(x) = 0$ then the x-axis ($y = 0$) is a horizontal asymptote. There are no vertical asymptotes (why?).
The graph is shown below:



If the Second Derivative Test fails then one alternative is the following test:

First Derivative Test: For a continuous function f on an interval I, let $x = c$ be a number in I such that $f(c)$ is defined, and either $f'(c) = 0$ or $f'(c)$ does not exist. Then:

(a) If $f'(x)$ changes from negative to positive around $x = c$ then f has a local minimum at $x = c$.

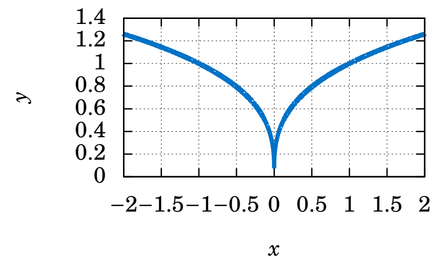
(b) If $f'(x)$ changes from positive to negative around $x = c$ then f has a local maximum at $x = c$.

This test merely states the obvious: a function decreases then increases around a minimum, and it increases then decreases around a maximum.

Example 4.10

Sketch the graph of $f(x) = x^{2/3}$.

Solution: Clearly $f(x)$ is continuous for all x, including $x = 0$ (since $f(0) = 0$), but $f'(x) = \frac{2}{3\sqrt[3]{x}}$ is not defined at $x = 0$. Since $f'(x)$ changes from negative to positive around $x = 0$ ($f'(x) < 0$ when $x < 0$ and $f'(x) > 0$ when $x > 0$), then by the First Derivative Test f has a local minimum at $x = 0$. Since $f''(x) = -\frac{2}{9x^{4/3}} < 0$ for all $x \neq 0$, then f is always concave down. There are no vertical or horizontal asymptotes. The graph is shown on the right.



Note that the Second Derivative Test could not be used for this function, since $f'(x) \neq 0$ for all x (notice also that $f''(x)$ is not defined at $x = 0$).

A more complete alternative to the Second Derivative Test is the following:[4]

N$^{\text{th}}$ Derivative Test: A non-constant function f with continuous derivatives of all orders up to and including $n > 1$ at $x = c$ has either a local minimum, local maximum or inflection point at $x = c$ if and only if

$$f^{(k)}(c) = 0 \text{ for } k = 1, 2, \ldots, n-1 \quad \text{and} \quad f^{(n)}(c) \neq 0$$

(i.e. the n^{th} derivative is the first nonzero derivative at $x = c$). If so, then:

(a) If $n > 1$ is even and $f^{(n)}(c) > 0$ then f has a local minimum at $x = c$.

(b) If $n > 1$ is even and $f^{(n)}(c) < 0$ then f has a local maximum at $x = c$.

(c) If $n > 1$ is odd then f has an inflection point at $x = c$.

Note that the Second Derivative Test is the special case where $n = 2$ in the N$^{\text{th}}$ Derivative Test. Though this test gives necessary and sufficient conditions for a local maximum, local minimum, and inflection point, calculating the first n derivatives can be complicated if n is large and the given function is not simple.

Example 4.11

The Second Derivative Test fails for $f(x) = x^4$ at the critical point $x = 0$, since $f''(0) = 0$. But the first 4 derivatives of $f(x) = x^4$ are $f'(x) = 4x^3$, $f''(x) = 12x^2$, $f^{(3)}(x) = 24x$, and $f^{(4)}(x) = 24$, which are all continuous and

$$f^{(k)}(0) = 0 \text{ for } k = 1, 2, 3 \quad \text{and} \quad f^{(4)}(0) = 24 \neq 0 .$$

So by the N$^{\text{th}}$ Derivative Test, since $n = 4$ is even and $f^{(4)}(0) = 24 > 0$ then $f(x) = x^4$ has a local minimum at $x = 0$. Note that $f(x) \geq 0 = f(0)$ for all x, so $x = 0$ is actually a global minimum for f.

[4] For a proof, see pp.10-11 in KOO, D., *Elements of Optimization*, New York: Springer-Verlag, 1977.

A common practice in many fields of science and engineering is to combine multiple named constants (e.g. π) or variables in a function into one variable and then sketch a graph of that function. The example below illustrates the technique.

Example 4.12

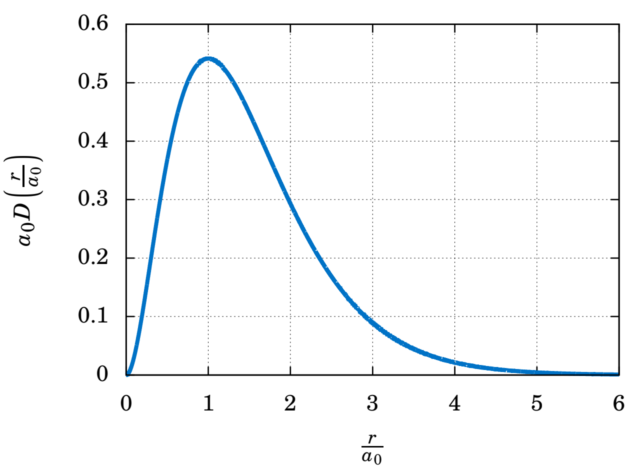
A hydrogen atom has one electron, and the probability of finding the electron in the ground state of the hydrogen atom between radii r and $r + dr$ is $D(r)\,dr$, where dr is an infinitesimal change in the radius r (the distance from the electron to the nucleus), $D(r)$ is the *radial probability density function*

$$D(r) \;=\; \frac{4}{a_0^3}\,r^2 e^{-2r/a_0}$$

and $a_0 \approx 5.291772 \times 10^{-11}$ m is the *Bohr radius*. It is useful to analyze this function in terms of $r \ge 0$ in relation to the Bohr radius a_0 (e.g. $r = 0.5a_0$, a_0, $2a_0$, $3a_0$). To do this, let $x = \frac{r}{a_0}$, so that

$$D(r) \;=\; \frac{4}{a_0}\left(\frac{r}{a_0}\right)^2 e^{-2\left(\frac{r}{a_0}\right)} \quad\Rightarrow\quad a_0\,D(x) \;=\; 4x^2 e^{-2x}$$

and then sketch the graph of $a_0\,D(x)$, which is shown below:



From the graph it looks like $x = 1$ (i.e. $r = a_0$) is a local (and global) maximum, so that the electron is most likely to be found near $r = a_0$, and the probability drops off dramatically past a distance $r = 3a_0$. In the exercises you will be asked to show that $r = a_0$ is indeed a local maximum and that the inflection points are $r = \left(1 \pm \frac{1}{\sqrt{2}}\right) a_0$.

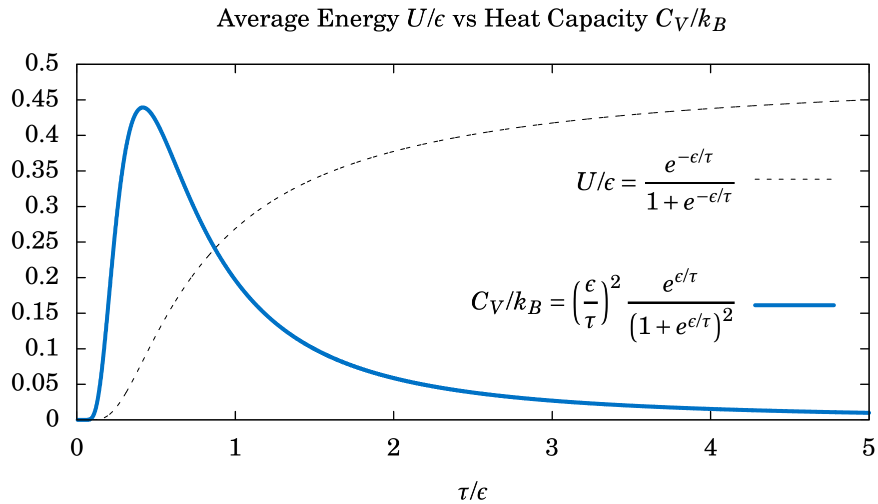
Note that the right side of the formula $a_0\,D(x) = 4x^2 e^{-2x}$ does not involve a_0, which was multiplied over to the left side. In general that is the strategy when dealing with these sorts of functions where variables and constants are combined. In this case the stray constant a_0 can be multiplied with D since that will not affect the location of critical and inflection points, nor fundamentally alter the general shape of the graph.

Example 4.13

For a single particle with two states—energy 0 and energy ϵ—in thermal contact with a reservoir at temperature τ, the average energy U and heat capacity C_V are given by

$$U = \epsilon \frac{e^{-\epsilon/\tau}}{1+e^{-\epsilon/\tau}} \quad \text{and} \quad C_V = k_B \left(\frac{\epsilon}{\tau}\right)^2 \frac{e^{\epsilon/\tau}}{\left(1+e^{\epsilon/\tau}\right)^2}$$

where $k_B \approx 1.38065 \times 10^{-23}$ J/K is the *Boltzmann constant*. The graph below shows both quantities as functions of τ/ϵ (*not* ϵ/τ, as you might expect). See Exercise 9.



Exercises

A

For Exercises 1-8 sketch the graph of the given function. Find all local maxima and minima, inflection points, where the function is increasing or decreasing, where the function is concave up or concave down, and indicate any asymptotes.

1. $f(x) = x^3 - 3x$ **2.** $f(x) = x^3 - 3x^2 + 1$ **3.** $f(x) = xe^{-x}$ **4.** $f(x) = x^2 e^{-x^2}$

5. $f(x) = \dfrac{1}{1 + x^2}$ **6.** $f(x) = \dfrac{x^2}{(x-1)^2}$ **7.** $f(x) = \dfrac{e^{-x} - e^{-2x}}{2}$ **8.** $f(x) = e^{-x} \sin x$

9. Write U/ϵ and C_V/k_B from Example 4.13 as functions of $x = \tau/\epsilon$. You do not need to sketch the graphs.

10. Show that the function $D(r) = \frac{4}{a_0^3} r^2 e^{-2r/a_0}$ from Example 4.12 has a local maximum at $r = a_0$ and inflection points at $r = \left(1 \pm \frac{1}{\sqrt{2}}\right) a_0$.

11. Sketch the graph of *Kratzer's molecular potential* $V(r) = -2D\left(\frac{a}{r} - \frac{1}{2}\frac{a^2}{r^2}\right)$ as a function of $x = \frac{r}{a}$, with $a > 0$ and $D > 0$ as constants.

12. Sketch the graph of $f(K) = \frac{2N\sqrt{K}e^{-\frac{K}{kT}}}{\sqrt{\pi}(kT)^{3/2}}$ as a function of $x = \frac{K}{kT}$, with N, k and T as positive constants.

13. Prove part (b) of the Concavity Theorem.

4.3 Numerical Approximation of Roots of Functions

When finding critical points of a function f, you encounter the problem of solving the equation $f'(x) = 0$. The examples and exercises so far were set up carefully so that solutions to that equation could be found in a simple closed form. But in practice this will not always be the case—in fact it is *almost never* the case. For example, finding the critical points of the function $f(x) = \sin x - \frac{x^2}{2}$ entails solving the equation $f'(x) = \cos x - x = 0$, for which there is no solution in a closed-form expression.

What should you do in such a situation?[5] One possibility is to use the bisection method mentioned in Section 3.3. In fact, in Example 3.25 the solution to the equation $\cos x = x$ (i.e. $\cos x - x = 0$) was shown to exist in the interval $[0,1]$, and then a demonstration of the bisection method was given to find that solution.

The bisection method is one of many *numerical methods* for finding roots of a function (i.e. where the function is zero). Finding the critical points of a function means finding the roots of its derivative. Though the bisection method could be used for that purpose, convergence to each root is usually slow. A far more efficient method is **Newton's method**[6], whose geometric interpretation is shown in Figure 4.3.1 below.

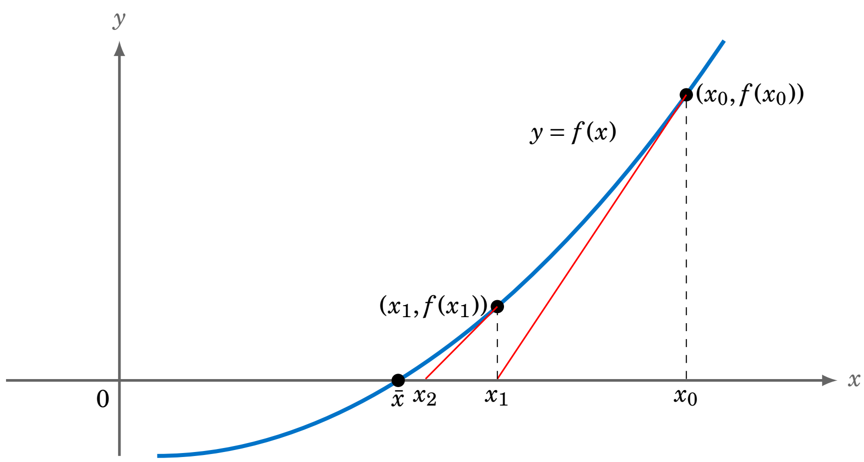


Figure 4.3.1 Newton's method for finding a root $\bar{x}$ of $f(x)$

The idea behind Newton's method is simple: to find a root $\bar{x}$ of a function f, choose an *initial guess* x_0 and then go up—or down—to the curve $y = f(x)$ and draw the tangent line to the curve at the point $(x_0, f(x_0))$. Let x_1 be where that tangent line intersects the x-axis, as shown above; repeat this procedure on x_1 to get the next number x_2, repeat on x_2 to get x_3, and so on. The resulting sequence of numbers x_0, x_1, x_2, x_3, ..., will approach the root $\bar{x}$. Convergence under certain conditions can be proved.[7]

[5]Note: To "just give up"—as suggested semi-seriously by some students I have had—is not an option.
[6]Sometimes called the *Newton-Raphson method*.
[7]See pp.58-62 in Saaty, T.L. and J. Bram, *Nonlinear Mathematics*, New York: McGraw-Hill, Inc., 1964.

The general formula for the number x_n obtained after $n \geq 1$ iterations in Newton's method can be determined by considering the formula for x_1. First, the tangent line to $y = f(x)$ at the point $(x_0, f(x_0))$ has slope $f'(x_0)$, so the equation of the line is

$$y - f(x_0) = f'(x_0)(x - x_0)\,. \tag{4.1}$$

The point $(x_1, 0)$ is (by design) also on that line, so that

$$0 - f(x_0) = f'(x_0)(x_1 - x_0) \quad \Rightarrow \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

provided that $f'(x_0) \neq 0$. The general formula for x_n is given by the following algorithm:

Newton's method: For an initial guess x_0, the numbers x_n for $n \geq 1$ are computed iteratively as:

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \qquad \text{for } n = 1,\, 2,\, 3,\, \ldots$$

That is, each "next" number x_n depends on the previous number x_{n-1}. The algorithm terminates whenever $f'(x_n) = 0$, or when the desired accuracy is reached. If $f'(x_n) = 0$ for some $n \geq 0$, then you could start over with a different initial guess x_0.

To implement this algorithm in a programming language (for which Newton's method is well-suited), the following language-independent *pseudocode* can be used as a guide:

Algorithm pseudocode for Newton's method

```
NEWTON'S METHOD
N ← NUMBER-OF-ITERATIONS          ▷ User supplies this value
x ← INITIAL-GUESS                 ▷ User supplies this value
for n ← 1 to N
      do
         if f'(x) ≠ 0
           then
                x ← x − f(x)/f'(x)
                print x
           else
                error "division by zero"
```

Example 4.14

Use Newton's method to find the root of $f(x) = \cos x - x$.

Solution: Since the root is already known to be in the interval $[0,1]$, choose $x_0 = 1$ as the initial guess. The numbers x_n for $n \geq 1$ can be computed with a hand-held scientific calculator, but the process is tedious and error-prone. Using a computer is far more efficient and allows more flexibility.

For example, the algorithm is easily implemented in the Java programming language. Save this code in a plain text file as `newton.java`:

Listing 4.1 Newton's method in Java (`newton.java`)

```
public class newton {
   public static void main(String[] args) {
      int N = Integer.parseInt(args[0]); //Number of iterations
      double x = 1.0; //initial guess
      System.out.println("n=0: " + x);
      for (int i = 1; i <= N; i++) {
         x =  x - f(x)/derivf(x);
         System.out.println("n=" + i + ": " + x);
      }
   }

   //Define the function f(x)
   public static double f(double x) {
      return Math.cos(x) - x;
   }

   //Define the derivative f'(x)
   public static double derivf(double x) {
      return -Math.sin(x) - 1.0;
   }
}
```

Though knowledge of Java would help, it should not be that difficult to figure out what the above code is doing. The number of iterations N is passed as a command-line parameter to the program, and x_n is computed and printed for $n = 0, 1, 2, \ldots, N$. Note that the derivative of $f(x)$ is "hard-coded" into the program.[8] There is also no error checking for the derivative being zero at any x_n. The program would simply halt on a division by zero error.

Compile the code, then run the program with 10 iterations:

```
javac newton.java
java newton 10
```

The output is shown below:

[8] There are some programming language libraries for calculating derivatives of functions "on the fly," i.e. dynamically. For example, the **GNU libmatheval** C/Fortran library can perform such symbolic operations. It is available at `http://www.gnu.org/software/libmatheval/`

```
n=0: 1.0
n=1: 0.7503638678402439
n=2: 0.7391128909113617
n=3: 0.739085133385284
n=4: 0.7390851332151607
n=5: 0.7390851332151607
n=6: 0.7390851332151607
n=7: 0.7390851332151607
n=8: 0.7390851332151607
n=9: 0.7390851332151607
n=10: 0.7390851332151607
```

Note that the solution $\bar{x} = 0.7390851332151607$ was found after only 4 iterations; the numbers x_n repeat for $n \geq 5$. This is much faster than the bisection method.

Another root-finding numerical method similar to Newton's method is the **secant method**, whose geometric interpretation is shown in Figure 4.3.2 below:

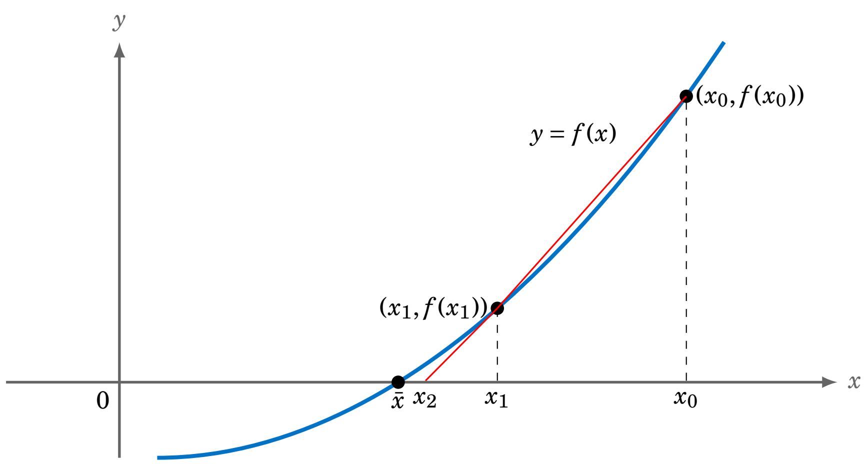


Figure 4.3.2 Secant method for finding a root $\bar{x}$ of $f(x)$

The idea behind the secant method is simple: to find a root $\bar{x}$ of a function f, choose *two* initial guesses x_0 and x_1, then go up—or down—to the curve $y = f(x)$ and draw the secant line through the points $(x_0, f(x_0))$ and $(x_1, f(x_1))$ on the curve. Let x_2 be where that secant line intersects the x-axis, as shown above; repeat this procedure on x_1 and x_2 to get the next number x_3, and keep repeating in this way. The resulting sequence of numbers x_0, x_1, x_2, x_3, ..., will approach the root $\bar{x}$, under the right conditions.[9]

[9]See pp.227-229 in DAHLQUIST, G. AND Å. BJÖRCK, *Numerical Methods*, Englewood Cliffs, NJ: Prentice-Hall, Inc., 1974.

Since the secant line through $(x_0, f(x_0))$ and $(x_1, f(x_1))$ has slope $\frac{f(x_1)-f(x_0)}{x_1-x_0}$, the equation of that secant line is:

$$y - f(x_1) = \frac{f(x_1)-f(x_0)}{x_1-x_0}(x - x_1)$$

The point $(x_2, 0)$ is on that line, so that

$$0 - f(x_1) = \frac{f(x_1)-f(x_0)}{x_1-x_0}(x_2 - x_1) \quad \Rightarrow \quad x_2 = x_1 - \frac{(x_1 - x_0)\cdot f(x_1)}{f(x_1) - f(x_0)}$$

provided that $x_1 \neq x_0$. The general formula for x_n is given by the following algorithm:

Secant method: For two initial guesses x_0 and x_1, the numbers x_n for $n \geq 2$ are computed iteratively as:

$$x_n = x_{n-1} - \frac{(x_{n-1} - x_{n-2})\cdot f(x_{n-1})}{f(x_{n-1}) - f(x_{n-2})} \qquad \text{for } n = 2, 3, 4, \ldots \tag{4.2}$$

That is, each "next" number x_n depends on the previous two numbers x_{n-1} and x_{n-2}. The algorithm terminates whenever $x_n = x_{n-1}$ (i.e. the numbers start repeating) or when the desired accuracy is reached.

Algorithm pseudocode for the secant method

```
SECANT METHOD
N ← NUMBER-OF-ITERATIONS        ▷ User supplies this value
x0 ← FIRST-INITIAL-GUESS        ▷ User supplies this value
x1 ← SECOND-INITIAL-GUESS       ▷ User supplies this value
f0 ← f(x0)
for n ← 1 to N
    do
       f1 ← f(x1)
       if f0 ≠ f1
         then
              x ← x1 − (x1 − x0)·f1 / (f1 − f0)
              print x
              x0 ← x1
              f0 ← f1            ▷ Re-use f1 as f0 in the next iteration
              x1 ← x
         else
              error "division by zero"
```

One difference you might have noticed between the secant method and Newton's method is that the secant method does not use derivatives. The secant method replaces the derivative in Newton's method with the slope of a secant line which approximates the derivative (recall how the tangent line is the limit of slopes of secant lines). This might seem like a drawback, perhaps giving a "less accurate" slope than the tangent line, but in practice it is not really a problem. In fact, in many cases the secant method is preferable, since computing derivatives can often be quite complicated.

Example 4.15

Use the secant method to find the root of $f(x) = \cos x - x$.

Solution: Since the root is already known to be in the interval $[0,1]$, choose $x_0 = 0$ and $x_1 = 1$ as the two initial guesses. The algorithm is easily implemented in the Java programming language. Save this code in a plain text file as `secant.java`:

Listing 4.2 Secant method in Java (`secant.java`)

```java
import java.math.*;
public class secant {
   public static void main(String[] args) {
      int N = Integer.parseInt(args[0]); //Number of iterations
      double x0 = 0.0; //first initial guess
      double x1 = 1.0; //second initial guess
      double f0 = f(x0);
      double f1;
      double x = 0.0;
      for (int i = 2; i <= N; i++) {
         f1 = f(x1);
         x = x1 - (x1 - x0)*f1/(f1 - f0);
         x0 = x1;
         f0 = f1; //Re-use f1 as f0 in the next iteration
         x1 = x;
         System.out.println("n=" + i + ": " + x);
      }
   }

   //Define the function f(x)
   public static double f(double x) {
      return Math.cos(x) - x;
   }
}
```

Compile the code, then run the program:

```
javac secant.java
java secant 10
```

The output is shown below:

```
n=2: 0.6850733573260451
n=3: 0.736298997613654
n=4: 0.7391193619116293
n=5: 0.7390851121274639
n=6: 0.7390851332150012
n=7: 0.7390851332151607
n=8: 0.7390851332151607
n=9: NaN
n=10: NaN
```

Notice that the root was found after 6 iterations ($n = 7$). The undefined number NaN (which stands for "Not a Number") was returned starting with the eighth iteration ($n = 9$) because $x_7 = x_8$, so that $f(x_8) - f(x_7) = 0$, causing a division by zero error in the term

$$x_9 = x_8 - \frac{(x_8 - x_7)\cdot f(x_8)}{f(x_8) - f(x_7)}.$$

For the function $f(x) = \cos x - x$, the table below summarizes the results of 10 iterations of the bisection method, Newton's method and the secant method:

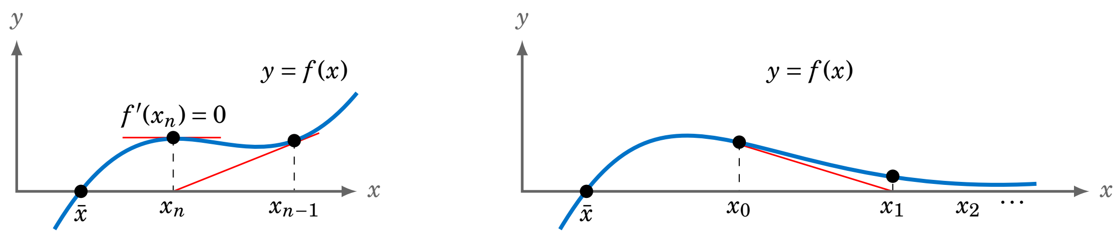
Term	Bisection	Newton	Secant
x_0	0.5	1.0	0.0
x_1	0.75	0.7503638678402439	1.0
x_2	0.625	0.7391128909113617	0.6850733573260451
x_3	0.6875	0.739085133385284	0.736298997613654
x_4	0.71875	0.7390851332151607	0.7391193619116293
x_5	0.734375	0.7390851332151607	0.7390851121274639
x_6	0.7421875	0.7390851332151607	0.7390851332150012
x_7	0.73828125	0.7390851332151607	0.7390851332151607
x_8	0.740234375	0.7390851332151607	0.7390851332151607
x_9	0.7392578125	0.7390851332151607	undefined
x_{10}	0.73876953125	0.7390851332151607	undefined

Newton's method found the root after 4 iterations, while the secant method needed 6 iterations. After 10 iterations the bisection method had yet to find the root to the same level of precision as the other methods—it would take 52 iterations (that is, x_{52}) to achieve similar accuracy to 16 decimal places.

In general Newton's method requires fewer iterations to find a root than the secant method does, but this does not necessarily mean that it will always be faster. Depending on the complexity of the function and its derivative, Newton's method could involve more "expensive" operations (i.e. *computing* values, as opposed to *assigning* values) than the secant method, so that the few more iterations possibly required by the secant method are made up for by fewer total computations.

To see this, notice that Newton's method always requires that both $f(x_{n-1})$ and $f'(x_{n-1})$ be computed for the nth term x_n in the sequence. The secant method needs $f(x_{n-1})$ and $f(x_{n-2})$ for the nth term, but a good programmer would save the value of $f(x_{n-1})$ so that it could be *re-used* (and hence not *re-computed*) as $f(x_{n-2})$ in the next iteration, resulting in potentially fewer total computations for the secant method.

There are occasional pitfalls in using Newton's method. For example, if $f'(x_n) = 0$ for some $n \ge 1$ then Newton's method fails, due to division by 0 in the algorithm. The geometric reason is clear: the tangent line to the curve at that point would be parallel to the x-axis and hence would not intersect it (assuming $f(x_n) \neq 0$). There would be no "next number" x_{n+1} in the iteration! See Figure 4.3.3(a) below.



(a) $f'(x_n) = 0$ **(b)** Moving away from a root

Figure 4.3.3 Newton's method: Potential pitfalls

Another possible problem is that Newton's method might move you *away* from the root, i.e. not get closer, typically by a poor choice of x_0. See Figure 4.3.3(b) above. In some extreme cases, it is possible that Newton's method simply loops back and forth endlessly between the same two numbers, as in Figure 4.3.4:

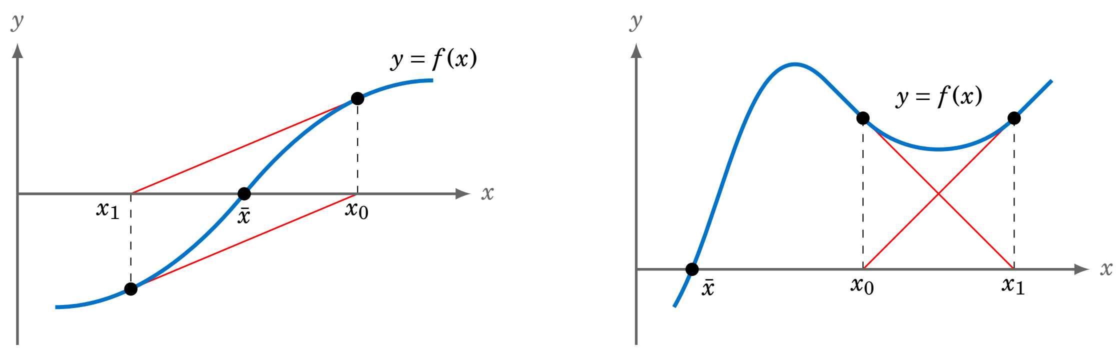


Figure 4.3.4 Newton's method: Infinite loop

In most cases a different choice for the initial guess x_0 will fix such problems. Most textbooks on the subject of *numerical analysis* discuss these issues.[10]

[10] For example, RALSTON, A. AND P. RABINOWITZ, *A First Course in Numerical Analysis*, 2nd ed., New York: McGraw-Hill, Inc., 1978.

There are conditions under which Newton's method is guaranteed to work, and convergence is fast. Newton's method has a *quadratic* rate of convergence, meaning roughly that the *error terms*—the differences between approximate roots and the actual root—are being squared in the long term. More precisely, if the numbers x_n for $n \ge 0$ converge to a root $\bar{x}$, then the error terms $\epsilon_n = x_n - \bar{x}$ for $n \ge 0$ satisfy the limit

$$\lim_{n\to\infty} \frac{|\epsilon_{n+1}|}{|\epsilon_n|^2} = C$$

for some constant C. Squared error terms might sound like a bad thing, but the x_n terms are converging to the root, making the error terms closer to 0 for large n. Squaring a number ϵ_n when $|\epsilon_n| < 1$ results in a smaller number, not a larger one.

The numerical methods that you have learned will make it possible to sketch the graphs of many more functions, since finding local minima and maxima involves finding roots of f', and finding inflection points involves finding roots of f''. You now know some methods for finding those roots.

Finally, despite being slower, the bisection method has the nice advantage of always working. The speed of modern computers makes the difference in algorithmic efficiency negligible in many cases.

Exercises

A

1. Use Newton's method to find the root of $f(x) = \cos x - 2x$.

2. Use Newton's method to find the positive root of $f(x) = \sin x - x/2$.

3. Use Newton's method to find the solution of the equation $e^{-x} = x$.

4. Use Newton's method to find the solution of the equation $e^{-x} = x^2$.

5. Use Newton's method and $f(x) = x^2 - 2$ to approximate $\sqrt{2}$ accurate to six decimal places.

6. Use Newton's method to approximate $\sqrt{3}$ accurate to six decimal places.

7. Repeat Exercise 1 with the secant method.

8. Repeat Exercise 3 with the secant method.

9. Repeat Exercise 5 with the secant method.

10. Repeat Exercise 6 with the secant method.

11. *Cosmic microwave background radiation* is described by a function similar to $f(x) = \frac{x^3}{-1+e^x}$ for $x \ge 0$. Use Newton's method to find the global maximum of f accurate to four decimal places.

12. Would a different choice for x_0 in either graph in Figure 4.3.4 eliminate the infinite loop? Explain.

13. Draw a graph without any symmetry that has the infinite loop problem for Newton's method.

14. The time $t(p)$ required for a p-way merge of a file on a single disk drive into memory is

$$t(p) = \frac{pa + m}{\ln p},$$

where $p > 1$, a is the disk access time, and m is the time to read in one segment of the size of memory. Find the integer closest to the value of p that minimizes $t(p)$ when $a = 24.3$ms and $m = 3500.3$ms.

4.4 The Mean Value Theorem

The difference between instantaneous and average rates of change has been discussed in earlier sections. Recall that there is no difference between the two for linear functions. For nonlinear functions the average rate of change over an interval $[a,b]$ of positive length (i.e. $b-a>0$) will not be the same as the instantaneous rate of change at *every* point in the interval. However, the following theorem guarantees that they will be the same at *some* point in the interval:

Mean Value Theorem: Let a and b be real numbers such that $a<b$, and suppose that f is a function such that

(a) f is continuous on $[a,b]$, and

(b) f is differentiable on (a,b).

Then there is at least one number c in the interval (a,b) such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \tag{4.3}$$

Figure 4.4.1 below shows the geometric interpretation of the theorem:

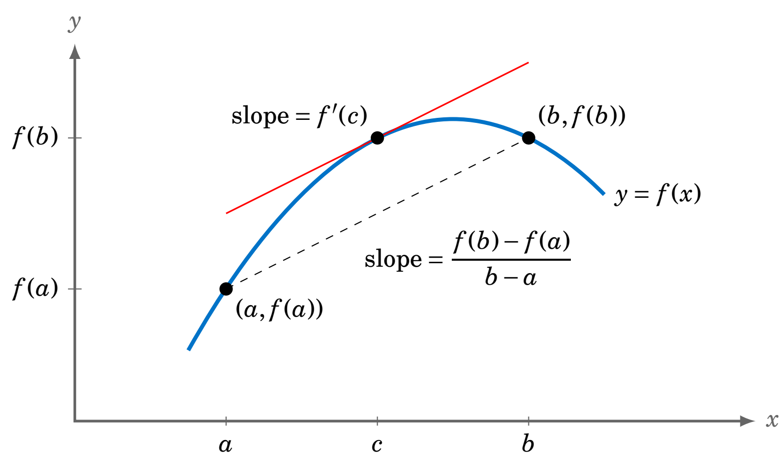


Figure 4.4.1 Mean Value Theorem: parallel tangent line and secant line

The idea is that there is at least one point on the curve $y = f(x)$ where the tangent line will be parallel to the secant line joining the points $(a, f(a))$ and $(b, f(b))$. For each c in (a,b) the tangent line has slope $f'(c)$, while the secant line has slope $\frac{f(b)-f(a)}{b-a}$. The Mean Value Theorem says that these two slopes will be equal somewhere in (a,b).

To prove the Mean Value Theorem (sometimes called *Lagrange's Theorem*), the following intermediate result is needed, and is important in its own right:

Rolle's Theorem: Let a and b be real numbers such that $a < b$, and suppose that f is a function such that
(a) f is continuous on $[a,b]$,

(b) f is differentiable on (a,b), and

(c) $f(a) = f(b) = 0$.

Then there is at least one number c in the interval (a,b) such that $f'(c) = 0$.

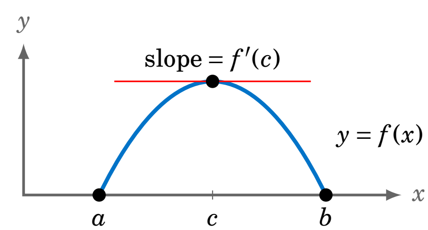


Figure 4.4.2

Figure 4.4.2 on the right shows the geometric interpretation of the theorem. To prove the theorem, assume that f is not the constant function $f(x) = 0$ for all x in $[a,b]$ (if it were then Rolle's Theorem would hold trivially). Then there must be at least one x_0 in (a,b) such that either $f(x_0) > 0$ or $f(x_0) < 0$. If $f(x_0) > 0$ then by the Extreme Value Theorem f attains a global maximum at some $x = c$ in the open interval (a,b), since f is zero at the endpoints $x = a$ and $x = b$ of the closed interval $[a,b]$. Then $f'(c) = 0$ since f has a maximum at $x = c$. Likewise if $f(x_0) < 0$ then f attains a global minimum at some $x = c$ in (a,b), and thus again $f'(c) = 0$. ✓

The Mean Value Theorem can now be proved by applying Rolle's Theorem to the function

$$F(x) \;=\; f(x) \;-\; f(a) \;-\; \frac{f(b)-f(a)}{b-a}(x-a)$$

where f satisfies the conditions of the Mean Value Theorem. Basically, the function F "tilts" the graph of f from Figure 4.4.1 to look like the graph in Figure 4.4.2. It is trivial to check that $F(a) = F(b) = 0$, and F is continuous on $[a,b]$ and differentiable on (a,b) since f is. Thus, by Rolle's Theorem, $F'(c) = 0$ for some c in (a,b). However,

$$F'(x) \;=\; f'(x) \;-\; \frac{f(b)-f(a)}{b-a}$$

and so

$$F'(c) \;=\; f'(c) \;-\; \frac{f(b)-f(a)}{b-a} \;=\; 0 \quad\Rightarrow\quad f'(c) \;=\; \frac{f(b)-f(a)}{b-a} \qquad ✓$$

Note that both the Mean Value Theorem and Rolle's Theorem are purely *existence* theorems; they tell you only that a certain number exists. The task of *finding* the numbers is left to you. For the Mean Value Theorem that task involves solving the equation $f'(x) = \frac{f(b)-f(a)}{b-a}$ (or $f'(x) = 0$ for Rolle's Theorem). The numerical root-finding methods from Section 4.3 could come in handy, since obtaining closed-form solutions might be impossible. For that reason, the Mean Value Theorem is more useful for theoretical purposes. One such application is the following important result:

If f is a differentiable function on an interval I such that $f'(x) = 0$ for all x in I, then f is a constant function on I.

Note that I can be any interval, even the entire real line $(-\infty,\infty)$. It is already known that $f = \text{constant} \Rightarrow f' = 0$; the above result says that the converse is true. The proof is by contradiction: assume that f is not a constant function and show this contradicts the Mean Value Theorem. If f is not constant then there exist numbers $a < b$ in I such that $f(a) \neq f(b)$. However, by the Mean Value Theorem there must exist a number c in the interval (a,b) such that

$$f'(c) = \frac{f(b)-f(a)}{b-a} .$$

Since the derivative of f is 0 everywhere in I, then $f'(c) = 0$ and so

$$\frac{f(b)-f(a)}{b-a} = 0 \quad \Rightarrow \quad f(b) - f(a) = 0 \quad \Rightarrow \quad f(a) = f(b) ,$$

a contradiction of $f(a) \neq f(b)$. Thus, f must be a constant function. ✓

Another theoretical result can be proved with the Mean Value Theorem:

Let f be a differentiable function on an interval I. Then:

(a) If $f' > 0$ on I then f is increasing on I.

(b) If $f' < 0$ on I then f is decreasing on I.

To prove part (a), assume that $f'(x) > 0$ for all x in I, and choose arbitrary numbers a and b in I with $a < b$. To prove that f is increasing on I it suffices to show that $f(a) < f(b)$. By the Mean Value Theorem there is a number c in (a,b) (and hence in I) such that

$$\frac{f(b)-f(a)}{b-a} = f'(c) \quad \text{, and so}$$

$$f(b)-f(a) = (b-a)f'(c) > 0$$

since $b-a > 0$ and $f'(c) > 0$. Thus, $f(b) > f(a)$, and so f is increasing on I. ✓

The proof of part (b) is similar and is left as an exercise. You might wonder why such a proof is necessary. After all, an intuitive explanation was provided in Section 1.2 for why positive or negative derivatives imply that a function is increasing or decreasing, respectively. That knowledge has been assumed and used in the subsequent sections. Intuitive so-called "hand-waving" explanations, in fact, often yield more insight than a "formal" proof, such as the one above. However, it is good to know that such intuition has a solid basis and can be proved, if needed.

The Mean Value Theorem can help in proving inequalities, often used in the sciences for establishing upper or lower bounds on a quantity (e.g. worst-case scenario).

Example 4.16

Show that $\sin x \le x$ for all $x \ge 0$.

Solution: The inequality holds trivially for $x = 0$, since $\sin 0 = 0 \le 0$. So assume that $x > 0$. Then by the Mean Value Theorem there is a number c in $(0, x)$ such that for $f(x) = \sin x$,

$$\frac{f(x) - f(0)}{x - 0} = f'(c) \quad \Rightarrow \quad \frac{\sin x - \sin 0}{x - 0} = \cos c$$

$$\Rightarrow \quad \sin x = x \cos c$$

$$\Rightarrow \quad \sin x \le x$$

since $\cos c \le 1$ and $x > 0$. Note that $\sin x \le x$ is a sharper inequality than $\sin x \le 1$ when $0 < x < 1$.

There is a useful alternative form of the Mean Value Theorem. If $a < b$ then let $h = b - a > 0$, so that $a + h = b$. Then any number c in (a, b) can be written as $c = a + \theta h$ for some number θ in $(0, 1)$. To see this, let c be in (a, b). Then $0 < c - a < b - a = h$ and so $0 < \frac{c-a}{h} < 1$. Thus, $\theta = \frac{c-a}{h}$ is in $(0, 1)$ and $a + \theta h = a + (c - a) = c$. Hence:

Mean Value Theorem (alternative form): Let a and $h > 0$ be real numbers, and suppose that f is a function such that

(a) f is continuous on $[a, a + h]$, and

(b) f is differentiable on $(a, a + h)$.

Then there is a number θ in the interval $(0, 1)$ such that

$$f(a + h) - f(a) = h\, f'(a + \theta h)\,. \tag{4.4}$$

The Mean Value Theorem is the special case of $g(x) = x$ in the following generalization:

Extended Mean Value Theorem: Let a and b be real numbers such that $a < b$, and suppose that f and g are functions such that

(a) f and g are continuous on $[a, b]$,

(b) f and g are differentiable on (a, b), and

(c) $g'(x) \ne 0$ for all x in (a, b).

Then there is at least one number c in the interval (a, b) such that

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}\,. \tag{4.5}$$

The Mean Value Theorem says that the derivative of a differentiable function will always attain one particular value on a closed interval: the function's average rate of change over the interval. It turns out that the derivative will take on *every* value between its values at the endpoints, similar to how the Intermediate Value Theorem applies to continuous functions:[11]

Darboux's Theorem: If f is a differentiable function on a closed interval $[a,b]$ then its derivative f' attains every value between $f'(a)$ and $f'(b)$.

In other words, if $f'(a) < \gamma < f'(b)$ (or $f'(b) < \gamma < f'(a)$) then there is a number c in (a,b) such that $f'(c) = \gamma$. If f' were continuous on $[a,b]$ then the result would follow trivially by the Intermediate Value Theorem for continuous functions. What is perhaps surprising is that Darboux's Theorem holds even for derivatives that are *not* continuous. This means that a discontinuous derivative cannot have the type of simple jump discontinuities that would allow it to "skip" over intermediate values—the points of discontinuity must be of a more complicated type. One rough interpretation of Darboux's Theorem is that even if a derivative is not a continuous function, it will behave *sort of* as if it were.

Exercises

A

1. Does Rolle's Theorem apply to the function $f(x) = 1 - |x|$ on the interval $[-1,1]$? If so, find the number in $(-1,1)$ that Rolle's Theorem guarantees to exist. If not, explain why not.

2. Suppose that two horses run a race starting together and ending in a tie. Show that, at some time during the race, they must have had the same speed.

3. Use the Mean Value Theorem to show that $|\sin A - \sin B| \le |A - B|$ for all A and B (in radians). Does $|\sin A + \sin B| \le |A + B|$ for all A and B? Explain.

4. Show that $|\cos A - \cos B| \le |A - B|$ for all A and B (in radians). Does $|\cos A + \cos B| \le |A + B|$ for all A and B? Explain.

5. Show that $\tan x \ge x$ for all $0 \le x < \frac{\pi}{2}$.

6. Show that $|\tan A - \tan B| \ge |A - B|$ for all A and B (in radians) in $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Can the inequality be extended to all A and B? Explain your answer.

B

7. Use Rolle's Theorem to show that for all constants a and b with $a > 0$, $f(x) = x^3 - ax + b$ can not have three positive roots. Also, show that it can not have three negative roots.

8. Use the Mean Value Theorem to show that if $f' < 0$ on an interval I then f is decreasing on I.

[11] For a proof see pp.16-17 in OSTROWSKI, A.M., *Solution of Equations and Systems of Equations*, 2nd ed., New York: Academic Press Inc., 1966.

9. Suppose that f and g are continuous on $[a,b]$ and differentiable on (a,b), and that $f'(x) > g'(x)$ for all $a < x < b$. Show that $f(b) - g(b) > f(a) - g(a)$.

10. Prove the Extended Mean Value Theorem, by applying Rolle's Theorem to the function

$$F(x) \;=\; f(x) \;-\; f(a) \;-\; \frac{f(b)-f(a)}{g(b)-g(a)}\,(g(x)-g(a))\;.$$

11. Show that $e^x \ge 1+x$ for all x. *(Hint: Consider $f(x) = e^x - x$.)*

12. Show that $\ln(1+x) < x$ for all $x > 0$.

13. Show that $\tan^{-1} x \;<\; x$ for all $x > 0$.

14. Show that for $0 < \alpha \le \beta < \frac{\pi}{2}$,

$$\frac{\beta-\alpha}{\cos^2\alpha} \;\le\; \tan\beta \;-\; \tan\alpha \;\le\; \frac{\beta-\alpha}{\cos^2\beta}\;.$$

15. Show that for $0 < a \le b$,

$$\frac{b-a}{b} \;\le\; \ln\frac{b}{a} \;\le\; \frac{b-a}{a}\;.$$

16. Show that for $n > 1$ and $a > b$,

$$nb^{n-1}(a-b) < a^n - b^n < na^{n-1}(a-b)\;.$$

C

17. Show that $\sqrt{a^2+b} \;<\; a + \dfrac{b}{2a}$ for all positive numbers a and b.

18. Show that $f(x) \;=\; \cos^2 x \;+\; \cos^2\left(\frac{\pi}{3}+x\right) \;-\; \cos x\cos\left(\frac{\pi}{3}+x\right)$ is a constant function. What is its value?

19. Suppose that $f(x)$ is a differentiable function and that $f(0) = 0$ and $f(1) = 1$. Show that $f'(x_0) = 2x_0$ for some x_0 in the interval $(0,1)$.

20. Prove the inequality

$$\left|\frac{x_1+x_2}{1+x_1x_2}\right| \;<\; 1 \quad \text{for } -1 \;<\; x_1, x_2 \;<\; 1$$

as follows:

(a) First prove the special case where $x_1 = x_2$.

(b) For the case $x_1 < x_2$ define

$$f(x) = \frac{x+a}{1+ax}$$

for $-1 \le x \le 1$, where $-1 < a < 1$. Show that f is increasing on $[-1,1]$, then use $a = x_2$ and $x = x_1$.

Note that proving the case $x_2 < x_1$ is unnecessary (why?).
This inequality is a generalization of the same inequality for $0 \le x_1, x_2 < 1$ in the *relativistic velocity addition law* from the theory of special relativity: if object 1 has velocity v_1 relative to a frame of reference F, and if object 2 has a velocity v_2 relative to object 1, so that $x_1 = v_1/c$ and $x_2 = v_2/c$ represent the fractions of the speed of light c at which the objects are moving, then the fraction of the speed of light at which object 2 is moving with respect to F is $x = (x_1+x_2)/(1+x_1x_2)$. So it should be true that $0 \le x < 1$, since nothing can move faster than the speed of light.

CHAPTER 5

The Integral

5.1 The Indefinite Integral

Derivatives appear in many physical phenomena, such as the motion of objects. Recall, for example, that given the position function $s(t)$ of an object moving along a straight line at time t, you could find the velocity $v(t) = s'(t)$ and the acceleration $a(t) = v'(t)$ of the object at time t by taking derivatives. Suppose the situation were reversed: given the velocity function how would you find the position function, or given the acceleration function how would you find the velocity function?

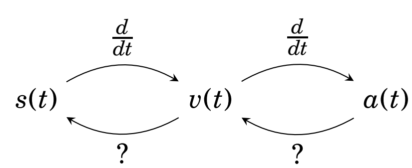


Figure 5.1.1 Differentiation vs antidifferentiation for motion functions

In this case calculating a derivative would not help, since the reverse process is needed: instead of differentiation you need a way of performing **antidifferentiation**, i.e. you would calculate an **antiderivative**.

An **antiderivative** $F(x)$ of a function $f(x)$ is a function whose derivative is $f(x)$. In other words, $F'(x) = f(x)$.

Differentiation is relatively straightforward. You have learned the derivatives of many classes of functions (e.g. polynomials, trigonometric functions, exponential and logarithmic functions), and with the various rules for differentiation you can calculate derivatives of complicated expressions involving those functions (e.g. sums, powers, products, quotients). Antidifferentiation, however, is a different story.

To see some of the issues involved, consider a simple function like $f(x) = 2x$. Of course you know that $\frac{d}{dx}(x^2) = 2x$, so it seems that $F(x) = x^2$ is *the* antiderivative of $f(x) = 2x$. But is it the *only* antiderivative of $f(x)$? No. For example, if $F(x) = x^2 + 1$ then $F'(x) = 2x = f(x)$, and so $F(x) = x^2 + 1$ is another antiderivative of $f(x) = 2x$. Likewise, so is $F(x) = x^2 + 2$. In fact, any function of the form $F(x) = x^2 + C$, where C is some constant, is an antiderivative of $f(x) = 2x$.

Another potential issue is that functions of the form $F(x) = x^2 + C$ are just the most *obvious* antiderivatives of $f(x) = 2x$. Could there be some other completely different function—one that cannot be simplified into the form $x^2 + C$—whose derivative also turns out to be $f(x) = 2x$? The answer, luckily, is no:

> Suppose that $F(x)$ and $G(x)$ are antiderivatives of a function $f(x)$. Then $F(x)$ and $G(x)$ differ only by a constant. That is, $F(x) = G(x) + C$ for some constant C.

To prove this, consider the function $H(x) = F(x) - G(x)$, defined for all x in the common domain I of F and G. Since $F'(x) = G'(x) = f(x)$, then

$$H'(x) = F'(x) - G'(x) = f(x) - f(x) = 0$$

for all x in I, so $H(x)$ is a constant function on I, as was shown in Section 4.4 on the Mean Value Theorem. Thus, there is a constant C such that

$$H(x) = C \quad \Rightarrow \quad F(x) - G(x) = C \quad \Rightarrow \quad F(x) = G(x) + C$$

for all x in I. ✓

The practical consequence of the above result can be stated as follows:

> To find *all* antiderivatives of a function, it is necessary only to find *one* antiderivative and then add a generic constant to it.

So for the function $f(x) = 2x$, since $F(x) = x^2$ is *one* antiderivative then *all* antiderivatives of $f(x)$ are of the form $F(x) = x^2 + C$, where C is a generic constant. Thus, functions do not have just one antiderivative but a whole *family* of antiderivatives, all differing only by a constant. The following notation makes all this easier to express:

> The **indefinite integral** of a function $f(x)$ is denoted by
>
> $$\int f(x)\,dx$$
>
> and represents the entire family of antiderivatives of $f(x)$.

The large S-shaped symbol before $f(x)$ is called an **integral sign**.

Though the indefinite integral $\int f(x)\,dx$ represents *all* antiderivatives of $f(x)$, the integral can be thought of as a single object or function in its own right, whose derivative is $f'(x)$:

$$\frac{d}{dx}\left(\int f(x)\,dx\right) = f(x)$$

You might be wondering what the integral sign in the indefinite integral represents, and why an infinitesimal dx is included. It has to do with what an infinitesimal represents: an infinitesimal "piece" of a quantity. For an antiderivative $F(x)$ of a function $f(x)$, the infinitesimal (or differential) dF is given by $dF = F'(x)\,dx = f(x)\,dx$, and so

$$F(x) = \int f(x)\,dx = \int dF\,.$$

The integral sign thus acts as a summation symbol: it sums up the infinitesimal "pieces" dF of the function $F(x)$ at each x so that they add up to the entire function $F(x)$. Think of it as similar to the usual summation symbol Σ used for *discrete* sums; the integral sign $\int$ takes the sum of a *continuum* of infinitesimal quantities instead.

Finding (or **evaluating**) the indefinite integral of a function is called **integrating** the function, and **integration** is antidifferentiation.

Example 5.1

Evaluate $\displaystyle\int 0\,dx$.

Solution: Since the derivative of any constant function is 0, then $\int 0\,dx = C$, where C is a generic constant.

Note: From now on C will simply be assumed to represent a generic constant, without having to explicitly say so every time.

Example 5.2

Evaluate $\displaystyle\int 1\,dx$.

Solution: Since the derivative of $F(x) = x$ is $F'(x) = 1$, then $\int 1\,dx = x + C$.

Example 5.3

Evaluate $\displaystyle\int x\,dx$.

Solution: Since the derivative of $F(x) = \frac{x^2}{2}$ is $F'(x) = x$, then $\int x\,dx = \frac{x^2}{2} + C$.

Since $\frac{d}{dx}\left(\frac{x^{n+1}}{n+1}\right) = x^n$ for any number $n \neq -1$, and $\frac{d}{dx}(\ln|x|) = \frac{1}{x} = x^{-1}$, then any power of x can be integrated:

Power Formula: $$\int x^n \, dx = \begin{cases} \dfrac{x^{n+1}}{n+1} + C & \text{if } n \neq -1 \\ \ln|x| + C & \text{if } n = -1 \end{cases}$$

The following rules for indefinite integrals are immediate consequences of the rules for derivatives:

Let f and g be functions and let k be a constant. Then:

1. $\displaystyle\int k\,f(x)\,dx = k\int f(x)\,dx$

2. $\displaystyle\int (f(x)+g(x))\,dx = \int f(x)\,dx + \int g(x)\,dx$

3. $\displaystyle\int (f(x)-g(x))\,dx = \int f(x)\,dx - \int g(x)\,dx$

The above rules are easily proved. For example, the first rule is a simple consequence of the Constant Multiple Rule for derivatives: if $F(x) = \int f(x)\,dx$, then

$$\frac{d}{dx}(k\,F(x)) = k\,\frac{d}{dx}(F(x)) = k\,f(x) \quad \Rightarrow \quad \int k\,f(x)\,dx = k\,F(x) = k\int f(x)\,dx\,. \quad \checkmark$$

The other rules are proved similarly and are left as exercises. Repeated use of the above rules along with the Power Formula shows that any polynomial can be integrated term by term—in fact any finite sum of functions can be integrated in that manner:

For any functions $f_1, \ldots, f_n$ and constants $k_1, \ldots, k_n$,

$$\int (k_1 f_1(x) + \cdots + k_n f_n(x))\,dx = k_1\int f_1(x)\,dx + \cdots + k_n\int f_n(x)\,dx\,. \tag{5.1}$$

Example 5.4

Evaluate $\displaystyle\int (x^7 - 3x^4)\,dx$.

Solution: Integrate term by term, pulling constant multiple outside the integral:

$$\int (x^7 - 3x^4)\,dx = \int x^7\,dx - 3\int x^4\,dx = \frac{x^8}{8} - \frac{3x^5}{5} + C$$

Example 5.5

Evaluate $\int \sqrt{x}\,dx$.

Solution: Use the Power Formula:

$$\int \sqrt{x}\,dx = \int x^{1/2}\,dx = \frac{x^{3/2}}{3/2} + C = \frac{2x^{3/2}}{3} + C$$

Example 5.6

Evaluate $\int \left(\frac{1}{x^2} + \frac{1}{x}\right) dx$.

Solution: Use the Power Formula and integrate term by term:

$$\int \left(\frac{1}{x^2} + \frac{1}{x}\right) dx = \int \left(x^{-2} + \frac{1}{x}\right) dx = \frac{x^{-1}}{-1} + \ln|x| + C = -\frac{1}{x} + \ln|x| + C$$

The following indefinite integrals are just re-statements of the corresponding derivative formulas for the six basic trigonometric functions:

$$\int \cos x\,dx = \sin x + C$$

$$\int \sin x\,dx = -\cos x + C$$

$$\int \sec^2 x\,dx = \tan x + C$$

$$\int \sec x \tan x\,dx = \sec x + C$$

$$\int \csc x \cot x\,dx = -\csc x + C$$

$$\int \csc^2 x\,dx = -\cot x + C$$

Since $\frac{d}{dx}(e^x) = e^x$, then:

$$\int e^x\,dx = e^x + C$$

Example 5.7

Evaluate $\int (3\sin x + 4\cos x - 5e^x)\, dx$.

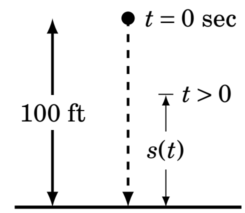
Solution: Integrate term by term:

$$\int (3\sin x + 4\cos x - 5e^x)\, dx = 3\int \sin x\, dx + 4\int \cos x\, dx - 5\int e^x\, dx$$

$$= -3\cos x + 4\sin x - 5e^x + C$$

Example 5.8

Recall from Section 1.1 the example of an object dropped from a height of 100 ft. Show that the height $s(t)$ of the object t seconds after being dropped is $s(t) = -16t^2 + 100$, measured in feet.



Solution: When the object is dropped at time $t = 0$ the only force acting on it is gravity, causing the object to accelerate downward at the known constant rate of 32 ft/s^2. The object's acceleration $a(t)$ at time t is thus $a(t) = -32$. If $v(t)$ is the object's velocity at time t, then $v'(t) = a(t)$, which means that

$$v(t) = \int a(t)\, dt = \int -32\, dt = -32t + C$$

for some constant C. The constant C here is *not* generic—it has a specific value determined by the *initial condition* on the velocity: the object was at rest at time $t = 0$. That is, $v(0) = 0$, which means

$$0 = v(0) = -32(0) + C = C \quad \Rightarrow \quad v(t) = -32t$$

for all $t \geq 0$. Likewise, since $s'(t) = v(t)$ then

$$s(t) = \int v(t)\, dt = \int -32t\, dt = -16t^2 + C$$

for some constant C, determined by the initial condition that the object was 100 ft above the ground at time $t = 0$. That is, $s(0) = 100$, which means

$$100 = s(0) = -16(0)^2 + C = C \quad \Rightarrow \quad s(t) = -16t^2 + 100$$

for all $t \geq 0$. ✓

The formula for $s(t)$ in Example 5.8 can be generalized as follows: denote the object's initial position at time $t = 0$ by s_0, let v_0 be the object's initial velocity (positive if thrown upward, negative if thrown downward), and let g represent the (positive) constant acceleration due to gravity. By Newton's First Law of motion the only acceleration imparted to the object *after* throwing it is due to gravity:

$$a(t) = -g \quad \Rightarrow \quad v(t) = \int a(t)\, dt = \int -g\, dt = -gt + C$$

for some constant C: $v_0 = v(0) = -g(0) + C = C$. Thus, $v(t) = -gt + v_0$ for all $t \geq 0$, and so

$$s(t) = \int v(t)\, dt = \int (-gt + v_0)\, dt = -\tfrac{1}{2}gt^2 + v_0 t + C$$

for some constant C: $s_0 = s(0) = -\frac{1}{2}g(0)^2 + v_0(0) + C = C$. To summarize:

Free fall motion: At time $t \geq 0$:

$$\begin{aligned} \text{acceleration:} \quad a(t) &= -g \\ \text{velocity:} \quad v(t) &= -gt + v_0 \\ \text{position:} \quad s(t) &= -\tfrac{1}{2}gt^2 + v_0 t + s_0 \\ \text{initial conditions:} \quad s_0 &= s(0),\ v_0 = v(0) \end{aligned}$$

Note that the units are not specified—they just need to be consistent. In metric units, $g = 9.8$ m/s^2, while $g = 32$ ft/s^2 in English units.

Thinking of the indefinite integral of a function as the sum of all the infinitesimal "pieces" of that function—for the purpose of retrieving that function—provides a handy way of integrating a differential equation to obtain the solution. The key idea is to transform the differential equation into an *equation of differentials*, which has the effect of treating functions as variables. Some examples will illustrate the technique.

Example 5.9

For any constant k, show that every solution of the differential equation $\frac{dy}{dt} = ky$ is of the form $y = Ae^{kt}$ for some constant A. You can assume that $y(t) > 0$ for all t.

Solution: Put the y terms on the left and the t terms on the right, i.e. separate the variables:

$$\frac{dy}{y} = k\,dt$$

Now integrate both sides (notice how the *function* y is treated as a *variable*):

$$\begin{aligned} \int \frac{dy}{y} &= \int k\,dt \\ \ln y + C_1 &= kt + C_2 \quad (C_1 \text{ and } C_2 \text{ are constants}) \\ \ln y &= kt + C \quad (\text{combine } C_1 \text{ and } C_2 \text{ into the constant } C) \\ y &= e^{kt+C} = e^{kt} \cdot e^C = Ae^{kt} \end{aligned}$$

where $A = e^C$ is a constant. Note that this is the formula for radioactive decay from Section 2.3.

Example 5.10

Recall from Section 3.6 the equation of differentials

$$\frac{dP}{P} + \frac{dV}{V} = \frac{dT}{T}$$

relating the pressure P, volume V and temperature T of an ideal gas. Integrate that equation to obtain the original ideal gas law $PV = RT$, where R is a constant. .

Solution: Integrating both sides of the equation yields

$$\begin{aligned}\int \frac{dP}{P} + \int \frac{dV}{V} &= \int \frac{dT}{T}\\ \ln P + \ln V &= \ln T + C \quad (C \text{ is a constant})\\ \ln(PV) &= \ln T + C\\ PV &= e^{\ln T + C} = e^{\ln T} \cdot e^{C} = T e^{C} = RT\end{aligned}$$

where $R = e^C$ is a constant. ✓

The integration formulas in this section depended on already knowing the derivatives of certain functions and then "working backward" from their derivatives to obtain the original functions. Without that prior knowledge you would be reduced to guessing, or perhaps recognizing a pattern from some derivative you have encountered. A number of integration techniques will be presented shortly, but there are many indefinite integrals for which no simple closed form exists (e.g. $\int e^{x^2}\,dx$ and $\int \sin(x^2)\,dx$).

Exercises

A

For Exercises 1-15, evaluate the given indefinite integral.

1. $\displaystyle\int \left(x^2 + 5x - 3\right) dx$ **2.** $\displaystyle\int 3\cos x\, dx$ **3.** $\displaystyle\int 4e^x\, dx$

4. $\displaystyle\int \left(x^5 - 8x^4 - 3x^3 + 1\right) dx$ **5.** $\displaystyle\int 5\sin x\, dx$ **6.** $\displaystyle\int \frac{3e^x}{5}\, dx$

7. $\displaystyle\int \frac{6}{x}\, dx$ **8.** $\displaystyle\int \frac{4}{3x}\, dx$ **9.** $\displaystyle\int \left(-2\sqrt{x}\right) dx$

10. $\displaystyle\int \frac{1}{3\sqrt{x}}\, dx$ **11.** $\displaystyle\int \left(x + x^{4/3}\right) dx$ **12.** $\displaystyle\int \frac{1}{3\sqrt[3]{x}}\, dx$

13. $\displaystyle\int 3\sec x \tan x\, dx$ **14.** $\displaystyle\int 5\sec^2 x\, dx$ **15.** $\displaystyle\int 7\csc^2 x\, dx$

16. Prove the sum and difference rules for indefinite integrals: $\int (f(x) \pm g(x))\,dx = \int f(x)\,dx \pm \int g(x)\,dx$

17. Integrate both sides of the equation

$$\frac{dP}{P} + \frac{dM}{M} = \frac{dT}{2T}$$

to obtain the ideal gas continuity relation: $\dfrac{PM}{\sqrt{T}} = \text{constant}$.

18. Use the free fall motion equation for position to show that the maximum height reached by an object launched straight up from the ground with an initial velocity v_0 is $\frac{v_0^2}{2g}$.

5.2 The Definite Integral

Recall from the last section that the integral sign in the indefinite integral

$$\int f(x)\,dx$$

represents a summation of the infinitesimals $f(x)\,dx = dF$ for an antiderivative $F(x)$ of $f(x)$. Why is the term "indefinite" used? Because the summation is indefinite: the x in $f(x)\,dx$ is defined generically, meaning "x in general," that is, not for x in a specific range of values. The same summation over a specific, definite range of values of x, say, over an interval $[a,b]$, is a different type of integral:

The **definite integral** of a function $f(x)$ over an interval $[a,b]$ is denoted by

$$\int_a^b f(x)\,dx$$

and represents the sum of the infinitesimals $f(x)\,dx$ for all x in $[a,b]$.

An indefinite integral yields a *generic function*, whereas a definite integral yields either a *number* or a *specific* function. There are many ways to calculate the specific summation in a definite integral, one of which is motivated by a geometric interpretation of the infinitesimal $f(x)\,dx$ as the area of a rectangle, as in Figure 5.2.1 below:

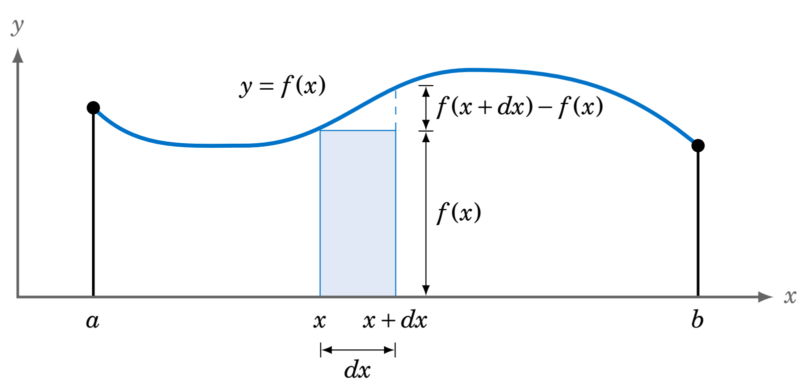


Figure 5.2.1 The infinitesimal $f(x)\,dx$ as the area of a rectangle

The shaded rectangle in the above picture has height $f(x)$ and width dx, and so its area is $f(x)\,dx$. In fact, it appears that that area is just a little bit smaller than the area under the curve $y = f(x)$ and above the x-axis between x and $x + dx$; there is a small gap between the curve and the top of the rectangle, accounting for the difference in the area. However, the area of that gap turns out to be zero, as shown below:

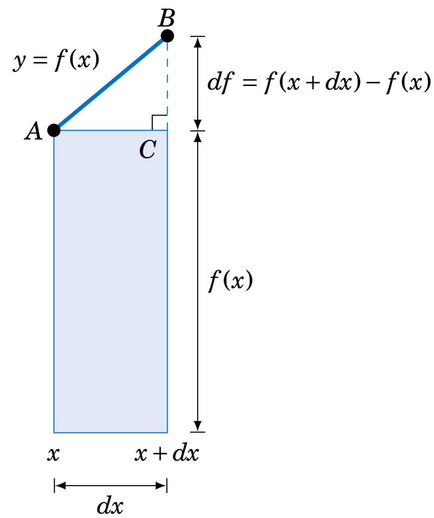


Figure 5.2.2 Area under the curve $y = f(x)$ over $[x, x + dx]$

By the Microstraightness Property, the curve $y = f(x)$ shown in Figure 5.2.1 is a straight line over the infinitesimal interval $[x, x + dx]$, as shown in Figure 5.2.2.[1] Thus, the part of the area between the curve and the x-axis over the interval $[x, x + dx]$ consists of two parts: the area $f(x)\,dx$ of the shaded rectangle and the area of the right triangle $\triangle ABC$, both of which are shown in Figure 5.2.2. However, the area of $\triangle ABC$ is zero:

$$\text{Area of } \triangle ABC \;=\; \frac{1}{2}(\text{base}) \times (\text{height}) \;=\; \frac{1}{2}(dx)(df) \;=\; \frac{1}{2}(dx)(f'(x)\,dx) \;=\; \frac{1}{2}f'(x)(dx)^2 \;=\; 0$$

The function f shown in Figure 5.2.2 is increasing at x, but a similar argument could be made if f were decreasing at x. Hence, the area between the curve $y = f(x)$ and the x-axis comes solely from the rectangles with area $f(x)\,dx$, as x varies from a to b. The sum of all those rectangular areas, though, equals the definite integral of $f(x)$ over $[a, b]$. The definite integral can thus be interpreted as an area:

For a function $f(x) \ge 0$ over $[a, b]$, the **area under the curve** $y = f(x)$ between $x = a$ and $x = b$, denoted by A, is given by

$$A \;=\; \int_a^b f(x)\,dx$$

and represents the area of the region R bounded above by $y = f(x)$, bounded below by the x-axis, and bounded on the sides by $x = a$ and $x = b$ (with $a < b$).

[1] The function f is assumed to be differentiable at x, in this case. If not then the points where f is not differentiable can be excluded without affecting the integral.

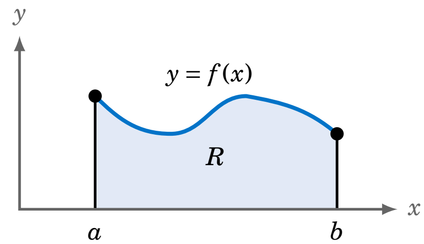


Figure 5.2.3 The area A of the region R equals $\int_a^b f(x)\,dx$

In Figure 5.2.3 the area under the curve $y = f(x)$ between $x = a$ and $x = b$ is the area A of the shaded region R, namely $A = \int_a^b f(x)\,dx$. To calculate that area for a specific function, rectangles can again be used, but this time with widths that are small positive numbers instead of infinitesimals. The procedure is as follows:

1. Create a **partition** $P = \{x_0 < x_1 < \cdots < x_{n-1} < x_n\}$ of the interval $[a,b]$ into $n \ge 1$ subintervals $[x_0,x_1]$, $[x_1,x_2]$, ..., $[x_{n-1},x_n]$, with $x_0 = a$ and $x_n = b$.

2. In each subinterval $[x_{i-1},x_i]$ of P pick a number x_i^*, so that $x_{i-1} \le x_i^* \le x_i$ for $i = 1$ to n.

3. For $i = 1$ to n, form a rectangle whose base is the subinterval $[x_{i-1},x_i]$ of length $\Delta x_i = x_i - x_{i-1} > 0$ and whose height is $f(x_i^*)$.

4. Take the sum $f(x_1^*)\Delta x_1 + f(x_2^*)\Delta x_2 + \cdots + f(x_n^*)\Delta x_n$ of the areas of these rectangles, called a **Riemann sum**.

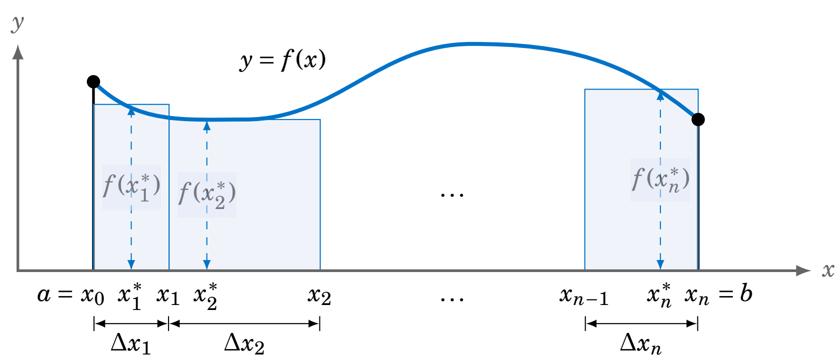


Figure 5.2.4 A Riemann sum of $f(x)$ over a partition of $[a,b]$

5. Take the limit of the Riemann sums as $n \to \infty$, so that the subinterval lengths approach 0. If the limit exists then that limit is the area A of the region R:

$$\text{Area } A \;=\; \int_a^b f(x)\,dx \;=\; \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\,\Delta x_i \tag{5.2}$$

The limit in Formula (5.2) should be taken over all partitions whose **norm**—the length of the largest subinterval—approaches 0. In practice, however, the partitions are usually chosen so that the subintervals are of equal length, and then simply make those equal lengths smaller and smaller by dividing the interval $[a,b]$ into more and more such subintervals. Note that the points x_i^* in each subinterval can be anywhere in the subinterval—often the midpoint of the subinterval is chosen, but the left and right endpoints are also typical choices.

In the above procedure the gaps between the rectangles and the curve will have areas approaching 0 as the number n of subintervals grows and the subinterval lengths approach 0. This is true if the function f is differentiable, and in fact even if f is merely continuous.[2] Thus, the area under the curve can be *defined* by the above procedure.

To calculate the area under a curve in this manner, the reader should have some familiarity with the summation notation in Formula (5.2).

For real numbers $a_1, a_2, \ldots, a_n$ and an integer $n \ge 1$,

$$\sum_{k=1}^{n} a_k = a_1 + a_2 + \cdots + a_n$$

is the sum of $a_1, \ldots, a_n$. The symbol Σ is called the **summation sign**, which is the Greek capital letter Sigma.

The following rules for this "Sigma notation" are intuitively obvious:

Let $a_1, a_2, \ldots, a_n$, and $b_1, b_2, \ldots, b_n$ be real numbers, and let c be a constant. Then:

(1) $$\sum_{k=1}^{n}(a_k + b_k) = \sum_{k=1}^{n} a_k + \sum_{k=1}^{n} b_k$$

(2) $$\sum_{k=1}^{n}(a_k - b_k) = \sum_{k=1}^{n} a_k - \sum_{k=1}^{n} b_k$$

(3) $$\sum_{k=1}^{n} c a_k = c \sum_{k=1}^{n} a_k$$

(4) $$\sum_{k=1}^{n} a_k = \sum_{i=1}^{n} a_i$$ (i.e. the sum is independent of the summation index letter)

[2]For a proof and fuller discussion of all this, see Ch.1-2 in KNOPP, M.I., *Theory of Area*, Chicago: Markham Publishing Co., 1969. The book attempts to define precisely what an "area" actually means, including that of a rectangle (showing agreement with the intuitive notion of width times height).

The following summation formulas can be helpful when calculating Riemann sums:

Let $n \geq 1$ be a positive integer. Then:

(1) $\displaystyle\sum_{k=1}^{n} 1 = n$

(2) $\displaystyle\sum_{k=1}^{n} k = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$

(3) $\displaystyle\sum_{k=1}^{n} k^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$

(4) $\displaystyle\sum_{k=1}^{n} k^3 = 1^3 + 2^3 + \cdots + n^3 = \frac{n^2(n+1)^2}{4}$

(5) $\displaystyle\sum_{k=1}^{n} k^4 = 1^4 + 2^4 + \cdots + n^4 = \frac{n(n+1)(6n^3+9n^2+n-1)}{30}$

Formula (1) is obvious: add the number 1 a total of n times and the sum is n.
Formula (2) can be proved by induction:

1. Show that $\displaystyle\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$ for $n = 1$:

$$\sum_{k=1}^{1} k = 1 = \frac{1(1+1)}{2} \quad \checkmark$$

2. Assume that $\displaystyle\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$ for some integer $n \geq 1$. Show that the formula holds for n replaced by $n+1$, that is:

$$\sum_{k=1}^{n+1} k = \frac{(n+1)((n+1)+1)}{2} = \frac{(n+1)(n+2)}{2}$$

   To show this, note that

$$\sum_{k=1}^{n+1} k = 1 + 2 + \cdots + n + (n+1) = \sum_{k=1}^{n} k + (n+1)$$

$$= \frac{n(n+1)}{2} + (n+1) = \frac{n(n+1) + 2(n+1)}{2} = \frac{(n+1)(n+2)}{2} \quad \checkmark$$

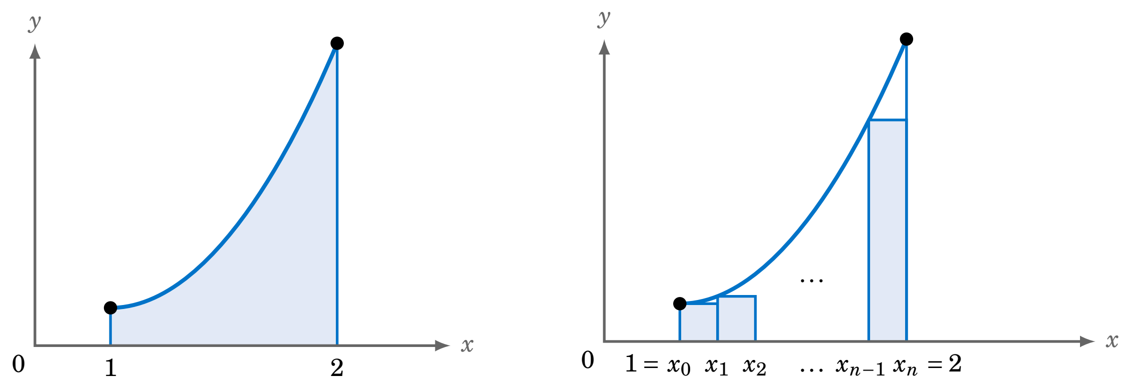
3. By induction, this proves the formula for all integers $n \geq 1$. **QED**

Formulas (3)-(5) can be proved similarly by induction (see the exercises). The example below shows how Formulas (2) and (3) are used in finding the limit of a Riemann sum.

Example 5.11

Use Riemann sums to calculate $\int_1^2 x^2\,dx$.

Solution: The definite integral is the area under the curve $y = f(x) = x^2$ between $x = 1$ and $x = 2$, as shown in Figure 5.2.5(a):



(a) Area under $y = x^2$ over $[1,2]$ **(b)** Riemann sums using left endpoints: $x_i^* = x_{i-1}$

Figure 5.2.5 Calculating $\int_1^2 x^2\,dx$

Divide the interval $[1,2]$ into n subintervals of equal length $\Delta x_i = (2-1)/n = 1/n$ for $i = 1$ to n, so that the partition P is $\{x_0 < x_1 < \ldots x_n\}$ where $x_i = 1 + \frac{i}{n}$ for $i = 0, 1, \ldots, n$ (and hence $x_0 = 1$ and $x_n = 2$). In each subinterval $[x_{i-1}, x_i]$ pick the point x_i^* to be the left endpoint x_{i-1}, so that the rectangles appear as in Figure 5.2.5(b). Then

$$\begin{aligned}
\int_1^2 x^2\,dx &= \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\,\Delta x_i = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_{i-1})\,\frac{1}{n} = \lim_{n\to\infty} \sum_{i=1}^{n} x_{i-1}^2\,\frac{1}{n}\\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \left(1 + \frac{i-1}{n}\right)^2 \frac{1}{n} = \lim_{n\to\infty} \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{2}{n^2}(i-1) + \frac{1}{n^3}(i-1)^2\right)\\
&= \lim_{n\to\infty} \left(\sum_{i=1}^{n} \frac{1}{n} + \frac{2}{n^2}\sum_{i=1}^{n}(i-1) + \frac{1}{n^3}\sum_{i=1}^{n}(i-1)^2\right) = \lim_{n\to\infty} \left(1 + \frac{2}{n^2}\sum_{i=1}^{n-1} i + \frac{1}{n^3}\sum_{i=1}^{n-1} i^2\right)\\
&= \lim_{n\to\infty} \left(1 + \frac{2}{n^2}\cdot\frac{(n-1)n}{2} + \frac{1}{n^3}\cdot\frac{(n-1)n(2n-1)}{6}\right) \quad \text{(replace } n \text{ by } n-1 \text{ in Formulas (2) and (3))}\\
&= \left(\lim_{n\to\infty} 1\right) + \left(\lim_{n\to\infty} \frac{n-1}{n}\right) + \left(\lim_{n\to\infty} \frac{2n^2 - 3n + 1}{6n^2}\right)\\
&= 1 + \frac{1}{1} + \frac{2}{6} = \frac{7}{3}
\end{aligned}$$

It it often simpler to use a computer to calculate approximations of a definite integral, by taking the Riemann sum of a sufficiently large number of rectangles in order to achieve the desired accuracy. Choosing subintervals of equal length, as in Example 5.11, makes it easier to use an algorithm to calculate the integral.

For example, the table below summarizes the calculations of Riemann sums for the function in Example 5.11—namely $f(x) = x^2$ over $[1,2]$—using different values for the points x_i^* in the subintervals (left endpoints, midpoints, and right endpoints):

# of rectangles	Left endpoint	Midpoint	Right endpoint
1	1	2.25	4
2	1.625	2.3125	3.125
3	1.851851851852	2.324074074074	2.851851851852
4	1.96875	2.328125	2.71875
5	2.04	2.33	2.64
10	2.185	2.3325	2.485
100	2.31835	2.333325	2.34835
1000	2.3318335	2.33333325	2.3348335
10000	2.333183335	2.3333333325	2.333483335
100000	2.33331833335	2.333333333325	2.33334833335
1000000	2.333331833333	2.333333333333	2.333334833333

Due to the concavity of the curve $y = x^2$, using the left endpoints underestimates the actual area, whereas using the right endpoints yields an overestimate. Using the midpoints usually gives better results (i.e. more accuracy in fewer iterations).

So far only definite integrals of nonnegative functions have been considered—that is, functions $f(x) \ge 0$ over an interval $[a,b]$. If $f(x)$ is either negative or changes sign over $[a,b]$, then the definite integral can be defined as follows:

Let R be the region bounded by $y = f(x)$ and the x-axis between $x = a$ and $x = b$. If $f(x) \le 0$ over $[a,b]$, then

$$\int_a^b f(x)\,dx \;=\; \text{the negative of the area of } R$$

If $f(x)$ changes sign over $[a,b]$, then

$$\int_a^b f(x)\,dx \;=\; \text{the net area of } R,$$

where the parts of R above the x-axis count as positive area and the parts below count as negative area.

Note: In the definite integral $\int_a^b f(x)\,dx$ the numbers a and b are called the **limits of integration**, with a being the **lower limit of integration** and b the **upper limit of integration**. The function $f(x)$ being integrated is called the **integrand**, in both definite and indefinite integrals.

Exercises

A

1. Explain why $\int_a^b c\,dx = c(b-a)$ for any constant c.

2. Would using left endpoints in the Riemann sums underestimate or overestimate $\int_1^2 \ln x\,dx$? Explain.

B

3. Use Riemann sums to calculate $\int_0^1 x\,dx$.

4. Use Riemann sums to calculate $\int_0^1 x^2\,dx$.

5. Use Riemann sums to calculate $\int_0^1 3x^2\,dx$.

6. Use Riemann sums to calculate $\int_0^1 x^3\,dx$.

7. Prove the formula $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$ by induction on $n \ge 1$.

8. Prove the formula $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$ as follows:

(a) Show that $\sum_{k=1}^{n} \left((k+1)^3 - k^3\right) = (n+1)^3 - 1$.

(b) Show that $(k+1)^3 - k^3 = 3k^2 + 3k + 1$.

(c) Use the formula $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$ and parts (a) and (b) to show that $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$.

9. Prove the formula $\sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4}$ by induction on $n \ge 1$.

10. The famous *quicksort algorithm* in computer science is a popular method for placing objects in some order (e.g. numerical, alphabetical). On average the algorithm needs $O(n \log n)$ comparisons to sort n objects (here $\log n$ means the natural logarithm of n). The proof of that average *complexity* depends on the inequality

$$\sum_{k=2}^{m-1} k \ln k \le \int_2^m x \ln x\,dx$$

for all integers $m > 2$. Explain why that inequality is true.

C

11. Prove the formula $\sum_{k=1}^{n} k^4 = 1^4 + 2^4 + \cdots + n^4 = \frac{n(n+1)(6n^3+9n^2+n-1)}{30}$ by induction on $n \ge 1$.

12. Calculate the following sum:

$$1 + (1+2) + (1+2+3) + (1+2+3+4) + \cdots + (1+2+3+4+\cdots+50)$$

5.3 The Fundamental Theorem of Calculus

Using Riemann sums to calculate definite integrals can be tedious, as was seen in the previous section. In fact the technique shown in that section depended on the function being a low-degree polynomial, which obviously will not always be the case. Luckily there is a better way, involving antiderivatives, given by the following theorem:

Fundamental Theorem of Calculus: Suppose that a function f is differentiable on $[a,b]$. Then:

(I) The function $A(x)$ defined on $[a,b]$ by

$$A(x) = \int_a^x f(t)\,dt$$

is differentiable on $[a,b]$, and

$$A'(x) = f(x)$$

for all x in $[a,b]$.

(II) If F is an antiderivative of f on $[a,b]$, i.e. $F'(x) = f(x)$ for all x in $[a,b]$, then

$$\int_a^b f(x)\,dx = F(b) - F(a)\,.$$

The function $A(x)$ in Part I of the theorem is sometimes called the **area function** because it represents the area under the curve $y = f(x)$ over the interval $[a,x]$, as shown in Figure 5.3.1 below.

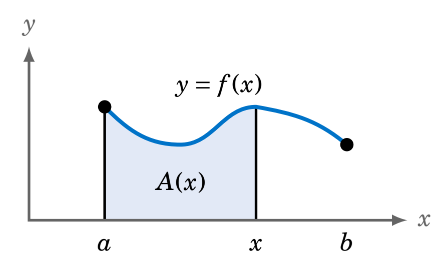


Figure 5.3.1 The area function $A(x) = \int_a^x f(t)\,dt$

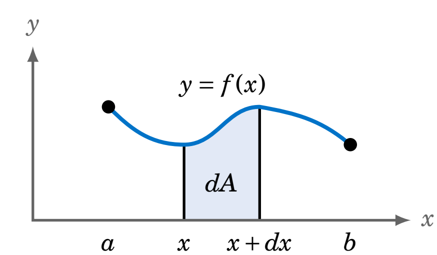
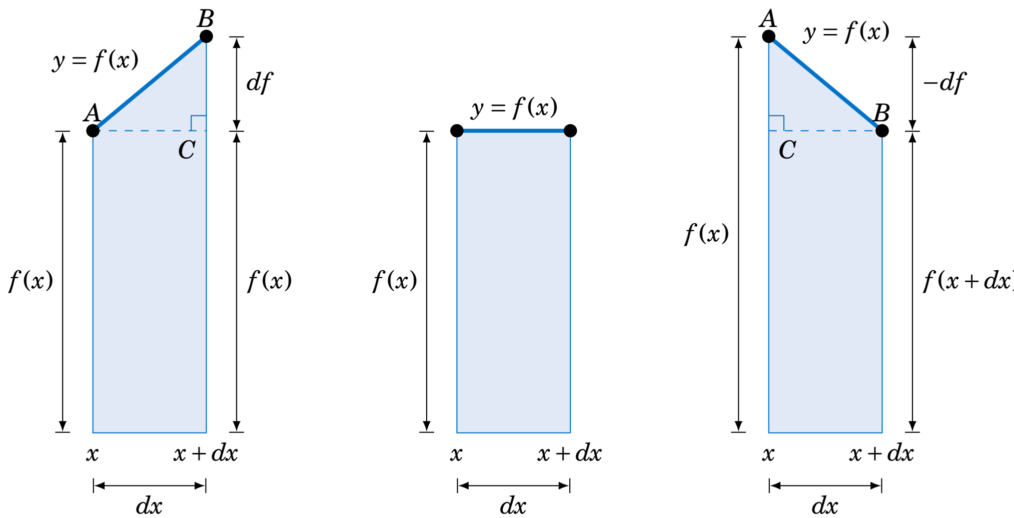


Figure 5.3.2 $dA = A(x+dx) - A(x)$

To prove Part I, assume that $f(x) \ge 0$ on $[a,b]$ as in Figure 5.3.1 (the proofs for $f(x)$ either negative or switching sign over $[a,b]$ are similar). The goal is to show that for any x in $[a,b]$ the differential dA exists and equals $f(x)\,dx$. First, $dA = A(x+dx) - A(x)$ is the area under the curve $y = f(x)$ over the interval $[x, x+dx]$, as shown in Figure 5.3.2 above.

By the Microstraightness Property the curve $y = f(x)$ is a straight line over the infinitesimal interval $[x, x+dx]$, so f must be either increasing, constant, or decreasing over that interval. The three possibilities are shown in Figure 5.3.3:



(a) f is increasing **(b)** f is constant **(c)** f is decreasing

Figure 5.3.3 The three possibilities for dA

In the case where f is increasing over $[x, x+dx]$, the infinitesimal area dA is the sum of the area of the rectangle of height $f(x)$ and width dx and the area of the right triangle $\triangle ABC$ shown in Figure 5.3.3(a). The area of $\triangle ABC$ is $\frac{1}{2}(df)(dx) = \frac{1}{2}f'(x)(dx)^2 = 0$, so $dA = f(x)\,dx$.

In the case where f is constant over $[x, x+dx]$, the infinitesimal area dA is the area of the rectangle of height $f(x)$ and width dx, as shown in Figure 5.3.3(b). So again, $dA = f(x)\,dx$.

In the case where f is decreasing over $[x, x+dx]$, the infinitesimal area dA is the sum of the area of the rectangle of height $f(x+dx)$ and width dx and the area of the right triangle $\triangle ABC$ shown in Figure 5.3.3(c). Note that $df < 0$ since f is decreasing, and so the area of $\triangle ABC$ is $\frac{1}{2}(-df)(dx) = -\frac{1}{2}f'(x)(dx)^2 = 0$. Thus,

$$dA = f(x+dx)\,dx = (f(x)+df)\,dx = f(x)\,dx + f'(x)(dx)^2 = f(x)\,dx + 0 = f(x)\,dx\,.$$

So in all three cases, $dA = f(x)\,dx$, and so $A'(x) = \frac{dA}{dx} = f(x)$, which shows that $A(x)$ is differentiable and has derivative $f(x)$. This proves Part I of the Fundamental Theorem of Calculus. ✓

To prove Part II of the theorem, let $F(x)$ be an antiderivative of $f(x)$ over $[a, b]$. Since the are afunction $A(x) = \int_a^x f(x)\,dx$ is also an antiderivative of $f(x)$ over $[a, b]$ by Part I of the theorem, then $A(x)$ and $F(x)$ differ by a constant C over $[a, b]$. In other words:

$$A(x) = F(x) + C \quad \text{for all } x \text{ in } [a, b]$$

By definition $A(a) = 0$, since it is the area under the curve over the interval $[a,a]$ of zero length. Thus,

$$0 \;=\; A(a) \;=\; F(a) + C \quad\Rightarrow\quad C \;=\; -F(a) \quad\Rightarrow\quad A(x) \;=\; F(x) - F(a) \quad \text{for all } x \text{ in } [a,b]$$

and so

$$\int_a^b f(x)\,dx \;=\; A(b) \;=\; F(b) - F(a)$$

which proves Part II of the theorem.[3] ✓

Note: In some textbooks Part I is called the *First Fundamental Theorem of Calculus* and Part II is called the *Second Fundamental Theorem of Calculus*. The following notation provides a shorthand way of writing $F(b) - F(a)$:

$$F(x)\,\Big|_a^b \;=\; F(b) - F(a)$$

Example 5.12

Calculate $\displaystyle\int_1^2 x^2\,dx$.

Solution: Recall from Example 5.11 in the previous section that the integral equals 7/3. In that example Riemann sums were used, but Part II of the Fundamental Theorem of Calculus makes the integral much easier to calculate. Since $F(x) = \frac{x^3}{3}$ is an antiderivative of $f(x) = x^2$, then

$$\int_1^2 x^2\,dx \;=\; \frac{x^3}{3}\,\Bigg|_1^2 \;=\; \frac{2^3}{3} - \frac{1^3}{3} \;=\; \frac{7}{3}\,.$$

Note in the above example that *any* antiderivative of $f(x) = x^2$ could have been used, e.g. $F(x) = \frac{x^3}{3} + 5$. Notice that the constant 5 cancels out when evaluating $F(2) - F(1)$. So you do *not* need to add a generic constant C to the antiderivative of $f(x)$ in a definite integral, as you would in an indefinite integral.

Example 5.13

Calculate $\displaystyle\int_0^\pi \sin x\,dx$.

Solution: Since $F(x) = -\cos x$ is an antiderivative of $f(x) = \sin x$, then

$$\int_0^\pi \sin x\,dx \;=\; -\cos x\,\Bigg|_0^\pi \;=\; -\cos\pi - (-\cos 0) \;=\; -(-1) - (-1) \;=\; 2\,.$$

[3] The theorem can be proved for the weaker condition that f is merely continuous on $[a,b]$. See p.173-175 in PARZYNSKI, W.R. AND P.W. ZIPSE, *Introduction to Mathematical Analysis*, New York: McGraw-Hill, Inc., 1982.

Example 5.14

Calculate $\int_{-1}^{1} x^3\,dx$.

Solution: Since $F(x) = \frac{x^4}{4}$ is an antiderivative of $f(x) = x^3$, then

$$\int_{-1}^{1} x^3\,dx \;=\; \frac{x^4}{4}\,\Bigg|_{-1}^{1} \;=\; \frac{1^4}{4} - \frac{(-1)^4}{4} \;=\; \frac{1}{4} - \frac{1}{4} \;=\; 0\,.$$

Example 5.14 is a special case of the following result for odd functions:

If f is an odd function, i.e. $f(-x) = -f(x)$ for all x, then

$$\int_{-a}^{a} f(x)\,dx \;=\; 0$$

for all $a > 0$ such that f is continuous on $[-a, a]$.

The idea is that since an odd function is symmetric around the origin, then the area between the curve and the x-axis over $[0, a]$ will cancel out the area between the curve and the x-axis over $[-a, 0]$. Both areas are the same but one gets counted as positive and the other negative, as shown in Figure 5.3.4 below:

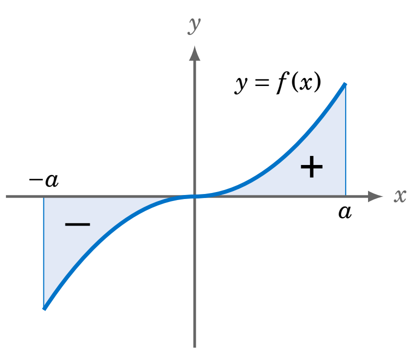


Figure 5.3.4 Odd function f over $[-a, a]$

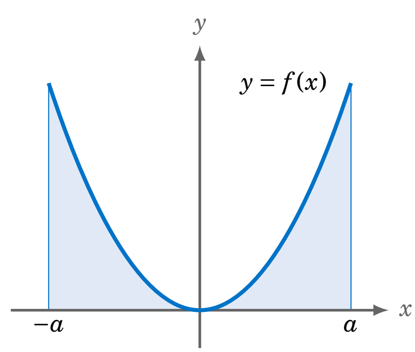


Figure 5.3.5 Even function f over $[-a, a]$

By symmetry around the y-axis, a similar result holds for even functions (see Figure 5.3.5):

If f is an even function, i.e. $f(-x) = f(x)$ for all x, then

$$\int_{-a}^{a} f(x)\,dx \;=\; 2\int_{0}^{a} f(x)\,dx$$

for all $a > 0$ such that f is continuous on $[-a, a]$.

The following rules for definite integrals are a consequence of the corresponding rules for indefinite integrals:

Let f and g be continuous functions on $[a,b]$ and let k be a constant. Then:

1. $\displaystyle\int_a^b k\,f(x)\,dx \;=\; k\int_a^b f(x)\,dx$

2. $\displaystyle\int_a^b (f(x)+g(x))\,dx \;=\; \int_a^b f(x)\,dx \;+\; \int_a^b g(x)\,dx$

3. $\displaystyle\int_a^b (f(x)-g(x))\,dx \;=\; \int_a^b f(x)\,dx \;-\; \int_a^b g(x)\,dx$

The following results for definite integrals are a consequence of the Fundamental Theorem of Calculus:

Let f be a continuous function on $[a,b]$ and suppose that $a<c<b$. Then:

(1) $\displaystyle\int_a^a f(x)\,dx \;=\; 0$

(2) $\displaystyle\int_b^a f(x)\,dx \;=\; -\int_a^b f(x)\,dx$

(3) $\displaystyle\int_a^b f(x)\,dx \;=\; \int_a^c f(x)\,dx \;+\; \int_c^b f(x)\,dx$

For example, if $F(x)$ is an antiderivative of $f(x)$ on $[a,b]$, then

$$\int_a^c f(x)\,dx \;+\; \int_c^b f(x)\,dx \;=\; (F(c)-F(a)) \;+\; (F(b)-F(c)) \;=\; F(b)\;-\;F(a) \;=\; \int_a^b f(x)\,dx$$

which proves rule (3).

The following result is a consequence of Part I of the Fundamental Theorem of Calculus along with the Chain Rule:

Chain Rule for integrals: Let f be a continuous function on an interval I containing $x=a$, and let $g(x)$ be a differentiable function on I. If

$$F(x) \;=\; \int_a^{g(x)} f(t)\,dt \quad \text{for all } x \text{ in } I$$

then $F'(x) = f(g(x))\cdot g'(x)$ for all x in I.

Example 5.15

Let $F(x) = \int_0^{x^2} e^{-t^2}\,dt$ for all $x > 0$. Find $F'(x)$.

Solution: By the Chain Rule for integrals, with $f(t) = e^{t^2}$ and $g(x) = x^2$:

$$F'(x) \;=\; f(g(x))\cdot g'(x) \;=\; e^{-(x^2)^2}\cdot(2x) \;=\; 2x\,e^{-x^4}$$

Exercises

A

For Exercises 1-12, evaluate the given definite integral.

1. $\int_0^1 x^2\,dx$ **2.** $\int_{-1}^1 x^2\,dx$ **3.** $\int_0^1 x^3\,dx$ **4.** $\int_{-1}^1 (x^2+3x-4)\,dx$

5. $\int_1^2 \frac{1}{x^2}\,dx$ **6.** $\int_2^3 \frac{1}{x^3}\,dx$ **7.** $\int_0^{\pi/2} \cos x\,dx$ **8.** $\int_0^1 e^x\,dx$

9. $\int_{-1}^1 2e^x\,dx$ **10.** $\int_{-\pi}^{\pi} \sin x\,dx$ **11.** $\int_0^4 \sqrt{x}\,dx$ **12.** $\int_{-2}^2 \frac{x^3 e^{x^2}}{\cos 2x}\,dx$

13. Show that $\ln x \;=\; \int_1^x \frac{1}{t}\,dt$ for all $x > 0$.

14. Show that $\int_b^a f(x)\,dx \;=\; -\int_a^b f(x)\,dx$.

15. Given $f(x) \;=\; \int_1^x \frac{t}{\sqrt{t^4+1}}\,dt$, find $f'(3)$ and $f'(-2)$.

B

16. Prove the Chain Rule for integrals.

17. Explain why for any continuous function f on $[a,b]$, $\left|\int_a^b f(x)\,dx\right| \;\le\; \int_a^b |f(x)|\,dx$.

18. Explain why if $f(x) \le g(x)$ on $[a,b]$ then $\int_a^b f(x)\,dx \;\le\; \int_a^b g(x)\,dx$.

C

19. Show that if $f(x)$ is continuous on $[a,b]$ then there is a number c in (a,b) such that

$$\int_a^b f(x)\,dx = f(c)\cdot(b-a)\;.$$

20. Let $f(t)$ be a continuous function for all $t \ge 0$, and for each $x \ge 0$ define a function $g(x)$ by

$$g(x) \;=\; \int_0^x (x-t)\,f(t)\,dt\;.$$

Show that $g'(x) \;=\; \int_0^x f(t)\,dt$ for all $x \ge 0$.

21. Show that for all $x > 0$,

$$\int_0^x \frac{dt}{1+t^2} \;+\; \int_0^{1/x} \frac{dt}{1+t^2}$$

is independent of x.

5.4 Integration by Substitution

The integrals encountered so far—whether indefinite or definite—have been the simplest kind, since the antiderivatives had known formulas. For example, $\int \cos x \, dx = \sin x + C$. What if the integral were $\int \cos 2x \, dx$ instead? No formula has been discussed yet for this integral, and the answer is not $\sin 2x + C$, since the derivative of $\sin 2x$ is $2\cos 2x$, not $\cos 2x$. But dividing $\sin 2x$ by 2 first and then taking the derivative *would* yield $\cos 2x$, so that $\int \cos 2x \, dx = \frac{1}{2}\sin 2x + C$.

Evaluating an integral in such a manner is often done when the function is not too complicated, as the one above. Usually it will not be quite that simple, and so a general technique called **substitution** is needed. The idea behind substitution is to replace part of the function being integrated by a new variable—typically u—so that a complicated function of x is now a simpler function of u that you know how to integrate.

Example 5.16

Evaluate $\displaystyle\int \cos 2x \, dx$ by substitution.

Solution: The $2x$ in the cosine function is what makes this integral unknown, so replace it by u: let $u = 2x$. The integral is now

$$\int \cos u \, dx$$

which is a problem because the point of doing substitution is to eliminate all references to the variable x, including in the infinitesimal dx. The entire integral needs to be in terms of u and du, but the dx is still there. So put dx in terms of du:

$$u = 2x \quad \Rightarrow \quad du = 2dx \quad \Rightarrow \quad dx = \frac{1}{2}du$$

The integral now becomes

$$\int \cos u \left(\frac{1}{2}du\right) = \frac{1}{2}\int \cos u \, du = \frac{1}{2}\sin u + C$$

by the formula already known, just with the letter u as the variable instead of x. The original integral was in terms of x, so the final answer—for an indefinite integral—should also be in terms of x. Thus, the final step is to substitute back into the answer what u equals in terms of x, namely $2x$:

$$\int \cos 2x \, dx = \frac{1}{2}\sin u + C = \frac{1}{2}\sin 2x + C$$

If the procedure in the above example seems similar to making a substitution when using the Chain Rule to take a derivative, that is because it is similar: you are basically doing the same thing only in reverse. Just as for differentiation, it is not always obvious what part of the function is the best candidate for substitution when performing integration. There is one obvious rule: *never* make the substitution $u = x$, because that changes nothing.

Example 5.17

Evaluate $\int e^{-3x}\,dx$.

Solution: The $-3x$ in the exponential function is what makes this integral unknown, so make the substitution $u = -3x$, which means that $du = -3dx$, and so $dx = -\frac{1}{3}du$. Thus:

$$\int e^{-3x}\,dx \;=\; \int e^{u}\left(-\frac{1}{3}du\right) \;=\; -\frac{1}{3}\int e^{u}\,du \;=\; -\frac{1}{3}e^{u} + C \;=\; -\frac{1}{3}e^{-3x} + C$$

The above example can be generalized as follows:

$$\int e^{ax}\,dx \;=\; \frac{1}{a}e^{ax} + C \quad \text{for any constant } a \neq 0$$

Example 5.17 was the special case with $a = -3$.

Example 5.18

Evaluate $\int (1+4x)^5\,dx$.

Solution: You might be tempted to make the substitution $u = 4x$, but that would then require finding the integral of $(1+u)^5$, for which there is not yet any formula. But there *is* a formula for the integral of u^5. Hence, let $u = 1+4x$, so that $du = 4dx \;\Rightarrow\; dx = \frac{1}{4}du$. Thus:

$$\int (1+4x)^5\,dx \;=\; \frac{1}{4}\int u^5\,du \;=\; \frac{1}{4}\frac{u^6}{6} + C \;=\; \frac{1}{24}(1+4x)^6 + C$$

Example 5.19

Evaluate $\int 2x\,e^{x^2}\,dx$.

Solution: It might be unclear whether you should make the substitution $u = 2x$ or $u = x^2$, but the hint here is that the derivative of the x^2 inside the exponential function is $2x$, which appears outside the exponential function. Indeed, you could check that letting $u = 2x$ would result in an integral no simpler than the current one (namely, $\frac{1}{2}\int u\,e^{u^2/4}\,du$). So let $u = x^2$, which means $du = 2x\,dx$. Thus:

$$\int 2x\,e^{x^2}\,dx \;=\; \int e^{u}\,du \;=\; e^{u} + C \;=\; e^{x^2} + C$$

Example 5.20

Evaluate $\int x\sqrt{1+3x^2}\,dx$.

Solution: Note that the derivative of the $1+3x^2$ term inside the square root function is $6x$, which is *almost* the function x outside the square root—all that is missing is the constant multiple 6. That is a hint to let $u = 1+3x^2$. Notice also that $du = 6x\,dx \;\Rightarrow\; x\,dx = \frac{1}{6}du$, and $x\,dx$ is the remaining part of the integral outside the square root. Thus:

$$\int x\sqrt{1+3x^2}\,dx \;=\; \frac{1}{6}\int \sqrt{u}\,du \;=\; \frac{1}{6}\frac{u^{3/2}}{3/2} + C \;=\; \frac{1}{9}(1+3x^2)^{3/2} + C$$

Example 5.21

Evaluate $\int \frac{x^2\,dx}{\sqrt{x^3+9}}$.

Solution: Let $u = x^3+9$, so that $du = 3x^2\,dx \Rightarrow x^2\,dx = \frac{1}{3}du$. Thus:

$$\int \frac{x^2\,dx}{\sqrt{x^3+9}} = \int \frac{\frac{1}{3}du}{\sqrt{u}} = \frac{1}{3}\int u^{-1/2}\,du = \frac{1}{3}\frac{u^{1/2}}{1/2} + C = \frac{2}{3}\sqrt{x^3+9} + C$$

Example 5.22

Evaluate $\int \frac{2x\,dx}{x^2-1}$.

Solution: Notice that the numerator $2x$ in the function is exactly the derivative of the denominator x^2-1. That is a hint to substitute on the denominator so that the integral is the natural logarithm function. Let $u = x^2-1$, so that $du = 2x\,dx$. Thus:

$$\int \frac{2x\,dx}{x^2-1} = \int \frac{du}{u} = \ln|u| + C = \ln|x^2-1| + C$$

Example 5.23

Evaluate $\int \tan x\,dx$.

Solution: Notice that $\tan x = \frac{\sin x}{\cos x}$ and that the numerator $\sin x$ is *almost* the derivative of the denominator $\cos x$; all that is missing is a negative sign. That is a hint to substitute on the denominator: $u = \cos x$, so that $du = -\sin x\,dx \Rightarrow \sin x\,dx = -du$. Thus:

$$\int \tan x\,dx = \int \frac{\sin x}{\cos x}\,dx$$
$$= \int \frac{-du}{u} = -\ln|u| + C = -\ln|\cos x| + C = \ln|\cos x|^{-1} + C = \ln|\sec x| + C$$

Example 5.24

Evaluate $\int \sec x\,dx$.

Solution: Notice that

$$\int \sec x\,dx = \int \frac{\sec x\,(\sec x + \tan x)}{\sec x + \tan x}\,dx = \int \frac{\sec x \tan x + \sec^2 x}{\sec x + \tan x}\,dx$$

and that the numerator in the last integral is the derivative of the denominator: let $u = \sec x + \tan x$, so that $du = (\sec x \tan x + \sec^2 x)\,dx$. Thus:

$$\int \sec x\,dx = \int \frac{du}{u} = \ln|u| + C = \ln|\sec x + \tan x| + C$$

The following formulas are straightforward consequences of substitution and the derivatives of inverse trigonometric functions discussed in Section 2.2:

For any constant $a > 0$:

$$\int \frac{dx}{\sqrt{a^2 - x^2}} = \sin^{-1}\left(\frac{x}{a}\right) + C \quad (\text{if } |x| < a) \tag{5.3}$$

$$\int \frac{dx}{a^2 + x^2} = \frac{1}{a}\tan^{-1}\left(\frac{x}{a}\right) + C \tag{5.4}$$

$$\int \frac{dx}{|x|\sqrt{x^2 - a^2}} = \frac{1}{a}\sec^{-1}\left(\frac{x}{a}\right) + C \quad (\text{if } |x| > a) \tag{5.5}$$

For example, to prove the second formula, recall that $\frac{d}{dx}\left(\tan^{-1} x\right) = \frac{1}{1+x^2}$. Make the substitution $u = x/a$, so that $x = au$ and $dx = a\,du$. Thus:

$$\int \frac{dx}{a^2 + x^2} = \int \frac{a\,du}{a^2 + a^2u^2} = \frac{1}{a}\int \frac{du}{1+u^2} = \frac{1}{a}\tan^{-1} u + C = \frac{1}{a}\tan^{-1}\left(\frac{x}{a}\right) + C \quad \checkmark$$

Example 5.25

Evaluate $\displaystyle\int \frac{dx}{\sqrt{4 - 9x^2}}$.

Solution: The $4 - 9x^2$ inside the square root is almost of the form $a^2 - x^2$, except for the 9. The goal is to have $u^2 = 9x^2$, so let $u = 3x$, which means that $dx = \frac{1}{3}\,du$ and $u^2 = 9x^2$. Thus,

$$\int \frac{dx}{\sqrt{4 - 9x^2}} = \frac{1}{3}\int \frac{du}{\sqrt{4 - u^2}} = \frac{1}{3}\sin^{-1}\left(\frac{u}{2}\right) + C = \frac{1}{3}\sin^{-1}\left(\frac{3x}{2}\right) + C$$

by Formula (5.2) with $a = 2$.

To use substitution with definite integrals, follow the same procedure as with indefinite integrals but add one extra step: replace the limits of integration $x = a$ and $x = b$ in the original integral $\int_a^b f(x)\,dx$ by $u = g(a)$ and $u = g(b)$, respectively, in the new integral involving u, where $u = g(x)$ is your substitution.

Example 5.26

Evaluate $\displaystyle\int_1^2 (2x+1)^3\,dx$.

Solution: Let $u = g(x) = 2x+1$, which means that $dx = \frac{1}{2}\,du$. The upper limit of integration $x = 2$ becomes $u = g(2) = 2(2)+1 = 5$ in the new u-based integral, while the lower limit of integration $x = 1$ becomes $u = g(1) = 2(1)+1 = 3$. Thus:

$$\int_1^2 (2x+1)^3\,dx = \frac{1}{2}\int_3^5 u^3\,du = \frac{1}{8}u^4\Big|_3^5 = \frac{1}{8}(5^4 - 3^4) = 68$$

Note that you could have put everything back in terms of x at the end, but there was no need to since you would get the same numerical answer.

The following property of definite integrals comes in handy for evaluating certain types of definite integrals:

For any constant a,

$$\int_0^a f(x)\,dx \;=\; \int_0^a f(a-x)\,dx\,. \tag{5.6}$$

This is simple to prove, using the substitution $u = a-x$, so $x = a-u$ and $dx = -du$, while $x = 0$ becomes $u = a$ and $x = a$ becomes $u = 0$ in the limits of integration:

$$\int_0^a f(x)\,dx \;=\; -\int_a^0 f(a-u)\,du \;=\; \int_0^a f(a-u)\,du \;=\; \int_0^a f(a-x)\,dx \quad \checkmark$$

Example 5.27

Evaluate $\displaystyle\int_0^{\pi} \frac{x \sin x}{1+\cos^2 x}\,dx$.

Solution: Let $I = \displaystyle\int_0^{\pi} \frac{x \sin x}{1+\cos^2 x}\,dx$. Then by the above property (with $a = \pi$):

$$I \;=\; \int_0^{\pi} \frac{(\pi-x)\sin(\pi-x)}{1+\cos^2(\pi-x)}\,dx \;=\; \int_0^{\pi} \frac{(\pi-x)\sin x}{1+\cos^2 x}\,dx \;=\; \pi\int_0^{\pi} \frac{\sin x}{1+\cos^2 x}\,dx \;-\; \int_0^{\pi} \frac{x\sin x}{1+\cos^2 x}\,dx$$

$$I \;=\; \pi\int_0^{\pi} \frac{\sin x}{1+\cos^2 x}\,dx \;-\; I$$

$$2I \;=\; \pi\int_0^{\pi} \frac{\sin x}{1+\cos^2 x}\,dx \;=\; -\pi\tan^{-1}(\cos x)\,\Bigg|_0^{\pi} \;=\; -\pi\left(-\frac{\pi}{4}-\frac{\pi}{4}\right) \;=\; \frac{\pi^2}{2}$$

$$I \;=\; \frac{\pi^2}{4}$$

Exercises

A

For Exercises 1-24 evaluate the given integral.

1. $\displaystyle\int (3\cos 5x \;+\; 4\sin 5x)\,dx$

2. $\displaystyle\int \frac{e^{2x} \;+\; e^{-2x}}{2}\,dx$

3. $\displaystyle\int \left(xe^{-x^2} \;+\; x^2\cos x^3\right)dx$

4. $\displaystyle\int \frac{x \;-\; 2}{x^2 \;-\; 4x \;+\; 9}\,dx$

5. $\displaystyle\int \frac{e^x}{1 \;+\; e^x}\,dx$

6. $\displaystyle\int \frac{1}{1 \;+\; e^x}\,dx$

7. $\displaystyle\int x\sqrt{x+4}\,dx$

8. $\displaystyle\int \cos^2 x\,dx$

9. $\displaystyle\int \tan^2 x\,dx$

10. $\displaystyle\int \frac{3}{\sqrt{4 \;-\; 25x^2}}\,dx$

11. $\displaystyle\int \frac{3}{4 \;+\; 25x^2}\,dx$

12. $\displaystyle\int \sin^2 x\,\cos^3 x\,dx$

13. $\int_0^1 (2x+1)^3\,dx$

14. $\int_0^1 (2x-1)^3\,dx$

15. $\int_0^8 x\sqrt{1+x}\,dx$

16. $\int_0^{\pi/2} 4\sin(x/2)\,dx$

17. $\int_0^{\pi/4} 4\sin x\,\cos x\,dx$

18. $\int_0^{\sqrt{\pi}} 5x\cos(x^2)\,dx$

19. $\int_{-2}^{-1} \frac{x}{(x^2+2)^3}\,dx$

20. $\int_{-\ln 3}^{\ln 3} \frac{e^x}{e^x+4}\,dx$

21. $\int_1^3 \frac{dx}{\sqrt{x}\,(x+1)}$

22. $\int_{-1}^1 \frac{x^2\,dx}{\sqrt{x^3+9}}$

23. $\int_1^2 \frac{dx}{x^2-6x+9}$

24. $\int_{-3}^3 \frac{x^5\,dx}{e^{x^2}}$

25. Evaluate the indefinite integral

$$\int \sin x\,\cos x\,dx$$

three different ways:

(a) Use the substitution $u = \sin x$.

(b) Use the substitution $u = \cos x$.

(c) Use the trigonometric identity $2\sin x\,\cos x = \sin 2x$.

(d) Are the three answers from parts (a)-(c) actually different? Explain.

26. For all positive constants L, show the following:

(a) $\int_0^L \left(1 - \frac{x}{L}\right)^2 dx = \frac{L}{3}$

(b) $\int_{-L/2}^{L/2} \left(\frac{1}{2} - \frac{x}{L}\right)^2 dx = \frac{L}{3}$

(c) $\int_0^L \left(1 - \frac{x}{L}\right)^3 \left(\frac{x}{L}\right)^2 dx = \frac{L}{60}$

B

27. Recall from trigonometry that $\sin^2 x = \frac{1}{2}(1 - \cos 2x)$ for all x.

(a) Use the Fundamental Theorem of Calculus to evaluate $\int_0^\pi \sin^2 x\,dx$.

(b) Approximate the integral from part (a) by dividing the interval $[0,\pi]$ into $n = 2$ subintervals of equal length, $[0,\pi/2]$ and $[\pi/2,\pi]$, and finding the **exact** value of the sum of the areas of the rectangles whose heights are determined at the right endpoints of the subintervals.

(c) Repeat part (b) with $n = 3$.

(d) Repeat part (b) with $n = 4$.

(e) Repeat part (b) with $n = 6$.

28. Show that $\int \csc x\,dx = -\ln|\csc x + \cot x| + C$. *(Hint: See Example 5.24.)*

C

29. Use the property $\int_0^a f(x)\,dx = \int_0^a f(a-x)\,dx$ to show that

$$\int_0^{\pi/2} \frac{\sin^2 x}{\sin x + \cos x}\,dx = \frac{1}{\sqrt{2}} \ln\left(\sqrt{2}+1\right) .$$

(Hint: Use Exercise 28 and the sine addition formula.)

5.5 Improper Integrals

Definite integrals so far have been defined only for continuous functions over finite closed intervals. There are times when you will need to perform integration despite those conditions not being met. For example, in quantum mechanics the *Dirac delta function*[4] δ is defined on $\mathbb{R}$ by four properties:

(1): $\delta(x) = 0$ for all $x \neq 0$

(2): $\delta(0) = \infty$

(3): $\int_{-\infty}^{\infty} \delta(x)\,dx = 1$

(4): For any continuous function f on $\mathbb{R}$,

$$\int_{-\infty}^{\infty} f(x)\,\delta(x)\,dx = f(0).$$

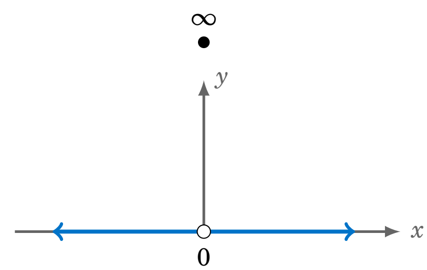


Figure 5.5.1 $y = \delta(x)$

Properties (3) and (4) provide examples of one type of **improper integral**: an integral over an infinite interval (in this case the entire real line $\mathbb{R} = (-\infty,\infty)$). Define this type of improper integral as follows:

For a continuous function f and a real number a, define the **improper integral** of f over $[a,\infty)$ by

$$\int_{a}^{\infty} f(x)\,dx = \lim_{b\to\infty} \int_{a}^{b} f(x)\,dx\,,$$

define the improper integral of f over $(-\infty,a]$ by

$$\int_{-\infty}^{a} f(x)\,dx = \lim_{b\to-\infty} \int_{b}^{a} f(x)\,dx\,,$$

and define the improper integral of f over $(-\infty,\infty)$ by

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{c} f(x)\,dx + \int_{c}^{\infty} f(x)\,dx\,,$$

for any real number c (typically $c = 0$). If the given limit exists (i.e. is a real number) then the improper integral is **convergent**; otherwise it is **divergent**.

[4]Created by the physicist P.A.M. Dirac (1902-1984), who won a Nobel Prize in physics in 1933. The function is neither real-valued nor continuous at $x = 0$. The "graph" in Figure 5.5.1 is perhaps misleading, as ∞ is not an actual point on the y-axis. One interpretation is that δ is an abstraction of an instantaneous pulse or burst of *something*, preceded and followed by *nothing*. To learn more about this fascinating and useful function, see §15 in DIRAC, P.A.M., *The Principles of Quantum Mechanics*, 4th ed., Oxford, UK: Oxford University Press, 1958.

The limits in the above definitions are always taken *after* evaluating the integral inside the limit. Just as for "proper" definite integrals, improper integrals can be interpreted as representing the area under a curve.

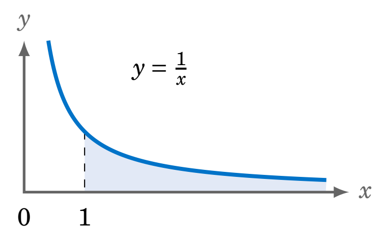
Example 5.28

Evaluate $\displaystyle\int_1^\infty \frac{dx}{x}$.

Solution: For all real numbers $b > 1$,

$$\int_1^\infty \frac{dx}{x} = \lim_{b\to\infty} \int_1^b \frac{dx}{x} = \lim_{b\to\infty} \left(\ln x \,\Bigg|_1^b \right)$$

$$= \lim_{b\to\infty} (\ln b - \ln 1) = \lim_{b\to\infty} b = \infty$$



and so the integral is divergent. This means that the area under the curve $y = 1/x$ over the interval $[1,\infty)$—as shown in the graph above—is infinite.

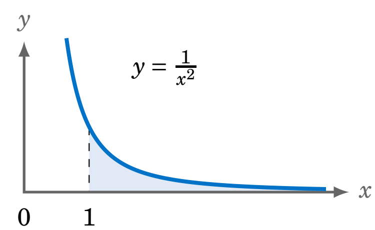
Example 5.29

Evaluate $\displaystyle\int_1^\infty \frac{dx}{x^2}$.

Solution: For all real numbers $b > 1$,

$$\int_1^\infty \frac{dx}{x^2} = \lim_{b\to\infty} \int_1^b \frac{dx}{x^2} = \lim_{b\to\infty} \left(-\frac{1}{x} \,\Bigg|_1^b \right)$$

$$= \lim_{b\to\infty} \left(-\frac{1}{b} - \left(-\frac{1}{1}\right) \right) = \lim_{b\to\infty} \left(1 - \frac{1}{b} \right) = 1 - 0 = 1 .$$



This means that the area under the curve $y = 1/x^2$ over the interval $[1,\infty)$—as shown in the graph above—equals 1. Thus, an infinite region can have a finite area. Length and area are different and not necessarily related concepts, as this example illustrates. Notice that $y = 1/x^2$ approaches the x-axis asymptote much faster than $y = 1/x$ does—fast enough to make the integral convergent.

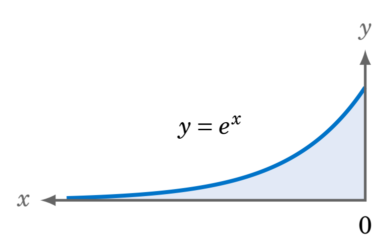
Example 5.30

Evaluate $\displaystyle\int_{-\infty}^0 e^x \, dx$.

Solution: For all real numbers $b < 0$,

$$\int_{-\infty}^0 e^x \, dx = \lim_{b\to-\infty} \int_b^0 e^x \, dx = \lim_{b\to-\infty} \left(e^x \,\Bigg|_b^0 \right)$$

$$= \lim_{b\to-\infty} (1 - e^b) = 1 - 0 = 1 .$$

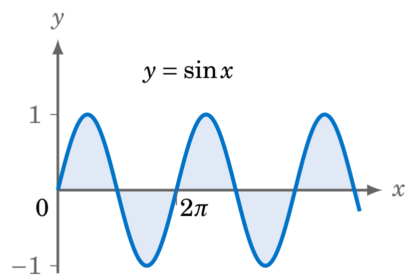


This means that the area under the curve $y = e^x$ over the interval $(-\infty, 0]$—as shown in the graph above—equals 1.

Example 5.31

Evaluate $\displaystyle\int_0^\infty \sin x\,dx$.

Solution: Since



$$\int_0^\infty \sin x\,dx = \lim_{b\to\infty}\int_0^b \sin x\,dx$$

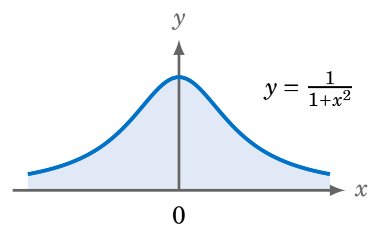
$$= \lim_{b\to\infty}\left(-\cos x\,\Big|_0^b\right) = \lim_{b\to\infty}\,(-\cos b + 1)$$

then the integral is divergent, since $\lim_{b\to\infty}\cos b$ does not exist ($\cos b$ oscillates between 1 and -1). This means that the net area over $[0,\infty)$—counted as positive above the x-axis and negative below—is indeterminate.

Example 5.32

Evaluate $\displaystyle\int_{-\infty}^\infty \frac{dx}{1+x^2}$.

Solution: Split the integral at $x = 0$:



$$\int_{-\infty}^\infty \frac{dx}{1+x^2} = \int_{-\infty}^0 \frac{dx}{1+x^2} + \int_0^\infty \frac{dx}{1+x^2}$$

$$= \left(\lim_{b\to-\infty}\int_b^0 \frac{dx}{1+x^2}\right) + \left(\lim_{b\to\infty}\int_0^b \frac{dx}{1+x^2}\right)$$

$$= \left(\lim_{b\to-\infty}\tan^{-1}x\,\Big|_b^0\right) + \left(\lim_{b\to\infty}\tan^{-1}x\,\Big|_0^b\right)$$

$$= \lim_{b\to-\infty}\,(\tan^{-1}0 - \tan^{-1}b) + \lim_{b\to\infty}\,(\tan^{-1}b - \tan^{-1}0)$$

$$= (0-(-\pi/2)) + (\pi/2-0) = \pi$$

This means that the area under the curve $y = \frac{1}{1+x^2}$ over the entire real line $(-\infty,\infty)$—as shown in the graph above—equals π. Note that if the integral were split at any number c then the answer would be the same. Another way to evaluate the integral would have been to use the symmetry around the y-axis—as $f(x) = \frac{1}{1+x^2}$ is an even function—so that

$$\int_{-\infty}^\infty \frac{dx}{1+x^2} = 2\int_0^\infty \frac{dx}{1+x^2} = \cdots = 2(\pi/2-0) = \pi\ .$$

Since the integrand is continuous over $\mathbb{R}$, a common way of evaluating the integral—especially among students—is to simply use $\pm\infty$ as actual limits of integration, thus avoiding the need to take a limit:

$$\int_{-\infty}^\infty \frac{dx}{1+x^2} = \tan^{-1}x\,\Big|_{-\infty}^{\infty} = \tan^{-1}(\infty) - \tan^{-1}(-\infty) = \frac{\pi}{2} - \frac{-\pi}{2} = \pi$$

This type of shortcut is fine as long as you are aware of what plugging $x = \pm\infty$ into $\tan^{-1}x$ actually means, and that there are no numbers for which the integrand is undefined (which would yield an improper integral of a different type, to be discussed shortly).

The second type of improper integral is of a function not continuous or not bounded over its interval of integration. For example, the integral in property (3) of the Dirac delta function is of that type, since δ is discontinuous at $x = 0$. Define this type of improper integral as follows:

For a function f that is continuous on $[a, b)$ but has either a discontinuity or vertical asymptote at $x = b$, define the **improper integral** of f over $[a, b)$ by

$$\int_a^b f(x)\,dx \;=\; \lim_{c\to b-} \int_a^c f(x)\,dx\ .$$

Likewise, if f is continuous on $(a, b]$ but has either a discontinuity or vertical asymptote at $x = a$, then define the improper integral of f over $(a, b]$ by

$$\int_a^b f(x)\,dx \;=\; \lim_{c\to a+} \int_c^b f(x)\,dx\ .$$

If f is continuous on $[a, b]$ but has either a discontinuity or vertical asymptote at $x = c$ for $a < c < b$, then define the improper integral of f over $[a, b]$ by

$$\int_a^b f(x)\,dx \;=\; \int_a^c f(x)\,dx \;+\; \int_c^b f(x)\,dx\ ,$$

where the integrals on the right are evaluated as in the first two definitions. If the given limit exists (i.e. is a real number) then the improper integral is **convergent**; otherwise it is **divergent**.

Adjust these definitions accordingly for infinite intervals—e.g. $[a, \infty)$, $(-\infty, b]$, or $(-\infty, \infty)$—to be consistent with the definitions of improper integrals of that type.

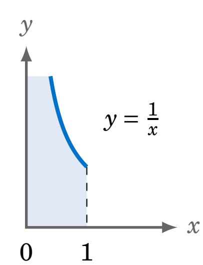
Example 5.33

Evaluate $\displaystyle\int_0^1 \frac{dx}{x}$.

Solution: Since $x = 0$ is a vertical asymptote for $y = \frac{1}{x}$,

$$\int_0^1 \frac{dx}{x} \;=\; \lim_{c\to 0+} \int_c^1 \frac{dx}{x} \;=\; \lim_{c\to 0+} \left(\ln x\,\Big|_c^1\right)$$

$$= \lim_{c\to 0+} (\ln 1 \;-\; \ln c) \;=\; 0 \;-\; (-\infty) \;=\; \infty$$

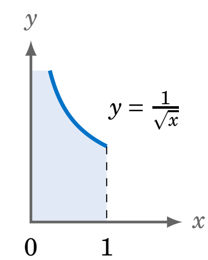


and so the integral is divergent. This means that the area under the curve $y = 1/x$ over the interval $(0, 1]$—as shown in the graph above—is infinite. The region is infinite in the y direction.

Example 5.34

Evaluate $\displaystyle\int_0^1 \frac{dx}{\sqrt{x}}$.

Solution: Since $x = 0$ is a vertical asymptote for $y = \frac{1}{\sqrt{x}}$,



$$\int_0^1 \frac{dx}{\sqrt{x}} = \lim_{c\to 0+} \int_c^1 \frac{dx}{\sqrt{x}} = \lim_{c\to 0+} \left(2\sqrt{x}\,\Big|_c^1\right)$$

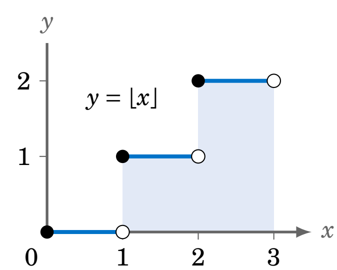
$$= \lim_{c\to 0+} (2 - 2\sqrt{c}) = 2 - 0 = 2 .$$

This means that the area under the curve $y = 1/\sqrt{x}$ over the interval $(0,1]$—as shown in the graph above—equals 2. The region is infinite in the y direction.

Example 5.35

Evaluate $\displaystyle\int_1^3 \lfloor x \rfloor\, dx$.

Solution: Recall from Example 3.22 in Section 3.3 that the floor function $y = \lfloor x \rfloor$ has jump discontinuities at each integer value of x, as shown in the graph on the right. The integral $\int_1^3 \lfloor x \rfloor\, dx$ is thus an improper integral over the interval $[1,3)$, which needs to be split at the point of discontinuity $x = 2$ within that interval:



$$\int_1^3 \lfloor x \rfloor\, dx = \int_1^2 \lfloor x \rfloor\, dx + \int_2^3 \lfloor x \rfloor\, dx$$

$$= \left(\lim_{b\to 2-} \int_1^b \lfloor x \rfloor\, dx\right) + \left(\lim_{c\to 3-} \int_2^c \lfloor x \rfloor\, dx\right)$$

$$= \left(\lim_{b\to 2-} \int_1^b 1\, dx\right) + \left(\lim_{c\to 3-} \int_2^c 2\, dx\right)$$

$$= \lim_{b\to 2-} \left(x\,\Big|_1^b\right) + \lim_{c\to 3-} \left(2x\,\Big|_2^c\right)$$

$$= \lim_{b\to 2-} (b-1) + \lim_{c\to 3-} (2c-4) = (2-1) + (6-4) = 3$$

Similar to some of the above examples, the following result is easy to prove (see the exercises):

For any real number $a > 0$, the improper integral

$$\int_a^\infty \frac{dx}{x^p}$$

is convergent if $p > 1$, and divergent if $0 < p \le 1$.

The following test for convergence or divergence is sometimes helpful:

Comparison Test for Improper Integrals:

(a) If $|f(x)| \le g(x)$ for all x in $[a,\infty)$, and if $\int_a^\infty g(x)\,dx$ is convergent, then $\int_a^\infty f(x)\,dx$ is convergent.

(b) If $f(x) \ge g(x) \ge 0$ for all x in $[a,\infty)$, and if $\int_a^\infty g(x)\,dx$ is divergent, then $\int_a^\infty f(x)\,dx$ is divergent.

The idea behind part (a) is that if $-g(x) \le f(x) \le g(x)$ over $[a,\infty)$, then—thinking of improper integrals as areas—the integral of f is "squeezed" between the two finite integrals for $\pm g$. There are, however, some subtle issues to prove about the limit in the integral of f—finite bounds might not necessarily mean the limit exists.[5]

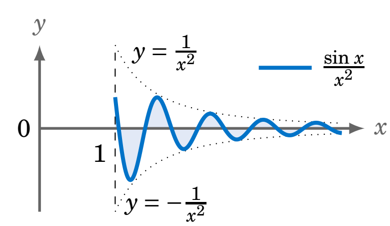
Example 5.36

Show that $\displaystyle\int_1^\infty \frac{\sin x}{x^2}\,dx$ is convergent.

Solution: By Example 5.29, the integral $\int_1^\infty \frac{1}{x^2}\,dx$ is convergent. So since $|\sin x| \le 1$ for all x, then

$$\left|\frac{\sin x}{x^2}\right| \le \frac{1}{x^2}$$

for all x in $[1,\infty)$. Thus, by the Comparison Test, $\int_1^\infty \frac{\sin x}{x^2}\,dx$ is convergent. The graph on the right shows how the curve $y = \frac{\sin x}{x^2}$ is bounded between the curves $y = \pm\frac{1}{x^2}$.



The rules and properties from Section 5.3 concerning definite integrals still apply to improper integrals, provided the improper integrals are convergent. For example, suppose a function f has a discontinuity or vertical asymptote at $x = c$. If both improper integrals $\int_a^c f(x)\,dx$ and $\int_c^b f(x)\,dx$ are convergent, then the improper integral $\int_a^b f(x)\,dx$ is convergent and

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx\,.$$

Likewise, if $\int_a^c f(x)\,dx$ and $\int_c^\infty f(x)\,dx$ are convergent, then so is $\int_a^\infty f(x)\,dx$, with

$$\int_a^\infty f(x)\,dx = \int_a^c f(x)\,dx + \int_c^\infty f(x)\,dx\,.$$

[5]See pp.140-141 in BUCK, R.C., *Advanced Calculus*, 2nd ed., New York: McGraw-Hill Book Co., 1965.

Exercises

A

For Exercises 1-15, evaluate the given improper integral.

1. $\int_1^\infty \frac{dx}{x^3}$ **2.** $\int_0^1 \frac{dx}{\sqrt[3]{x}}$ **3.** $\int_0^\infty e^{-x}\,dx$ **4.** $\int_0^\infty e^{-2x}\,dx$ **5.** $\int_{-1}^1 \frac{dx}{x}$

6. $\int_0^\infty xe^{-x^2}\,dx$ **7.** $\int_{-\infty}^0 2^x\,dx$ **8.** $\int_0^{\pi/2} \tan x\,dx$ **9.** $\int_0^1 \frac{\ln x}{x}\,dx$ **10.** $\int_{-1}^1 \frac{dx}{\sqrt{1-x^2}}$

11. $\int_0^3 \lceil x \rceil\,dx$ **12.** $\int_{-\infty}^\infty \frac{dx}{x^2+4}$ **13.** $\int_0^1 \frac{dx}{(x-1)^3}$ **14.** $\int_2^\infty \frac{dx}{x\ln x}$ **15.** $\int_1^\infty \frac{dx}{x\sqrt{x^2-1}}$

16. In a *standby system* of two non-identical components, the normal operating component A has a failure rate of $\lambda_A > 0$ failures per unit time, while the standby component B—which takes over when A fails—has a failure rate $\lambda_B > 0$ (with $\lambda_A \neq \lambda_B$).

(a) Find the standby system's *reliability* $R(t)$ beyond time $t \ge 0$, where

$$R(t) = \int_t^\infty \frac{\lambda_A\lambda_B}{\lambda_A-\lambda_B}\left(e^{-\lambda_B x} - e^{-\lambda_A x}\right)dx\,.$$

(b) Show that the system's *mean time to failure* (MTTF) m, where $m = \int_0^\infty R(t)\,dt$, is $m = \frac{1}{\lambda_A} + \frac{1}{\lambda_B}$.

17. Show that for all $a > 0$, $\int_a^\infty \frac{dx}{x^p}$ is convergent if $p > 1$, and divergent if $0 < p \le 1$.

18. Show that for all $a > 0$, $\int_0^a \frac{dx}{x^p}$ is convergent if $0 < p < 1$, and divergent if $p \ge 1$.

19. Is $\int_1^\infty \frac{dx}{x+x^4}$ convergent? Explain.

20. Is $\int_2^\infty \frac{dx}{x-\sqrt{x}}$ convergent? Explain.

B

21. Example 5.31 showed that $\int_0^\infty \sin x\,dx$ is divergent. What is the flaw in the argument that the integral must be 0 since each "hump" of $\sin x$ above the x-axis is canceled by one below the x-axis?

22. This exercise concerns the subtraction rule $\int_a^\infty (f(x)-g(x))\,dx = \int_a^\infty f(x)\,dx - \int_a^\infty g(x)\,dx$.

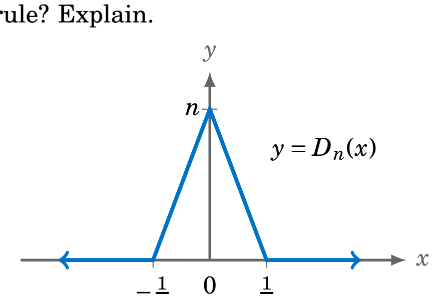
(a) Show that $\frac{1}{x(x+1)} = \frac{1}{x} - \frac{1}{x+1}$ for all x except 0 and -1

(b) Show that $\int_1^\infty \frac{dx}{x(x+1)}$ is convergent.

(c) Show that both $\int_1^\infty \frac{dx}{x}$ and $\int_1^\infty \frac{dx}{x+1}$ are divergent.

(d) Does part (c) contradict parts(a)-(b) and the subtraction rule? Explain.

23. The improper integral $\int_{-\infty}^\infty \delta(x)\,dx = 1$ is one of the notable "improprieties" of the Dirac delta function δ. One way to think of that integral is by approximating δ by triangular "pulse" functions D_n (for $n \ge 1$), as in the picture on the right.



(a) Write a formula for each $D_n(x)$ over all of $\mathbb{R}$.

(b) Show that $\int_{-\infty}^\infty D_n(x)\,dx = 1$ for all integers $n \ge 1$.

(c) Show that $\lim_{n\to\infty} D_n(0) = \infty = \delta(0)$.

(d) Do the D_n functions begin to resemble δ as $n \to \infty$?

CHAPTER 6

Methods of Integration

6.1 Integration by Parts

In physics and engineering the **Gamma function**[1] $\Gamma(t)$, defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dx \quad \text{for all } t > 0,$$

has found many uses. Evaluating $\Gamma(2)$ entails integrating the function $f(x) = x e^{-x}$. No formula or substitution you have learned so far would be of help. *Differentiating* that function, on the other hand, is easy. By the Product Rule for differentials,

$$\begin{aligned}
d(xe^{-x}) &= x\,d(e^{-x}) + d(x)e^{-x} \\
&= -xe^{-x}\,dx + e^{-x}\,dx \\
d(xe^{-x}) &= -xe^{-x}\,dx - d(e^{-x}) \\
xe^{-x}\,dx &= -d(xe^{-x}) - d(e^{-x}), \quad \text{so integrate both sides to get}
\end{aligned}$$

$$\int xe^{-x}\,dx = -\int d(xe^{-x}) - \int d(e^{-x}) = -xe^{-x} - e^{-x} + C$$

since $\int dF = F + C$. Generalizing this process for functions u and v,

$$\begin{aligned}
d(uv) &= u\,dv + v\,du \\
u\,dv &= d(uv) - v\,du
\end{aligned}$$

so that integrating both sides yields the **integration by parts** formula:

For differentiable functions u and v:

$$\int u\,dv = uv - \int v\,du \tag{6.1}$$

[1]Created by the Swiss mathematician, physicist and astronomer Leonhard Euler (1707-1783). The use of the Greek capital letter Γ for this function is due to the French mathematician Adrien-Marie Legendre (1752-1833).

Integration by parts is just the Product Rule for derivatives in integral form, typically used when the integral $\int v\,du$ would be simpler than the original integral $\int u\,dv$.

Example 6.1

Use integration by parts to evaluate $\displaystyle\int x\,e^{-x}\,dx$. Use the answer to evaluate $\Gamma(2)$.

Solution: The original integral is always of the form $\int u\,dv$, so you must decide which parts of $xe^{-x}\,dx$ will represent u and dv. Typically you would choose dv to be a differential that you could integrate easily (since you will need to integrate dv to get v) and choose u to be a function whose derivative is simpler than u (since that derivative will appear in $v\,du$, which you hope to be a simpler integral). In this case, pick $u = x$ and $dv = e^{-x}\,dx$. Then $du = dx$ and $v = \int dv = \int e^{-x}\,dx = -e^{-x}$ (you can omit the generic constant C for now—include it when you have evaluated $\int v\,du$). Thus,

$$\int u\,dv = uv - \int v\,du$$

$$\int \underbrace{x}_{u}\,\underbrace{e^{-x}\,dx}_{dv} = \underbrace{x}_{u}\,\underbrace{(-e^{-x})}_{v} - \int \underbrace{-e^{-x}}_{v}\,\underbrace{dx}_{du}$$

$$\int x\,e^{-x}\,dx = -x\,e^{-x} - e^{-x} + C$$

which agrees with the example at the beginning of this section. Note that $\int v\,du = \int -e^{-x}\,dx$, which indeed is simpler than the original integral. The Gamma function value $\Gamma(2)$ can now be evaluated:

$$\begin{aligned}
\Gamma(2) &= \int_0^\infty x\,e^{-x}\,dx \\
&= -x\,e^{-x} - e^{-x}\,\Big|_0^\infty \\
&= \lim_{x\to\infty}\,(-x\,e^{-x} - e^{-x}) - (-0\,e^0 - e^0) \\
&= -\left(\lim_{x\to\infty}\,\frac{x}{e^x}\right) - \left(\lim_{x\to\infty}\,e^{-x}\right) - (0-1) \\
&= 0 - 0 + 1 = 1
\end{aligned}$$

Example 6.2

What would happen in Example 6.1 if you let $u = e^{-x}$ and $dv = x\,dx$?

Solution: In this case $du = -e^{-x}\,dx$ and $v = \int dv = \int x\,dx = \frac{1}{2}x^2$, so that

$$\begin{aligned}
\int x\,e^{-x}\,dx &= \int \underbrace{e^{-x}}_{u}\,\underbrace{x\,dx}_{dv} \\
&= \underbrace{e^{-x}}_{u}\,\underbrace{\frac{1}{2}x^2}_{v} - \int \underbrace{\frac{1}{2}x^2}_{v}\,\underbrace{(-e^{-x})\,dx}_{du} \\
&= \frac{1}{2}x^2\,e^{-x} + \frac{1}{2}\int x^2\,e^{-x}\,dx
\end{aligned}$$

which leads you in the wrong direction: a more difficult integral than the original.

Example 6.2 showed the importance of an appropriate choice for u and dv. There are some rough guidelines for that choice—as in Example 6.1—but no rules that are guaranteed to always work. It might not be clear when you should even attempt integration by parts.

Example 6.3

Evaluate $\int \ln x \, dx$.

Solution: Integration by parts ostensibly requires two functions in the integral, whereas here $\ln x$ appears to be the only one. However, the choice for dv is a *differential*, and one exists here: dx. Choosing $dv = dx$ obliges you to let $u = \ln x$. Then $du = \frac{1}{x}\,dx$ and $v = \int dv = \int dx = x$. Now integrate by parts:

$$\begin{aligned}
\int u\,dv &= uv - \int v\,du \\
\int \ln x \, dx &= (\ln x)(x) - \int x \cdot \frac{1}{x}\,dx \\
&= x \ln x - \int 1\,dx \\
&= x \ln x - x + C
\end{aligned}$$

Note that choosing $dv = \ln x\,dx$ would be pointless, as integrating dv to get v is the original problem!

Example 6.4

Evaluate $\int x^3 e^{x^2}\,dx$.

Solution: One frequently useful guideline for integration by parts is to eliminate the most complicated function in the integral by integrating it—as dv—into something simpler (which becomes v). In this integral, e^{x^2} is somewhat complicated but has no closed form antiderivative. However, $x e^{x^2}$ appears in the integral and can be integrated easily (using a substitution as in Section 5.4). So choose $dv = x e^{x^2}\,dx$, which means $u = x^2$. Then $du = 2x\,dx$ and $v = \int dv = \int x e^{x^2}\,dx = \frac{1}{2}e^{x^2}$. Now integrate by parts:

$$\begin{aligned}
\int u\,dv &= uv - \int v\,du \\
\int x^3 e^{x^2}\,dx &= x^2 \cdot \frac{1}{2}e^{x^2} - \int \frac{1}{2}e^{x^2} \cdot 2x\,dx \\
&= \frac{x^2}{2}e^{x^2} - \int x e^{x^2}\,dx \\
&= \frac{x^2}{2}e^{x^2} - \frac{1}{2}e^{x^2} + C \\
&= \frac{(x^2-1)}{2}e^{x^2} + C
\end{aligned}$$

Sometimes multiple rounds of integration by parts are needed, as in the following example.

Example 6.5

Evaluate $\int x^2 e^{-x}\,dx$.

Solution: This integral appears similar to the one in Example 6.1, so choose $dv = e^{-x}\,dx$ and $u = x^2$. Then $du = 2x\,dx$ and $v = \int e^{-x}\,dv = -e^{-x}$. Now integrate by parts:

$$\int u\,dv = uv - \int v\,du$$

$$\int x^2 e^{-x}\,dx = x^2 \cdot (-e^{-x}) - \int -e^{-x} \cdot 2x\,dx$$

$$= -x^2 e^{-x} + 2 \int x e^{-x}\,dx \quad \text{(integrate by parts } \textit{again}\text{)}$$

$$\int x^2 e^{-x}\,dx = -x^2 e^{-x} + 2(-x e^{-x} - e^{-x}) + C \quad \text{(by Example 6.1)}$$

$$= -x^2 e^{-x} - 2x e^{-x} - 2e^{-x} + C$$

In the above example, notice that the $u = 2x$ in the second integral came from the derivative of the $u = x^2$ in the first integral. Likewise, the $dv = -e^{-x}\,dx$ in the second integral came from integrating the $dv = e^{-x}\,dx$ in the first integral. In general, if n rounds of integration by parts were needed, with u_i and v_i representing the u and v, respectively, for round $i = 1, 2, \ldots, n$, then the repeated integration by parts would look like this:

$$\int u_1\,dv_1 = u_1 v_1 - \int v_1\,du_1$$

$$= u_1 v_1 - \int u_2\,dv_2$$

$$= u_1 v_1 - \left(u_2 v_2 - \int v_2\,du_2\right) = u_1 v_1 - u_2 v_2 + \int u_3\,dv_3$$

$$= u_1 v_1 - u_2 v_2 + \left(u_3 v_3 - \int u_4\,dv_4\right)$$

$$= u_1 v_1 - u_2 v_2 + u_3 v_3 - \left(u_4 v_4 - \int u_5\,dv_5\right)$$

$$= \cdots$$

$$= u_1 v_1 - u_2 v_2 + u_3 v_3 - u_4 v_4 + u_5 v_5 - \cdots \int u_n\,dv_n$$

The last integral $\int u_n\,dv_n$ is one you could presumably integrate easily.

The above procedure is called the **tabular method** for integration by parts, since it can be shown in a table (the arrows indicate multiplication):

$$
\begin{array}{cccccc}
u & & dv & & & \\
\hline
u_1 & \searrow & dv_1 & & & \\
u_2 & \searrow & v_1 & (+) & \longrightarrow & +\,u_1 v_1 \\
u_3 & \searrow & v_2 & (-) & \longrightarrow & -\,u_2 v_2 \\
u_3 & \searrow & v_3 & (+) & \longrightarrow & +\,u_3 v_3 \\
u_4 & & v_4 & (-) & \longrightarrow & -\,u_4 v_4 \\
\vdots & & \vdots & \vdots & &
\end{array}
$$

The idea is to differentiate down the u column and integrate down the dv column. If the u in the original integral is a polynomial of degree n, then you know from Section 1.6 that its $(n+1)$-st derivative will be 0, at which point the tabular method terminates. The integral is then the sum of the indicated products with alternating signs.

For example, the tabular method on the integral from Example 6.5 looks like this:

$$
\begin{array}{ccccccc}
 & u & & dv & & & \\
\hline
 & x^2 & \searrow & e^{-x}\,dx & & & \\
 & 2x & \searrow & -e^{-x} & (+) & \longrightarrow & +\,(x^2)(-e^{-x}) \\
 & 2 & \searrow & e^{-x} & (-) & \longrightarrow & -\,(2x)(e^{-x}) \\
\text{STOP} \longrightarrow & 0 & & -e^{-x} & (+) & \longrightarrow & +\,(2)(-e^{-x})
\end{array}
$$

The integral is the sum of the products, and agrees with the result in Example 6.5:

$$\int x^2\,e^{-x}\,dx \;=\; +\,(x^2)(-e^{-x}) \;-\; (2x)(e^{-x}) \;+\; (2)(-e^{-x}) \;+\; C \;=\; -x^2\,e^{-x} \;-\; 2x\,e^{-x} \;-\; 2\,e^{-x} \;+\; C$$

Example 6.6

Evaluate $\int x^3\,e^{-x}\,dx$.

Solution: Use the tabular method with $u = x^3$ and $dv = e^{-x}\,dx$:

$$
\begin{array}{ccccccc}
 & u & & dv & & & \\
\hline
 & x^3 & \searrow & e^{-x}\,dx & & & \\
 & 3x^2 & \searrow & -e^{-x} & (+) & \longrightarrow & +\,(x^3)(-e^{-x}) \\
 & 6x & \searrow & e^{-x} & (-) & \longrightarrow & -\,(3x^2)(e^{-x}) \\
 & 6 & \searrow & -e^{-x} & (+) & \longrightarrow & +\,(6x)(-e^{-x}) \\
\text{STOP} \longrightarrow & 0 & & e^{-x} & (-) & \longrightarrow & -\,(6)(e^{-x})
\end{array}
$$

$$\int x^3\,e^{-x}\,dx \;=\; -x^3\,e^{-x} \;-\; 3x^2\,e^{-x} \;-\; 6x\,e^{-x} \;-\; 6\,e^{-x} \;+\; C$$

Integration by parts can sometimes result in the original integral reappearing, allowing it to be combined with the original integral.

Example 6.7

Evaluate $\int \sec^3 x\, dx$.

Solution: Let $u = \sec x$ and $dv = \sec^2 x\, dx$, so that $du = \sec x\, \tan x\, dx$ and $v = \int dv = \int \sec^2 x\, dx = \tan x$. Then

$$\int u\, dv = uv - \int v\, du$$

$$\int \sec^3 x\, dx = \sec x \tan x - \int \sec x \tan^2 x\, dx$$

$$\int \sec^3 x\, dx = \sec x \tan x - \int \sec x\, (\sec^2 x - 1)\, dx$$

$$\int \sec^3 x\, dx = \sec x \tan x + \int \sec x\, dx - \int \sec^3 x\, dx$$

$$2\int \sec^3 x\, dx = \sec x \tan x + \ln|\sec x + \tan x| + C$$

$$\int \sec^3 x\, dx = \frac{1}{2}(\sec x \tan x + \ln|\sec x + \tan x|) + C$$

Example 6.8

Evaluate $\int e^x \sin x\, dx$.

Solution: Let $u = e^x$ and $dv = \sin x\, dx$, so that $du = e^x\, dx$ and $v = \int dv = \int \sin x\, dx = -\cos x$. Then

$$\int u\, dv = uv - \int v\, du$$

$$\int e^x \sin x\, dx = -e^x \cos x + \int e^x \cos x\, dx$$

and so integration by parts is needed again, for the integral on the right: let $u = e^x$ and $dv = \cos x\, dx$, so that $du = e^x\, dx$ and $v = \int dv = \int \cos x\, dx = \sin x$. Then

$$\int e^x \sin x\, dx = -e^x \cos x + \left(uv - \int v\, du\right)$$

$$\int e^x \sin x\, dx = -e^x \cos x + \left(e^x \sin x - \int e^x \sin x\, dx\right)$$

$$2\int e^x \sin x\, dx = -e^x \cos x + e^x \sin x$$

$$\int e^x \sin x\, dx = \frac{e^x}{2}(\sin x - \cos x) + C$$

In Example 6.1 integration by parts was used in evaluating an improper integral. In general, in definite or improper integrals where a and b are real numbers or $\pm\infty$,

$$\int_a^b u\,dv \;=\; uv\,\Big|_a^b \;-\; \int_a^b v\,du\;.$$

Example 6.9

Evaluate $\displaystyle\int_0^1 x^3\sqrt{1-x^2}\;dx$.

Solution: Since $x^3\sqrt{1-x^2} = x^2 \cdot x\sqrt{1-x^2}$, and $x\sqrt{1-x^2}$ is easy to integrate (via a substitution), let $u = x^2$ and $dv = x\sqrt{1-x^2}\,dx$. Then $du = 2x\,dx$ and $v = \int dv = \int x\sqrt{1-x^2}\,dx = -\frac{1}{3}(1-x^2)^{3/2}$, and so:

$$\int_a^b u\,dv \;=\; uv\,\Big|_a^b \;-\; \int_a^b v\,du$$

$$\int_0^1 x^3\sqrt{1-x^2}\;dx \;=\; -\frac{x^2}{3}(1-x^2)^{3/2}\,\Big|_0^1 \;+\; \int_0^1 \frac{2x}{3}(1-x^2)^{3/2}\;dx$$

$$=\;(0-0)\;+\;\left(\frac{-2}{15}(1-x^2)^{5/2}\,\Big|_0^1\right) \;=\; 0\;+\;\frac{2}{15} \;=\; \frac{2}{15}$$

Exercises

A

For Exercises 1-25, evaluate the given integral.

1. $\int x\ln x\;dx$ **2.** $\int x^2e^x\,dx$ **3.** $\int x\cos x\;dx$ **4.** $\int x\,3^x\,dx$ **5.** $\int x^2a^x\,dx\;\;(a>0)$

6. $\int \ln 4x\;dx$ **7.** $\int \ln x^2\;dx$ **8.** $\int x^2\sin x\;dx$ **9.** $\int x\cos^2 x\;dx$ **10.** $\int \sin x\cos 2x\;dx$

11. $\int \sin^{-1}x\;dx$ **12.** $\int \cos^{-1}2x\;dx$ **13.** $\int \tan^{-1}3x\;dx$ **14.** $\int x\sec^2 x\;dx$ **15.** $\int \sin x\sin 3x\;dx$

16. $\int \frac{\ln x}{x^3}\,dx$ **17.** $\int x^3\ln^2 x\;dx$ **18.** $\int x^5e^x\,dx$ **19.** $\int_0^2 \frac{x^3\,dx}{\sqrt{4-x^2}}$ **20.** $\int_0^1 x^3\sqrt{1+x^2}\,dx$

21. $\int \sin(\ln x)\;dx$ **22.** $\int \ln(1+x^2)\,dx$ **23.** $\int x\tan^{-1}x\;dx$ **24.** $\int \cot^{-1}\sqrt{x}\;dx$ **25.** $\int e^{\sqrt{x}}\,dx$

26. Evaluate the integral $\int e^x\sin x\,dx$ from Example 6.8 by using two rounds of the tabular method and the formula $\int u_1\,dv_1 = u_1v_1 - u_2v_2 + \int v_2\,du_2$ from p.162.

B

27. Show that for all constants a and $b \neq 0$:

$$\int e^{ax}\cos bx\;dx \;=\; \frac{e^{ax}(a\cos bx\,+\,b\sin bx)}{a^2+b^2}\;+\;C \quad\text{and}\quad \int e^{ax}\sin bx\;dx \;=\; \frac{e^{ax}(a\sin bx\,-\,b\cos bx)}{a^2+b^2}\;+\;C$$

28. For the Gamma function $\Gamma(t)$ show the following:

(a) $\Gamma(t+1) = t\,\Gamma(t)$ for all $t > 0$. *(Hint: Use integration by parts.)*

(b) $\Gamma(n) = (n-1)!$ for all positive integers n. *(Hint: Use part (a) and induction.)*

Note that by part (b) the Gamma function can be thought of as an extension of the factorial operation to all positive real numbers. In fact, the Gamma function was created for that purpose.

29. Use Exercise 28 to prove for all integers $n \ge 1$:

$$\int_0^\infty r^n e^{-r} \ln r\, dr = (n-1)! + n \int_0^\infty r^{n-1} e^{-r} \ln r\, dr$$

30. By the *Maxwell speed distribution* for gas molecules, the average speed $\langle v \rangle$ of molecules of mass m in a gas at temperature T is

$$\langle v \rangle = 4\pi \left(\frac{m}{2\pi kT}\right)^{3/2} \int_0^\infty v^3 e^{-mv^2/2kT}\, dv\ ,$$

where $k \approx 1.38056 \times 10^{-23}$ J/K is the Boltzmann constant. Show that

$$\langle v \rangle = \sqrt{\frac{8kT}{\pi m}}\ .$$

31. Some physics texts write integrals in a form like this energy integral from statistical mechanics,

$$\int_0^\infty \ln\left(1 - \alpha e^{-x^2}\right)\, d(x^3)$$

which uses the differential of a *function*—in this case $d(x^3)$—instead of a variable (e.g. not just plain dx). This often signals that integration by parts is on the way, with the added benefit of having the $v = \int dv$ calculation done for you—in the above integral $d(x^3)$ means that $v = x^3$, with $dv = d(x^3)$ not really being needed for anything else. With that understanding, show that for $0 < \alpha < 1$,

$$\int_0^\infty \ln\left(1 - \alpha e^{-x^2}\right)\, d(x^3) = -2\alpha \int_0^\infty \frac{x^4 e^{-x^2}\, dx}{1 - \alpha e^{-x^2}}\ .$$

32. Find the flaw in the following "proof" that $0 = 1$:

Evaluating the integral $\int \frac{dx}{x}$ *using integration by parts with* $u = \frac{1}{x}$ *and* $dv = dx$*, so that* $du = -\frac{dx}{x^2}$ *and* $v = x$*, shows that*

$$\int u\, dv = uv - \int v\, du$$

$$\int \frac{dx}{x} = \left(\frac{1}{x}\right) \cdot x - \int x \cdot \left(-\frac{dx}{x^2}\right)$$

$$\int \frac{dx}{x} = 1 + \int \frac{dx}{x}$$

$$0 = 1 \quad \checkmark$$

6.2 Trigonometric Integrals

In engineering applications you sometimes encounter integrals of the form

$$\int \cos(\alpha t + \phi_1)\, \cos(\beta t + \phi_2)\, dt$$

where $\alpha t + \phi_1$ and $\beta t + \phi_2$ are different angles (e.g. when the voltage and current are out of phase in an AC circuit). In general, integrals involving products of sines and cosines with "mixed" angles can be simplified with the useful product-to-sum formulas:[2]

$$\sin A\, \cos B = \tfrac{1}{2}(\sin(A+B) + \sin(A-B)) \tag{6.2}$$

$$\cos A\, \sin B = \tfrac{1}{2}(\sin(A+B) - \sin(A-B)) \tag{6.3}$$

$$\cos A\, \cos B = \tfrac{1}{2}(\cos(A+B) + \cos(A-B)) \tag{6.4}$$

$$\sin A\, \sin B = -\tfrac{1}{2}(\cos(A+B) - \cos(A-B)) \tag{6.5}$$

Example 6.10

Evaluate $\int 0.5 \sin x \sin 12x\, dx$.

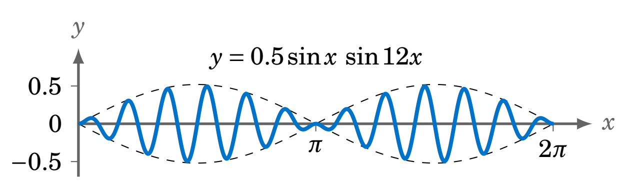
Solution: Using the product-to-sum formula (6.5) with $A = x$ and $B = 12x$,

$$\begin{aligned}
\sin A\, \sin B &= -\tfrac{1}{2}(\cos(A+B) - \cos(A-B))\\
\sin x\, \sin 12x &= -\tfrac{1}{2}(\cos(x+12x) - \cos(x-12x))\\
\sin x\, \sin 12x &= -\tfrac{1}{2}(\cos 13x - \cos 11x)
\end{aligned}$$

since $\cos(-11x) = \cos 11x$. Then

$$\begin{aligned}
\int 0.5 \sin x \sin 12x\, dx &= -\frac{1}{4}\int (\cos 13x - \cos 11x)\, dx\\
&= -\frac{1}{52}\sin 13x + \frac{1}{44}\sin 11x + C
\end{aligned}$$

Notice how the product-to-sum formula turned an integral of products of sines into integrals of individual cosines, which are easily integrated. The integrand is an example of a *modulated wave*, commonly used in electronic communications (e.g. radio broadcasting). The graph is shown below:



The curves $y = \pm 0.5 \sin x$ (shown in dashed lines) form an *amplitude envelope* for the modulated wave.

[2]See Section 3.4 in CORRAL, M., *Trigonometry*, `http://mecmath.net/trig/`, 2009.

On occasion you might need to integrate trigonometric functions raised to powers higher than two. For the sine function raised to odd powers of the form $2n+1$ (for $n \ge 1$), the trick is to replace $\sin^2 x$ by $1-\cos^2 x$, so that

$$\begin{aligned}\int \sin^{2n+1} x\,dx &= \int (\sin^2 x)^n \sin x\,dx\\ &= \int (1-\cos^2 x)^n \sin x\,dx\\ &= \int p(u)\,du\end{aligned}$$

where $p(u)$ is a polynomial in the variable $u = \cos x$, and the remaining single $\sin x$ is now part of $du = -\sin x\,dx$. You can then use the Power Formula to integrate that polynomial.

Example 6.11

Evaluate $\int \sin^3 x\,dx$.

Solution: Let $u = \cos x$ so that $du = -\sin x\,dx$:

$$\begin{aligned}\int \sin^3 x\,dx &= \int (\sin^2 x) \sin x\,dx\\ &= \int (1-\cos^2 x) \sin x\,dx\\ &= \int (1-u^2)(-du)\\ &= \int (u^2-1)\,du\\ &= \frac{1}{3}u^3 - u + C\\ &= \frac{1}{3}\cos^3 x - \cos x + C\end{aligned}$$

In general $\int \sin^{2n+1} x\,dx$ will be a polynomial of degree $2n+1$ in terms of $\cos x$. Similarly, use $\cos^2 x = 1-\sin^2 x$ to integrate odd powers of $\cos x$, with the substitution $u = \sin x$:

$$\begin{aligned}\int \cos^{2n+1} x\,dx &= \int (\cos^2 x)^n \cos x\,dx\\ &= \int \underbrace{(1-\sin^2 x)^n}_{p(u)}\ \underbrace{\cos x\,dx}_{du}\end{aligned}$$

Integrals of the form $\int \sin^m x \cos^n x\, dx$, where either m or n is odd, can be evaluated using the above trick for the function having the odd power.

Example 6.12

Evaluate $\displaystyle\int \sin^2 x \cos^3 x\, dx$.

Solution: Replace $\cos^2 x$ by $1 - \sin^2 x$, then let $u = \sin x$ so that $du = \cos x\, dx$:

$$\begin{aligned}
\int \sin^2 x \cos^3 x \; dx &= \int \sin^2 x\,(\cos^2 x) \cos x\, dx \\
&= \int \underbrace{\sin^2 x\,(1 - \sin^2 x)}_{p(u)}\,\underbrace{\cos x\, dx}_{du} \\
&= \int (u^2 - u^4)\, du \\
&= \frac{1}{3}u^3 - \frac{1}{5}u^5 + C \\
&= \frac{1}{3}\sin^3 x - \frac{1}{5}\sin^5 x + C
\end{aligned}$$

For even powers of $\sin x$ or $\cos x$. You would replace $\sin^2 x$ or $\cos^2 x$ with either

$$\sin^2 x = \frac{1 - \cos 2x}{2} \qquad \text{or} \qquad \cos^2 x = \frac{1 + \cos 2x}{2},$$

respectively, as often as necessary, then proceed as before if odd powers occur.

Example 6.13

Evaluate $\displaystyle\int \sin^4 x\, dx$.

Solution: Replace $\sin^2 x$ by $\frac{1-\cos 2x}{2}$:

$$\begin{aligned}
\int \sin^4 x\, dx &= \int \left(\frac{1 - \cos 2x}{2}\right)^2 dx \\
&= \frac{1}{4}\int (1 - 2\cos 2x + cos^2 2x)\, dx \\
&= \frac{1}{4}\int \left(1 - 2\cos 2x + \frac{1 + \cos 4x}{2}\right) dx \\
&= \frac{1}{8}\int (3 - 4\cos 2x + \cos 4x)\, dx \\
&= \frac{3x}{8} - \frac{1}{4}\sin 2x + \frac{1}{32}\sin 4x + C
\end{aligned}$$

Similar methods can be used for integrals of the form $\int \sec^m x \tan^n x\,dx$ when either m is even or n is odd. For an even power $m = 2k+2$, use $\sec^2 x = 1+\tan^2 x$ for all but two of the m powers of $\sec x$, then use the substitution $u = \tan x$, so that $du = \sec^2 x\,dx$. This results in an integral of a polynomial $p(u)$ in terms of $u = \tan x$:

$$\begin{aligned}\int \sec^{2k+2} x\ \tan^n x\ dx &= \int (\sec^2 x)^k \sec^2 x\ \tan^n x\ dx\\ &= \int \underbrace{(1+\tan^2 x)^k \tan^n x}_{p(u)}\ \underbrace{\sec^2 x\ dx}_{du}\end{aligned}$$

Likewise for an odd power $n = 2k+1$, use $\tan^2 x = \sec^2 x - 1$ for all but one of the n powers of $\tan x$, then use the substitution $u = \sec x$, so that $du = \sec x\,\tan x\,dx$. This results in an integral of a polynomial $p(u)$ in terms of $u = \sec x$:

$$\begin{aligned}\int \sec^m x\ \tan^{2k+1} x\ dx &= \int \sec^{m-1} x\ \sec x\ (\tan^2 x)^k\ \tan x\ dx\\ &= \int \underbrace{\sec^{m-1} x\ (\sec^2 x - 1)^k}_{p(u)}\ \underbrace{\sec x\ \tan x dx}_{du}\end{aligned}$$

Mimic the above procedure for integrals of the form $\int \csc^m x \cot^n x\,dx$ when either m is even or n is odd, using the identity $\csc^2 x = 1+\cot^2 x$ in a similar manner.

Example 6.14

Evaluate $\displaystyle\int \sec^4 x\ \tan x\ dx$.

Solution: Use $\sec^2 x = 1+\tan^2 x$ for one $\sec^2 x$ term, then substitute $u = \tan x$, so that $du = \sec^2 x\,dx$:

$$\begin{aligned}\int \sec^4 x\ \tan x\ dx &= \int \sec^2 x\ \sec^2 x\ \tan x\ dx\\ &= \int (1+\tan^2 x)\tan x\ \sec^2 x\ dx\\ &= \int (1+u^2)u\ du\\ &= \int (u+u^3)\ du\\ &= \frac{1}{2}u^2 + \frac{1}{4}u^4 + C\\ &= \frac{1}{2}\tan^2 x + \frac{1}{4}\tan^4 x + C\end{aligned}$$

For some trigonometric integrals try putting everything in terms of sines and cosines.

Example 6.15

Evaluate $\int \frac{\cot^4 x}{\csc^5 x}\,dx$.

Solution: Put $\cot x$ and $\csc x$ in terms of $\sin x$ and $\cos x$:

$$\int \frac{\cot^4 x}{\csc^5 x}\,dx = \int \frac{\cos^4 x\ \sin^5 x}{\sin^4 x}\,dx$$

$$= \int \cos^4 x\ \sin x\,dx \quad (\text{now let } u = \cos x,\ du = -\sin x\,dx)$$

$$= -\int u^4\,dx = -\frac{1}{5}\cos^5 x + C$$

Exercises

A

For Exercises 1-12, evaluate the given integral.

1. $\int \sin 2x\ \cos 5x\,dx$ **2.** $\int \cos 2x\ \cos 5x\,dx$ **3.** $\int \sin 2x\ \sin 5x\,dx$ **4.** $\int \cos 2\pi x\ \sin 3\pi x\,dx$

5. $\int \sin^3 x\ \cos^{3/2} x\,dx$ **6.** $\int \cos^4 x\,dx$ **7.** $\int \sin^6 x\,dx$ **8.** $\int \sin x\ \sin 2x\ \sin 3x\,dx$

9. $\int \sec^4 x\,dx$ **10.** $\int \sec^2 x\ \tan^3 x\,dx$ **11.** $\int \frac{\tan^3 x}{\sec^4 x}\,dx$ **12.** $\int \frac{dx}{\csc^2 x\ \cot x}$

B

13. Evaluate $\int \sin^3 x\ \cos^3 x\,dx$ in two different ways:

(a) Use $\sin^3 x = (1-\cos^2 x)\sin x$ and the substitution $u = \cos x$.

(b) Use $\cos^3 x = (1-\sin^2 x)\cos x$ and the substitution $u = \sin x$.

Are the answers from parts(a) and (b) equivalent? Explain.

14. Evaluate $\int \sec^4 x\ \tan x\,dx$ by using $\sec^4 x\ \tan x = \sec^3 x\,(\sec x\ \tan x)$ and the substitution $u = \sec x$. Is your answer equivalent to the answer in Example 6.14? Explain.

15. Show that $\int \frac{4\tan x\,(1-\tan^2 x)}{(1+\tan^2 x)^2}\,dx = -\frac{1}{4}\cos 4x + C$.

16. The *autocorrelation function* $R_x(\tau)$ of the periodic function $x(t) = A\cos(\omega t + \theta)$ is given by

$$R_x(\tau) = \frac{\omega}{2\pi}\int_0^{2\pi/\omega} x(t)\,x(t-\tau)\,dt$$

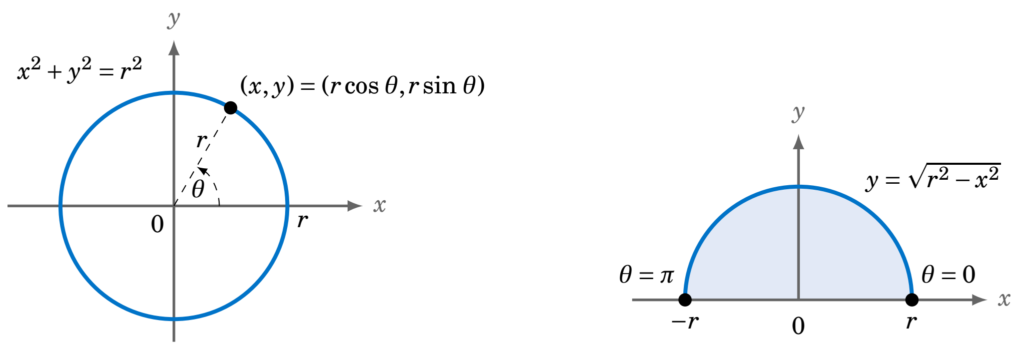
where A, ω and θ are constants. Show that

$$R_x(\tau) = \frac{A^2}{2}\cos\omega\tau .$$

6.3 Trigonometric Substitutions

One of the fundamental formulas in geometry is for the area A of a circle of radius r: $A = \pi r^2$. The calculus-based proof of that formula uses a definite integral evaluated by means of a **trigonometric substitution**, as will now be demonstrated.

Use the circle of radius $r > 0$ centered at the origin $(0,0)$ in the xy-plane, whose equation is $x^2 + y^2 = r^2$ (see Figure 6.3.1(a) below).



(a) Full circle **(b)** Upper hemisphere

Figure 6.3.1 Circle of radius r

By symmetry about the x-axis, the area A of the circle is twice the area of its upper hemisphere (see Figure 6.3.1(b) above), which is the area under the curve $y = \sqrt{r^2 - x^2}$:

$$A \;=\; 2\int_{-r}^{r} \sqrt{r^2 - x^2}\;dx$$

To evaluate this integral, recall from trigonometry that any point (x, y) on the circle can be written as $(x, y) = (r\cos\theta, r\sin\theta)$, where $0 \le \theta < 2\pi$ (in radians) is the angle shown in Figure 6.3.1(a). Figure 6.3.1(b) shows that as x goes from $x = -r$ to $x = r$, the angle θ goes from $\theta = \pi$ to $\theta = 0$. Now substitute $x = r\cos\theta$ and $dx = -r\sin\theta\,d\theta$ into the integral and change the limits of integration from $x = -r$ and $x = r$ to $\theta = \pi$ and $\theta = 0$, respectively:

$$\begin{aligned}
A &= 2\int_{\pi}^{0} \sqrt{r^2 - r^2\cos^2\theta}\;(-r\sin\theta)\;d\theta \;=\; -2\int_{0}^{\pi} \sqrt{r^2(1-\cos^2\theta)}\;(-r\sin\theta)\;d\theta \\
&= 2\int_{0}^{\pi} r\sqrt{\sin^2\theta}\;r\sin\theta\;d\theta \;=\; 2r^2\int_{0}^{\pi}\sin^2\theta\;d\theta \;=\; \cancel{2}r^2\int_{0}^{\pi}\frac{1-\cos 2\theta}{\cancel{2}}\;d\theta \\
&= r^2\left(\theta - \frac{1}{2}\sin 2\theta\right)\Bigg|_{0}^{\pi} \;=\; r^2\left(\pi - \frac{1}{2}\sin 2\pi - \left(0 - \frac{1}{2}\sin 0\right)\right) \\
&= \pi r^2 \quad \checkmark
\end{aligned}$$

For an indefinite integral of the general form $\int \sqrt{a^2 - u^2}\, du$, the same calculation as above with the substitutions $u = a\cos\theta$ and $du = -a\sin\theta\, d\theta$ yields

$$\int \sqrt{a^2 - u^2}\, du = \int \sqrt{a^2 - a^2\cos^2\theta}\,(-a\sin\theta)\, d\theta = -a^2 \int \sin^2\theta\, d\theta$$

$$= -a^2 \int \frac{1-\cos 2\theta}{2}\, d\theta = -\frac{a^2}{2}\theta + \frac{a^2\sin 2\theta}{4} + C\,,$$

which is still in terms of θ. To put this back in terms of u, use $\theta = \cos^{-1}(\frac{u}{a})$, the double-angle formula $\sin 2\theta = 2\sin\theta\cos\theta$, and $\sqrt{a^2 - u^2} = \sqrt{a^2\sin^2\theta} = a\sin\theta$. Then

$$\int \sqrt{a^2 - u^2}\, du = -\frac{a^2}{2}\cos^{-1}\left(\frac{u}{a}\right) + \frac{2a^2\sin\theta\cos\theta}{4}$$

$$= -\frac{a^2}{2}\cos^{-1}\left(\frac{u}{a}\right) + \frac{(a\cos\theta)(a\sin\theta)}{2} + C$$

which results in the following formula:

$$\int \sqrt{a^2 - u^2}\, du = -\frac{a^2}{2}\cos^{-1}\left(\frac{u}{a}\right) + \frac{1}{2}u\sqrt{a^2 - u^2} + C \qquad (6.6)$$

It is left as an exercise to show that the substitution $u = a\sin\theta$ gives:

$$\int \sqrt{a^2 - u^2}\, du = \frac{a^2}{2}\sin^{-1}\left(\frac{u}{a}\right) + \frac{1}{2}u\sqrt{a^2 - u^2} + C \qquad (6.7)$$

That these two seemingly different antiderivatives are equivalent follows immediately from the identity $\sin^{-1}x + \cos^{-1}x = \frac{\pi}{2}$ for all $-1 \le x \le 1$, which shows that the antiderivatives differ by the constant $\frac{\pi a^2}{4}$ (absorbed in the generic constant C):

$$\frac{a^2}{2}\sin^{-1}\left(\frac{u}{a}\right) + \frac{1}{2}u\sqrt{a^2 - u^2} + C = \frac{a^2}{2}\left(\frac{\pi}{2} - \cos^{-1}\left(\frac{u}{a}\right)\right) + \frac{1}{2}u\sqrt{a^2 - u^2} + C$$

$$= -\frac{a^2}{2}\cos^{-1}\left(\frac{u}{a}\right) + \frac{1}{2}u\sqrt{a^2 - u^2} + \underbrace{C + \frac{\pi a^2}{4}}_{C}$$

Thus, either substitution—$u = a\cos\theta$ or $u = a\sin\theta$—can be used when evaluating the integral $\int \sqrt{a^2 - u^2}\, du$. The latter choice is sometimes preferred, to avoid the negative sign in du and the resulting formula.

Example 6.16

Evaluate $\int \sqrt{9-4x^2}\,dx$.

Solution: The integrand is of the form $\sqrt{a^2-u^2}$ with $a = 3$ and $u = 2x$, so that $du = 2dx$. Then $dx = \frac{1}{2}du$ and so:

$$\begin{aligned}\int \sqrt{9-4x^2}\,dx &= \frac{1}{2}\int \sqrt{a^2-u^2}\,du \\ &= \frac{1}{2}\left(\frac{a^2}{2}\sin^{-1}\left(\frac{u}{a}\right) + \frac{1}{2}u\sqrt{a^2-u^2}\right) + C \\ &= \frac{9}{4}\sin^{-1}\left(\frac{2x}{3}\right) + \frac{1}{2}x\sqrt{9-4x^2} + C\end{aligned}$$

In general, when other methods fail, use the table below as a guide for certain types of integrals, making use of the specified substitution and trigonometric identity:

Integral contains	Substitution	Identity
$\sqrt{a^2 - u^2}$	$u = a\sin\theta$	$1 - \sin^2\theta = \cos^2\theta$
$\sqrt{a^2 + u^2}$	$u = a\tan\theta$	$1 + \tan^2\theta = \sec^2\theta$
$\sqrt{u^2 - a^2}$	$u = a\sec\theta$	$\sec^2\theta - 1 = \tan^2\theta$

For example, the substitution $u = a\tan\theta$ leads to the following formula:

$$\int \sqrt{a^2+u^2}\,du = \frac{1}{2}u\sqrt{a^2+u^2} + \frac{a^2}{2}\ln\left|u+\sqrt{a^2+u^2}\right| + C \tag{6.8}$$

Similarly, the substitution $u = a\sec\theta$ yields this formula:

$$\int \sqrt{u^2-a^2}\,du = \frac{1}{2}u\sqrt{u^2-a^2} - \frac{a^2}{2}\ln\left|u+\sqrt{u^2-a^2}\right| + C \tag{6.9}$$

The proof of each formula requires this result from Example 6.7 in Section 6.1:

$$\int \sec^3\theta\,d\theta = \frac{1}{2}(\sec\theta\tan\theta + \ln|\sec\theta + \tan\theta|) + C \tag{6.10}$$

The above substitutions can be used even if no square roots are present.

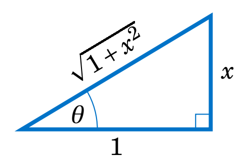
Example 6.17

Evaluate $\int \frac{dx}{(1+x^2)^2}$.

Solution: Notice that this integral cannot be evaluated by using the Power Formula with the substitution $u = 1+x^2$ (why?). Integration by parts does not look promising, either. So try a trigonometric substitution. The integrand contains a term of the form $a^2 + u^2$ (with $a = 1$ and $u = x$), so use the substitution $x = \tan\theta$. Then $dx = \sec^2\theta\, d\theta$ and so

$$\begin{aligned}
\int \frac{dx}{(1+x^2)^2} &= \int \frac{\sec^2\theta\; d\theta}{(1+\tan^2\theta)^2} \\
&= \int \frac{\sec^2\theta\, d\theta}{(\sec^2\theta)^2} \\
&= \int \frac{d\theta}{\sec^2\theta^2} \\
&= \int \cos^2\theta\; d\theta \\
&= \int \frac{1+\cos 2\theta}{2}\, d\theta \\
&= \frac{\theta}{2} + \frac{1}{4}\sin 2\theta + C \\
&= \frac{\theta}{2} + \frac{1}{2}\sin\theta\,\cos\theta + C
\end{aligned}$$

by the trigonometric double-angle identity $\sin 2\theta = 2\sin\theta\cos\theta$.



The simplest way to get expressions for $\sin\theta$ and $\cos\theta$ in terms of x is to draw a right triangle with an angle θ such that $\tan\theta = x = \frac{x}{1}$, as in the drawing on the right. The hypotenuse must then be $\sqrt{1+x^2}$ (by the Pythagorean Theorem), which makes it easy to read off the values of $\sin\theta$ and $\cos\theta$:

$$\sin\theta = \frac{x}{\sqrt{1+x^2}} \qquad \text{and} \qquad \cos\theta = \frac{1}{\sqrt{1+x^2}}$$

Since $\theta = \tan^{-1} x$, putting the integral back in terms of x yields:

$$\begin{aligned}
\int \frac{dx}{(1+x^2)^2} &= \frac{1}{2}\tan^{-1} x + \frac{1}{2}\,\frac{x}{\sqrt{1+x^2}}\,\frac{1}{\sqrt{1+x^2}} + C \\
&= \frac{1}{2}\tan^{-1} x + \frac{x}{2(1+x^2)} + C
\end{aligned}$$

Note: An alternative method for getting $\sin\theta$ and $\cos\theta$ in terms of x would be to put $\tan\theta = x$ in the identity $\sec^2\theta = 1+\tan^2\theta$ to solve for $\cos\theta$, then use the identity $\sin^2\theta = 1-\cos^2\theta$ to solve for $\sin\theta$.

By completing the square, quadratic expressions in x can be put in one of the forms $a^2 \pm u^2$ or $u^2 - a^2$, enabling the use of the corresponding trigonometric substitution.

Example 6.18

Evaluate $\displaystyle\int \frac{dx}{(4x^2+8x-5)^{3/2}}$.

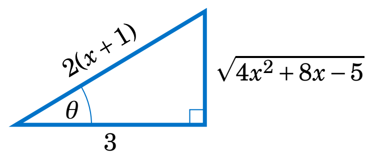
Solution: This integral cannot be evaluated by using the Power Formula, so try a trigonometric substitution. Complete the square on the expression $4x^2+8x-5$:

$$4x^2 + 8x - 5 = 4(x^2+2x) - 5 = 4(x^2+2x+1) - 5 - 4 = 4(x+1)^2 - 9$$

This expression is now of the form u^2-a^2 for $u=2(x+1)$ and $a=3$. Use the substitution $u=a\sec\theta$, which means $2(x+1)=3\sec\theta$. Then $2\,dx=3\sec\theta\tan\theta\,d\theta$ and so:

$$\begin{aligned}\int \frac{dx}{(4x^2+8x-5)^{3/2}} &= \int \frac{dx}{(4(x+1)^2 - 9)^{3/2}}\\ &= \frac{3}{2}\int \frac{\sec\theta\tan\theta\,d\theta}{(9\sec^2\theta - 9)^{3/2}} = \frac{3}{2}\int \frac{\sec\theta\tan\theta\,d\theta}{(9(\sec^2\theta - 1))^{3/2}}\\ &= \frac{1}{18}\int \frac{\sec\theta\tan\theta\,d\theta}{\tan^3\theta} = \frac{1}{18}\int \frac{\sec\theta\,d\theta}{\tan^2\theta} = \frac{1}{18}\int \frac{\cos\theta\,d\theta}{\sin^2\theta}\\ &= \frac{1}{18}\int \csc\theta\cot\theta\,d\theta = -\frac{1}{18}\csc\theta + C\end{aligned}$$

To get an expression for $\csc\theta$ in terms of x, draw a right triangle with an angle θ such that $\sec\theta = \frac{2(x+1)}{3}$, as in the drawing on the right. The side opposite θ must then be $\sqrt{4x^2+8x-5}$ (by the Pythagorean Theorem), and hence:



$$\csc\theta = \frac{2(x+1)}{\sqrt{4x^2+8x-5}}$$

Putting the integral back in terms of x yields:

$$\begin{aligned}\int \frac{dx}{(4x^2+8x-5)^{3/2}} &= -\frac{1}{18}\,\frac{2(x+1)}{\sqrt{4x^2+8x-5}} + C\\ &= -\frac{x+1}{9\sqrt{4x^2+8x-5}} + C\end{aligned}$$

Note: Trigonometric identities could have been used to obtain $\csc\theta$ by knowing $\sec\theta$.

The following integrals from Section 5.4 might be helpful for the exercises:

$$\int \tan u\,du = \ln|\sec u| + C \qquad (6.11)$$

$$\int \sec u\,du = \ln|\sec u + \tan u| + C \qquad (6.12)$$

$$\int \csc u\,du = -\ln|\csc u + \cot u| + C \qquad (6.13)$$

Exercises

A

For Exercises 1-16, evaluate the given integral.

1. $\int \sqrt{9+4x^2}\,dx$ **2.** $\int \sqrt{2-3x^2}\,dx$ **3.** $\int \sqrt{4x^2-9}\,dx$ **4.** $\int \sqrt{x^2+2x+10}\,dx$

5. $\int \frac{\sqrt{1-x^2}}{x^2}\,dx$ **6.** $\int \frac{x^2\,dx}{\sqrt{x^2-9}}$ **7.** $\int \frac{dx}{x\sqrt{1+x^2}}$ **8.** $\int \frac{dx}{x^2\sqrt{a^2+x^2}}$ $(a>0)$

9. $\int \frac{x^3\,dx}{\sqrt{x^2+4}}$ **10.** $\int \frac{dx}{(4x^2-9)^{3/2}}$ **11.** $\int \frac{dx}{(9+4x^2)^2}$ **12.** $\int \frac{x^2\,dx}{\sqrt{a^2-x^2}}$ $(a>0)$

13. $\int \frac{x^3\,dx}{\sqrt{9-x^2}}$ **14.** $\int \frac{\sqrt{4-x^2}}{x}\,dx$ **15.** $\int \frac{(x-4)\,dx}{\sqrt{-9x^2+36x-32}}$ **16.** $\int \frac{dx}{(4x^2+16x+15)^{3/2}}$

17. Prove formula (6.7) directly by using the substitution $u = a\sin\theta$.

18. Prove formula (6.8). **19.** Prove formula (6.9).

20. Show that using the substitution $u = a\cot\theta$ to evaluate the integral $\int \sqrt{a^2+u^2}\,dx$ leads to an antiderivative equivalent to the one in formula (6.8).

21. Show that using the substitution $u = a\csc\theta$ to evaluate the integral $\int \sqrt{u^2-a^2}\,dx$ leads to an antiderivative equivalent to the one in formula (6.9).

B

22. The integrals $\int \frac{dx}{\sqrt{x^2 \pm a^2}}$ can be evaluated without the use of trigonometric substitutions, by using differentials:

(a) For $u^2 = x^2 \pm a^2$, show that

$$\frac{dx}{u} = \frac{d\,(x+u)}{x+u}\;.$$

(b) Integrate both sides of the result from part (a).

Note: In general, many integrals involving $\sqrt{x^2 \pm a^2}$ can be handled with a similar manipulation of differentials, with varying complexity.

23. According to Newtonian physics the path of a photon grazing the surface of the Sun should be deflected by the Sun's gravitational field by an angle θ, given approximately by

$$\theta = \left|\frac{2GMR}{c^2}\int_{\infty}^{0} \frac{dy}{(R^2+y^2)^{3/2}}\right|$$

where $c = 2.998\times10^8$ m/s is the speed of light, $G = 6.67\times10^{-11}$ N/m^2/kg^2 is the gravitational constant, $M = 1.99\times10^{30}$ kg is the mass of the Sun, and $R = 6.96\times10^8$ m is the radius of the Sun. Show that

$$\theta = \frac{2GM}{c^2R} = 4.24\times10^{-6}\text{ radians} = 2.43\times10^{-4}\text{ degrees} \approx 0.875\text{ seconds of arc,}$$

where 1 second of arc = 1/3600 of 1 degree.[3]

[3] Albert Einstein published this result in 1911, then showed in 1915 that the true angle should be double that amount, due to the curvature of space. Experiments verified Einstein's prediction. See pp.69-71 in SERWAY, R.A., C.J. MOSES AND C.A. MOYER, *Modern Physics*, Orlando, FL: Harcourt Brace Jovanovich Publishers, 1989.

6.4 Partial Fractions

In the last two sections some trigonometric integrals were simplified by using various trigonometric identities. For integrals of rational functions—quotients of polynomials—some algebraic identities (e.g. $x^2 - a^2 = (x-a)(x+a)$) will be useful in the method of **partial fractions**. The idea behind this method is simple: replace a complicated rational function with simpler ones that are easy to integrate.

For example, there is no formula for evaluating the integral

$$\int \frac{dx}{x^2+x}$$

but notice that you can write

$$\frac{1}{x^2+x} = \frac{1}{x(x+1)} = \frac{1}{x} - \frac{1}{x+1},$$

called the **partial fraction decomposition** of $\frac{1}{x^2+x}$, so that

$$\int \frac{dx}{x^2+x} = \int \left(\frac{1}{x} - \frac{1}{x+1}\right) dx$$

$$= \ln|x| - \ln|x+1| + C .$$

There is a systematic way to find this decomposition. First, assume that

$$\frac{1}{x(x+1)} = \frac{A}{x} + \frac{B}{x+1}$$

for some constants A and B. Get a common denominator on the right side:

$$\frac{1}{x(x+1)} = \frac{A(x+1) + Bx}{x(x+1)}$$

$$\frac{0x + 1}{x(x+1)} = \frac{(A+B)x + A}{x(x+1)}$$

Now equate coefficients in the numerators of both sides to solve for A and B:

$$\text{constant term}: \quad A = 1$$

$$\text{coefficient of } x: \quad A + B = 0 \quad \Rightarrow \quad B = -A = -1$$

Thus,

$$\frac{1}{x(x+1)} = \frac{A}{x} + \frac{B}{x+1} = \frac{1}{x} + \frac{-1}{x+1} = \frac{1}{x} - \frac{1}{x+1}$$

as before.

The partial fraction method can be discussed in general, and its assumptions proved[4], but only the simplest cases—linear and quadratic factors— will be considered here. In all cases it will be assumed that the degree of the polynomial in the numerator of the rational function is less than the degree of the polynomial in the denominator.

Start with the most basic case, similar to the example above:

Case 1 - Distinct linear factors: A rational function $\frac{p(x)}{q(x)}$ such that degree$(p(x))$ < degree$(q(x))$, where $q(x)$ is a product of $n > 1$ distinct linear factors

$$q(x) = (a_1x+b_1)(a_2x+b_2)\cdots(a_nx+b_n)\,,$$

can be written as a sum of partial fractions

$$\frac{p(x)}{q(x)} = \frac{A_1}{a_1x+b_1} + \frac{A_2}{a_2x+b_2} + \cdots + \frac{A_n}{a_nx+b_n}$$

for some constants $A_1, A_2, \ldots, A_n$. Those constants can be solved for by getting a common denominator on the right side of the equation and then equating the coefficients of the numerators of both sides.

Example 6.19

Evaluate $\displaystyle\int \frac{dx}{x^2-7x+10}$.

Solution: Since $x^2-7x+10 = (x-2)(x-5)$, then

$$\frac{1}{x^2-7x+10} = \frac{1}{(x-2)(x-5)} = \frac{A}{x-2} + \frac{B}{x-5}$$
$$= \frac{(A+B)x + (-5A-2B)}{(x-2)(x-5)}$$

so that

$$\text{coefficient of } x: \quad A + B = 0 \quad\Rightarrow\quad B = -A$$
$$\text{constant term}: \quad -5A - 2B = 1 \quad\Rightarrow\quad -5A + 2A = 1 \quad\Rightarrow\quad A = -\frac{1}{3} \text{ and } B = \frac{1}{3}$$

Thus,

$$\int \frac{dx}{x^2-7x+10} = \int \left(\frac{-\frac{1}{3}}{x-2} + \frac{\frac{1}{3}}{x-5}\right) dx$$
$$= -\frac{1}{3}\ln|x-2| + \frac{1}{3}\ln|x-5| + C$$

[4]For example, see Section 5.10 in HILLMAN, A.P., AND G.L. ALEXANDERSON, *A First Undergraduate Course in Abstract Algebra*, 3rd ed., Belmont, CA: Wadsworth Publishing Co., 1983.

If one linear factor in the denominator is repeated more than once, and all other factors are distinct, then use the following decomposition:

Case 2 - One repeated linear factor + distinct linear factors: A rational function $\frac{p(x)}{q(x)}$ such that degree$(p(x)) <$ degree$(q(x))$, where $q(x)$ is a product of $n > 1$ distinct linear factors and one linear factor repeated $m > 1$ times

$$q(x) = (ax+b)^m (a_1x+b_1)(a_2x+b_2) \cdots (a_nx+b_n),$$

can be written as a sum of partial fractions

$$\frac{p(x)}{q(x)} = \frac{A_1}{ax+b} + \frac{A_2}{(ax+b)^2} + \cdots + \frac{A_m}{(ax+b)^m} + \frac{B_1}{a_1x+b_1} + \cdots + \frac{B_n}{a_nx+b_n}$$

for some constants $A_1, A_2, \ldots, A_m$ and $B_1, \ldots, B_n$. Those constants can be solved for by the same method as in Case 1.

Example 6.20

Evaluate $\int \frac{x^2+x-1}{x^3+x^2}\,dx$.

Solution: Since $x^3+x^2 = x^2(x+1)$, then

$$\begin{aligned}\frac{x^2+x-1}{x^3+x^2} = \frac{x^2+x-1}{x^2(x+1)} &= \frac{A}{x} + \frac{B}{x^2} + \frac{C}{x+1}\\ &= \frac{Ax(x+1) + B(x+1) + Cx^2}{x^2(x+1)}\\ &= \frac{(A+C)x^2 + (A+B)x + B}{x^2(x+1)}\end{aligned}$$

so that

$$\begin{aligned}\text{constant term:}\quad & B = -1\\ \text{coefficient of } x:\quad & A + B = 1 \quad\Rightarrow\quad A = 1 - B = 2\\ \text{coefficient of } x^2:\quad & A + C = 1 \quad\Rightarrow\quad C = 1 - A = -1\end{aligned}$$

Thus,

$$\begin{aligned}\int \frac{x^2+x-1}{x^3+x^2}\,dx &= \int \left(\frac{2}{x} + \frac{-1}{x^2} + \frac{-1}{x+1}\right) dx\\ &= 2\ln|x| + \frac{1}{x} - \ln|x+1| + C_0 \quad (C_0 = \text{ generic constant})\end{aligned}$$

Case 2 can be extended to more than one repeated factor—the partial fraction decomposition would then have more terms similar to those for the first repeated factor.

Example 6.21

Evaluate $\displaystyle\int \frac{dx}{x^2(x+1)^2}$.

Solution: Expanding Case 2 to two repeated factors,

$$\frac{1}{x^2(x+1)^2} = \frac{A}{x} + \frac{B}{x^2} + \frac{C}{x+1} + \frac{D}{(x+1)^2}$$

$$= \frac{Ax(x+1)^2 + B(x+1)^2 + Cx^2(x+1) + Dx^2}{x^2(x+1)^2}$$

$$= \frac{(A+C)x^3 + (2A+B+C+D)x^2 + (A+2B)x + B}{x^2(x+1)}$$

so that

$$\begin{array}{lrcl} \text{constant term:} & B = 1 & & \\ \text{coefficient of } x: & A + 2B = 0 & \Rightarrow & A = -2B = -2 \\ \text{coefficient of } x^3: & A + C = 0 & \Rightarrow & C = -A = 2 \\ \text{coefficient of } x^2: & 2A + B + C + D = 0 & \Rightarrow & D = -2A - B - C = 1 \end{array}$$

Thus,

$$\int \frac{dx}{x^2(x+1)^2} = \int \left(\frac{-2}{x} + \frac{1}{x^2} + \frac{2}{x+1} + \frac{1}{(x+1)^2} \right) dx$$

$$= -2\ln|x| - \frac{1}{x} + 2\ln|x+1| - \frac{1}{x+1} + C_0 \quad (C_0 = \text{ generic constant})$$

The partial fraction decompositions for quadratic factors are similar to those for linear factors, except the numerators in each partial fraction can now contain linear terms. A factor of the form ax^2+bx+c is considered quadratic only if it cannot be factored into a product of linear terms (i.e. has no real roots) and $a \neq 0$.

Case 3 - Distinct quadratic factors: A rational function $\frac{p(x)}{q(x)}$ such that degree$(p(x)) <$ degree$(q(x))$, and $q(x)$ is a product of $n > 1$ distinct quadratic factors

$$q(x) = (a_1x^2 + b_1x + c_1)(a_2x^2 + b_2x + c_2) \cdots (a_nx^2 + b_nx + c_n),$$

can be written as a sum of partial fractions

$$\frac{p(x)}{q(x)} = \frac{A_1x + B_1}{a_1x^2 + b_1x + c_1} + \frac{A_2x + B_2}{a_2x^2 + b_2x + c_2} + \cdots + \frac{A_nx + B_n}{a_nx^2 + b_nx + c_n}$$

for some constants $A_1, A_2, \ldots, A_n$ and $B_1, B_2, \ldots, B_n$. Those constants can be solved for by the same method as in Case 1.

Example 6.22

Evaluate $\displaystyle\int \frac{dx}{(x^2+1)(x^2+4)}$.

Solution: Neither x^2+1 nor x^2+4 has real roots, so by Case 3,

$$\begin{aligned}\frac{1}{(x^2+1)(x^2+4)} &= \frac{Ax+B}{x^2+1} + \frac{Cx+D}{x^2+4}\\ &= \frac{(Ax+B)(x^2+4) + (Cx+D)(x^2+1)}{(x^2+1)(x^2+4)}\\ &= \frac{(A+C)x^3 + (B+D)x^2 + (4A+C)x + (4B+D)}{(x^2+1)(x^2+4)}\end{aligned}$$

so that

$$\begin{aligned}&\text{coefficient of } x^3: & A + C &= 0 &\Rightarrow\quad & C = -A\\ &\text{coefficient of } x^2: & B + D &= 0 &\Rightarrow\quad & D = -B\\ &\text{coefficient of } x: & 4A + C &= 0 &\Rightarrow\quad & 4A - A = 0 \quad\Rightarrow\quad A = 0\\ &\text{constant term}: & 4B + D &= 1 &\Rightarrow\quad & 4B - B = 1 \quad\Rightarrow\quad B = \frac{1}{3} \text{ and } D = -\frac{1}{3}\end{aligned}$$

Thus,

$$\begin{aligned}\int \frac{dx}{(x^2+1)(x^2+4)} &= \int \left(\frac{\frac{1}{3}}{x^2+1} + \frac{-\frac{1}{3}}{x^2+4}\right) dx\\ &= \frac{1}{3}\tan^{-1}x - \frac{1}{6}\tan^{-1}\left(\frac{x}{2}\right) + C_0\end{aligned}$$

by formula (5.4) in Section 5.4.

A repeated quadratic factor is handled in the same way as a repeated linear factor:

Case 4 - One repeated quadratic factor + distinct quadratic factors: A rational function $\frac{p(x)}{q(x)}$ such that degree$(p(x)) <$ degree$(q(x))$, where $q(x)$ is a product of $n > 1$ distinct quadratic factors and one quadratic factor repeated $m > 1$ times

$$q(x) = (ax^2+bx+c)^m (a_1x^2+b_1x+c_1)(a_2x^2+b_2x+c_2)\cdots(a_nx^2+b_nx+c_n),$$

can be written as a sum of partial fractions

$$\frac{p(x)}{q(x)} = \frac{A_1x+B_1}{ax^2+bx+c} + \cdots + \frac{A_mx+B_m}{(ax^2+bx+c)^m} + \frac{C_1x+D_1}{a_1x^2+b_1x+c_1} + \cdots + \frac{C_nx+D_n}{a_nx^2+b_nx+c_n}$$

for some constants $A_1, \ldots, A_m$, $B_1, \ldots, B_m$, $C_1, \ldots, C_n$, and $D_1, \ldots, D_n$. Those constants can be solved for by the same method as in Case 1.

Example 6.23

Evaluate $\displaystyle\int \frac{dx}{(x^2+1)^2(x^2+4)}$.

Solution: Neither x^2+1 nor x^2+4 has real roots, and x^2+1 is repeated, so by Case 4,

$$\frac{1}{(x^2+1)^2(x^2+4)} = \frac{Ax+B}{x^2+1} + \frac{Cx+D}{(x^2+1)^2} + \frac{Ex+F}{x^2+4}$$

$$= \frac{(Ax+B)(x^2+1)(x^2+4) + (Cx+D)(x^2+4) + (Ex+F)(x^2+1)^2}{(x^2+1)(x^2+4)}$$

with the right side of the equation expanded as

$$\frac{(A+E)x^5 + (B+F)x^4 + (5A+C+2E)x^3 + (5B+D+2F)x^2 + (4A+4C+E)x + (4B+4D+F)}{(x^2+1)^2(x^2+4)}$$

so that equating coefficients of both sides gives

coefficient of x^5: $A+E=0 \Rightarrow E=-A$

coefficient of x^4: $B+F=0 \Rightarrow F=-B$

coefficient of x^3: $5A+C+2E=0 \Rightarrow 5A+C-2A=0 \Rightarrow C=-3A$

coefficient of x^2: $5B+D+2F=0 \Rightarrow 5B+D-2F=0 \Rightarrow D=-3B$

coefficient of x: $4A+4C+E=0 \Rightarrow 4A-12A-A=0 \Rightarrow A=0 \Rightarrow C=0$ and $E=0$

constant term: $4B+4D+F=1 \Rightarrow 4B-12B-B=1 \Rightarrow B=-\frac{1}{9} \Rightarrow D=\frac{1}{3}$ and $F=\frac{1}{9}$

Thus,

$$\int \frac{dx}{(x^2+1)^2(x^2+4)} = \int\left(\frac{-\frac{1}{9}}{x^2+1} + \frac{\frac{1}{3}}{(x^2+1)^2} + \frac{\frac{1}{9}}{x^2+4}\right)dx$$

$$= -\frac{1}{9}\tan^{-1}x + \frac{1}{3}\left(\frac{1}{2}\tan^{-1}x + \frac{x}{2(x^2+1)}\right) + \frac{1}{18}\tan^{-1}\left(\frac{x}{2}\right) + C_0$$

$$= \frac{1}{18}\tan^{-1}x + \frac{x}{6(x^2+1)} + \frac{1}{18}\tan^{-1}\left(\frac{x}{2}\right) + C_0$$

where the middle integral on the right is from Example 6.17 in Section 6.3.

When a rational function has a numerator with degree larger than its denominator, dividing the numerator by the denominator leaves the sum of a polynomial and a new rational function perhaps satisfying the conditions for Cases 1-4. When the numerator and denominator have the same degree, a trick like this might be easier.

$$\frac{x^2+2}{(x+1)(x+2)} = \frac{(x^2+3x+2)-3x}{(x+1)(x+2)} = \frac{x^2+3x+2}{(x+1)(x+2)} - \frac{3x}{(x+1)(x+2)} = 1 - \frac{3x}{(x+1)(x+2)}$$

The last rational function on the right can be integrated using partial fractions.

Exercises

A

For Exercises 1-12, evaluate the given integral.

1. $\displaystyle\int \frac{dx}{x^2 - x}$ **2.** $\displaystyle\int \frac{x+1}{x^2 - x}\,dx$ **3.** $\displaystyle\int \frac{dx}{2x^2 + 3x - 2}$ **4.** $\displaystyle\int \frac{dx}{x^2 + x - 6}$

5. $\displaystyle\int \frac{dx}{x^4 - x^2}$ **6.** $\displaystyle\int \frac{x}{(x-2)^3}\,dx$ **7.** $\displaystyle\int \frac{x-1}{x^2(x+1)}\,dx$ **8.** $\displaystyle\int \frac{x^2}{(x-1)^2}\,dx$

9. $\displaystyle\int \frac{x-2}{x^2(x-1)^2}\,dx$ **10.** $\displaystyle\int \frac{dx}{x^4 + x^2}$ **11.** $\displaystyle\int \frac{dx}{x^4 + 5x^2 + 4}$ **12.** $\displaystyle\int \frac{(x-1)^2}{(x^2+1)^2}\,dx$

B

13. For all numbers $a \neq b$ show that

$$\int \frac{dx}{(x-a)(x-b)} = \frac{1}{a-b} \ln\left|\frac{x-a}{x-b}\right| + C\,.$$

14. Let $q(x) = (x-a_1)(x-a_2)$, where $a_1 \neq a_2$. Show that

$$\frac{q'(x)}{q(x)} = \frac{1}{x-a_1} + \frac{1}{x-a_2}\,.$$

15. For $q(x)$ as in Exercise 14, show that

$$\frac{1}{q(x)} = \frac{1}{q'(a_1)(x-a_1)} + \frac{1}{q'(a_2)(x-a_2)}\,.$$

16. Extend Exercise 15 to three distinct linear factors: if $q(x) = (x-a_1)(x-a_2)(x-a_3)$ then

$$\frac{1}{q(x)} = \frac{1}{q'(a_1)(x-a_1)} + \frac{1}{q'(a_2)(x-a_2)} + \frac{1}{q'(a_3)(x-a_3)}\,.$$

This result can be extended to any $n \geq 2$ distinct factors, though you do not need to prove that.

17. Find $\displaystyle\frac{d^{100}}{dx^{100}}\left(\frac{1}{x^2 - 9x + 20}\right)$. **18.** Find $\displaystyle\frac{d^{2020}}{dx^{2020}}\left((x^2-1)^{-1}\right)$.

19. It is possible to use a form of partial fractions to evaluate integrals that are not rational functions. For example, evaluate the integral

$$\int \frac{dx}{x + x^{4/3}}$$

by finding constants A, B and C such that

$$\frac{1}{x + x^{4/3}} = \frac{1}{x(1+x^{1/3})} = \frac{A}{x} + \frac{B x^{-2/3} + C}{1 + x^{1/3}}\,.$$

Notice how in the second partial fraction the highest power of x in the numerator is one less than in the denominator, similar to a partial fraction for a quadratic factor in a rational function.

C

20. Find a different way to evaluate the integral in Exercise 19.

6.5 Miscellaneous Integration Methods

The integration methods presented so far are considered "standard," meaning every calculus student should know them. This section will discuss a few additional methods, some more common than others. One such method is the **Leibniz integral rule** for "differentiation under the integral sign."[5] This powerful and useful method is best explained with a simple example.

Recall from Section 5.4 that

$$\int e^{\alpha x}\,dx = \tfrac{1}{\alpha}e^{\alpha x} + C \tag{6.14}$$

for any constant $\alpha \neq 0$. This antiderivative involves the variable x and the constant α. The idea behind the Leibniz rule is to reverse those roles: view α as the variable and x as a constant. The antiderivative is then seen as a function of α, and so its derivative can be taken with respect to α. The big leap that this method makes is to move the differentiation operation *inside* the integral:[6]

$$\frac{d}{d\alpha}\int e^{\alpha x}\,dx = \int \frac{d}{d\alpha}(e^{\alpha x})\,dx = \int x\,e^{\alpha x}\,dx$$

However, differentiating the right side of formula (6.14) shows that

$$\frac{d}{d\alpha}\int e^{\alpha x}\,dx = \frac{d}{d\alpha}\left(\tfrac{1}{\alpha}e^{\alpha x} + C\right) = \frac{\alpha\,(x\,e^{\alpha x}) - 1\cdot e^{\alpha x}}{\alpha^2} = \tfrac{1}{\alpha}x\,e^{\alpha x} - \tfrac{1}{\alpha^2}e^{\alpha x}$$

Thus,

$$\int x\,e^{\alpha x}\,dx = \tfrac{1}{\alpha}x\,e^{\alpha x} - \tfrac{1}{\alpha^2}e^{\alpha x} + C$$

which can be verified via integration by parts with the tabular method:

	u	dv		
	x	$e^{\alpha x}\,dx$		
	1	$\frac{1}{\alpha}e^{\alpha x}$	(+)	$\longrightarrow + (x)(\frac{1}{\alpha}e^{\alpha x})$
STOP $\longrightarrow$	0	$\frac{1}{\alpha^2}e^{\alpha x}$	(–)	$\longrightarrow - (1)(\frac{1}{\alpha^2}e^{\alpha x})$

$$\int x\,e^{\alpha x}\,dx = \tfrac{1}{\alpha}x\,e^{\alpha x} - \tfrac{1}{\alpha^2}e^{\alpha x} + C \quad \checkmark$$

[5]The renowned physicist Richard Feynman (1918-1988) famously lamented that the technique was no longer being taught. See p.72 in FEYNMAN, R.P., *Surely You're Joking, Mr. Feynman!*, New York: Bantam Books, 1986.

[6]It can be proved that this is valid when the derivative of the integrand is a continuous function of α, which will always be the case in this book. See pp.121-122 in SOKOLNIKOFF, I.S., *Advanced Calculus*, New York: McGraw-Hill Book Company, Inc., 1939.

What was actually done in the above example? A *known* integral,

$$\int e^{\alpha x}\,dx \;=\; \tfrac{1}{\alpha}\,e^{\alpha x} \;+\; C\;,$$

was differentiated with respect to α via the Leibniz rule to produce a *new* integral,

$$\int x\,e^{\alpha x}\,dx \;=\; \tfrac{1}{\alpha}\,x\,e^{\alpha x} \;-\; \tfrac{1}{\alpha^2}\,e^{\alpha x} \;+\; C\;,$$

with the constant α treated temporarily—only during the differentiation—as a variable. In general, that is how the Leibniz rule is used. Typically this means if you want to evaluate a certain integral with the Leibniz rule, then you "work backwards" to figure out which integral you need to differentiate with respect to some constant (e.g. α) in the integrand.

Example 6.24

Use the Leibniz rule to evaluate $\displaystyle\int \frac{dx}{(1+x^2)^2}$.

Solution: By formula (5.4) in Section 5.4,

$$\int \frac{dx}{a^2+x^2} \;=\; \tfrac{1}{a}\,\tan^{-1}\left(\tfrac{x}{a}\right) \;+\; C$$

for any constant $a > 0$. So differentiate both sides with respect to a:

$$\begin{aligned}
\frac{d}{da}\int \frac{dx}{a^2+x^2} \;&=\; \frac{d}{da}\left(\tfrac{1}{a}\,\tan^{-1}\left(\tfrac{x}{a}\right) \;+\; C\right)\\
\int \frac{d}{da}\left(\frac{1}{a^2+x^2}\right)dx \;&=\; -\tfrac{1}{a^2}\,\tan^{-1}\left(\tfrac{x}{a}\right) \;+\; \tfrac{1}{a}\cdot\frac{1}{1+\left(\tfrac{x}{a}\right)^2}\cdot -\tfrac{x}{a^2}\\
\int -\frac{2a}{(a^2+x^2)^2}\,dx \;&=\; -\tfrac{1}{a^2}\,\tan^{-1}\left(\tfrac{x}{a}\right) \;-\; \frac{x}{a\,(a^2+x^2)}\\
\int \frac{dx}{(a^2+x^2)^2} \;&=\; \tfrac{1}{2a^3}\,\tan^{-1}\left(\tfrac{x}{a}\right) \;+\; \frac{x}{2a^2\,(a^2+x^2)} \;+\; C
\end{aligned}$$

That general formula is useful in itself. In particular, for $a = 1$,

$$\int \frac{dx}{(1+x^2)^2} \;=\; \tfrac{1}{2}\,\tan^{-1}x \;+\; \frac{x}{2(1+x^2)} \;+\; C\;,$$

which agrees with the result from Example 6.17 in Section 6.3.
Notice that there was no generic constant (e.g. a or α) in the statement of the problem. When that happens, you will need to figure out where the constant *should* be in order to use the Leibniz rule.

You can also use differentiation under the integral sign to evaluate definite integrals.

Example 6.25

Show that $\displaystyle\int_0^\infty e^{-x^2}\,dx \;=\; \tfrac{1}{2}\sqrt{\pi}$.

Solution: Let $I = \int_0^\infty e^{-x^2}\,dx$. The integral is convergent, since by Exercise 11 in Section 4.4, for all x

$$e^{x^2} \;\ge\; 1 + x^2 \quad\Rightarrow\quad 0 \;\le\; e^{-x^2} \;\le\; \frac{1}{1+x^2}$$

implies I is convergent by the Comparison Test, since $\int_0^\infty \frac{1}{1+x^2}\,dx$ is convergent (and equals $\frac{1}{2}\pi$) by Example 5.32 in Section 5.5. For $\alpha \ge 0$, define

$$\phi(\alpha) \;=\; \int_0^\infty \frac{\alpha\, e^{-\alpha^2 x^2}}{1+x^2}\,dx \;.$$

Then clearly $\phi(0) = 0$, and differentiating under the integral sign shows

$$\phi'(\alpha) \;=\; \int_0^\infty \frac{-2\alpha^2 e^{-\alpha^2x^2} + e^{-\alpha^2x^2}}{1+x^2}\,dx \qquad\Rightarrow\qquad \phi'(0) \;=\; \int_0^\infty \frac{dx}{1+x^2} \;=\; \tfrac{1}{2}\pi\;.$$

The substitution $y = \alpha x$, so that $dy = \alpha\,dx$, shows $\phi(\alpha)$ can be written as

$$\phi(\alpha) \;=\; \int_0^\infty \frac{e^{-y^2}}{1+\left(\frac{y}{\alpha}\right)^2}\,dy \qquad\Rightarrow\qquad 0 \;\le\; \lim_{\alpha\to\infty}\,\phi(\alpha) \;\le\; I \;<\; \infty\;.$$

Also, for $\alpha > 0$,

$$\begin{aligned}
\frac{d}{d\alpha}\left(\frac{1}{\alpha}\,e^{-\alpha^2}\,\phi(\alpha)\right) \;&=\; \frac{d}{d\alpha}\int_0^\infty \frac{e^{-\alpha^2(1+x^2)}}{1+x^2}\,dx \;=\; \int_0^\infty \frac{-2\alpha(1+x^2)e^{-\alpha^2(1+x^2)}}{1+x^2}\,dx\\
&=\; -2\alpha\, e^{-\alpha^2}\int_0^\infty e^{-\alpha^2x^2}\,dx \quad \text{, now substitute } u = \alpha x \text{ and } du = \alpha dx \text{ to get}\\
&=\; -2\alpha\, e^{-\alpha^2}\,\frac{1}{\alpha}\int_0^\infty e^{-u^2}\,du \;=\; -2e^{-\alpha^2}\,I \quad \text{, and so integrating both sides yields}
\end{aligned}$$

$$\int_0^\infty \frac{d}{d\alpha}\left(\frac{1}{\alpha}\,e^{-\alpha^2}\,\phi(\alpha)\right)d\alpha \;=\; -2I\int_0^\infty e^{-\alpha^2}\,d\alpha \;=\; -2I^2\;.$$

However, by the Fundamental Theorem of Calculus.

$$\begin{aligned}
\int_0^\infty \frac{d}{d\alpha}\left(\frac{1}{\alpha}\,e^{-\alpha^2}\,\phi(\alpha)\right)d\alpha \;&=\; \frac{1}{\alpha}\,e^{-\alpha^2}\,\phi(\alpha)\,\Bigg|_0^\infty \;=\; \left(\lim_{\alpha\to\infty}\frac{\phi(\alpha)}{\alpha\, e^{\alpha^2}}\right) - \left(\lim_{\alpha\to 0}\frac{\phi(\alpha)}{\alpha\, e^{\alpha^2}}\right)\\
&=\; 0 \;-\; \left(\lim_{\alpha\to 0}\frac{\phi(\alpha)}{\alpha\, e^{\alpha^2}}\right) \;\to\; \frac{0}{0} \quad \text{, so by L'Hôpital's Rule}\\
&=\; -\lim_{\alpha\to 0}\frac{\phi'(\alpha)}{e^{\alpha^2} + 2\alpha^2\,e^{\alpha^2}} \;=\; -\frac{\phi'(0)}{1+0} \;=\; -\tfrac{1}{2}\pi\;.
\end{aligned}$$

Thus,

$$-2I^2 \;=\; -\tfrac{1}{2}\pi \qquad\Rightarrow\qquad I \;=\; \tfrac{1}{2}\sqrt{\pi}$$

which is the desired result.

One immediate consequence of Example 6.25 is that

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx \;=\; \sqrt{\pi}$$

since e^{-x^2} is an even function. The following example shows another consequence, as well as how useful substitutions can be in writing integrals in a different form.

Example 6.26

Show that the Gamma function $\Gamma(t)$ can be written as

$$\Gamma(t) \;=\; 2\int_0^{\infty} y^{2t-1} e^{-y^2}\,dy \quad \text{for all } t > 0,$$

and that $\Gamma\left(\frac{1}{2}\right) \;=\; \sqrt{\pi}$.

Solution: Let $x = y^2$, so that $dx = 2y\,dy$. Then $x = 0 \;\Rightarrow\; y = 0$ and $x = \infty \;\Rightarrow\; y = \infty$, so

$$\Gamma(t) \;=\; \int_0^{\infty} x^{t-1} e^{-x}\,dx \;=\; \int_0^{\infty} (y^2)^{t-1} e^{-y^2}\, 2y\,dy \;=\; 2\int_0^{\infty} y^{2t-1} e^{-y^2}\,dy\,.$$

In this form, with the help of Example 6.25 it is now easy to evaluate $\Gamma\left(\frac{1}{2}\right)$:

$$\Gamma\left(\tfrac{1}{2}\right) \;=\; 2\int_0^{\infty} y^{1-1} e^{-y^2}\,dy \;=\; 2\int_0^{\infty} e^{-y^2}\,dy \;=\; 2\left(\tfrac{1}{2}\sqrt{\pi}\right) \;=\; \sqrt{\pi}$$

A function closely related to the Gamma function is the **Beta function** $B(x,y)$, defined by:

$$B(x,y) \;=\; \int_0^1 t^{x-1}(1-t)^{y-1}\,dt \qquad \text{for all } x > 0 \text{ and } y > 0 \tag{6.15}$$

It can be shown that[7]

$$B(x,y) \;=\; \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)} \qquad \text{for all } x > 0 \text{ and } y > 0. \tag{6.16}$$

Example 6.27

Show that the Beta function $B(x,y)$ can be written as

$$B(x,y) \;=\; \int_0^{\infty} \frac{u^{x-1}}{(1+u)^{x+y}}\,du\,.$$

Solution: Let $u = \frac{t}{1-t}$, so that $t = \frac{u}{1+u}$, $1-t = \frac{1}{1+u}$, and $dt = \frac{du}{(1+u)^2}$. Then $t = 0 \;\Rightarrow\; u = 0$ and $t = 1 \;\Rightarrow\; u = \infty$, so

$$B(x,y) \;=\; \int_0^1 t^{x-1}(1-t)^{y-1}\,dt \;=\; \int_0^{\infty} \left(\frac{u}{1+u}\right)^{x-1} \left(\frac{1}{1+u}\right)^{y-1} \frac{du}{(1+u)^2} \;=\; \int_0^{\infty} \frac{u^{x-1}}{(1+u)^{x+y}}\,du\,.$$

---

[7] See p.18-19 in RAINVILLE, E.D., *Special Functions*, New York: Chelsea Publishing Company, 1971.

Another application of substitutions in integrals is in the evaluation of **fractional derivatives**. Recall from Section 1.6 that the zero-th derivative of a function is just the function itself, and that derivatives of order n are well-defined for integer values $n \geq 1$. It turns out that derivatives of fractional orders (e.g. $n = \frac{1}{2}$) can be defined, with the *Riemann-Louiville* definition being the most common:

For all $0 < \alpha < 1$, the **fractional derivative** of order α of a function $f(x)$ is

$$\frac{d^\alpha}{dx^\alpha} f(x) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dx} \int_0^x \frac{f(t)}{(x-t)^\alpha} \, dt \,. \tag{6.17}$$

Example 6.28

Calculate $\dfrac{d^{1/2}}{dx^{1/2}}(x)$.

Solution: Here $\alpha = \frac{1}{2}$ and $f(x) = x$, so that

$$\frac{d^{1/2}}{dx^{1/2}}(x) = \frac{1}{\Gamma(1-1/2)} \frac{d}{dx} \int_0^x \frac{t}{(x-t)^{1/2}} \, dt = \frac{1}{\sqrt{\pi}} \frac{d}{dx} \int_0^x \frac{t \, dt}{\sqrt{x-t}}$$

since $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ by Example 6.26. Use the substitution $u = \sqrt{x-t}$, so that $t = x - u^2$ and $dt = -2u \, du$. Then $t = 0 \Rightarrow u = \sqrt{x}$ and $t = x \Rightarrow u = 0$, so

$$\begin{aligned} \frac{d^{1/2}}{dx^{1/2}}(x) &= \frac{1}{\sqrt{\pi}} \frac{d}{dx} \int_{\sqrt{x}}^0 \frac{(x-u^2)(-2u \, du)}{u} = \frac{2}{\sqrt{\pi}} \frac{d}{dx} \int_0^{\sqrt{x}} (x - u^2) \, du \\ &= \frac{2}{\sqrt{\pi}} \frac{d}{dx} \left(xu - \tfrac{1}{3} u^3 \right) \Bigg|_{u=0}^{u=\sqrt{x}} = \frac{2}{\sqrt{\pi}} \frac{d}{dx} \left(x^{3/2} - \tfrac{1}{3} x^{3/2} \right) \\ &= \frac{2}{\sqrt{\pi}} \frac{d}{dx} \left(\tfrac{2}{3} x^{3/2} \right) = \frac{2}{\sqrt{\pi}} \sqrt{x} \end{aligned}$$

Part of the motivation for creating fractional derivatives was to find if it were possible to take two "half" derivatives to form a "whole" derivative:

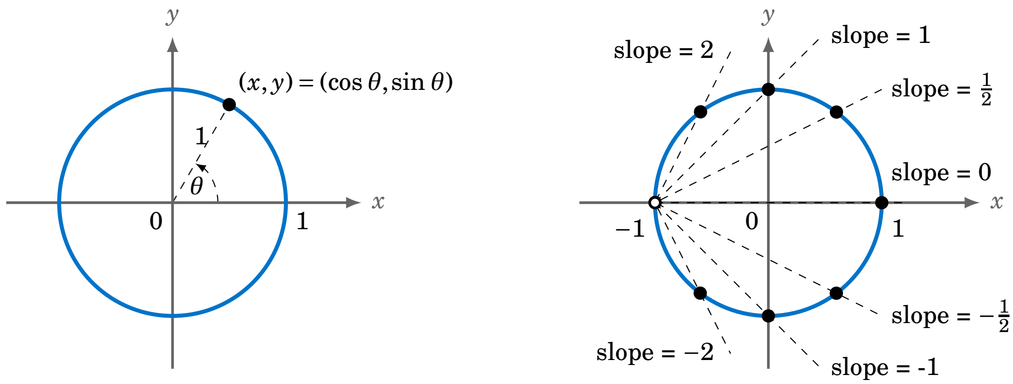
$$\frac{d^{1/2}}{dx^{1/2}} \left(\frac{d^{1/2}}{dx^{1/2}} f(x) \right) = \frac{d}{dx} f(x)$$

It is left as an exercise to show that the above relation does hold for the function $f(x) = x$. Derivatives with fractional order $0 < \alpha < 1$ and integer order $n \geq 1$ can be combined by taking the derivative of integer order first:[8]

$$\frac{d^{n+\alpha}}{dx^{n+\alpha}} f(x) = \frac{d^\alpha}{dx^\alpha} \left(\frac{d^n}{dx^n} f(x) \right)$$

[8]For more details about fractional derivatives, as well as examples of their applications in physics and engineering, see OLDHAM, K.B. AND J. SPANIER, *The Fractional Calculus*, New York: Academic Press, 1974.

Recall from Section 6.3 that the trigonometric substitution $x = r\cos\theta$—or its sister substitution $x = r\sin\theta$—was motivated by trying to find the area of a circle of radius r. To simplify matters, let $r = 1$ so that points on the unit circle can be identified with the angle θ via that substitution, with θ as shown in Figure 6.5.1(a) below.

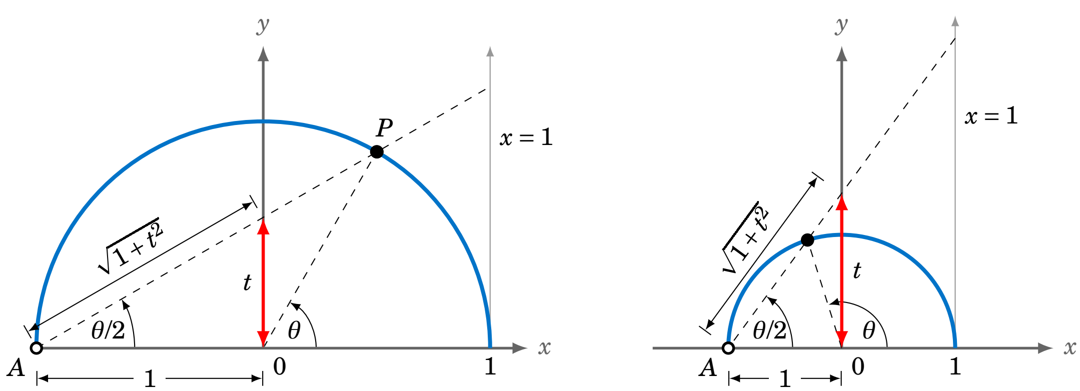


(a) Identify points by angle **(b)** Identify points by slope

Figure 6.5.1 Points on the unit circle $x^2 + y^2 = 1$

Figure 6.5.1(b) shows a different identification of points on the unit circle—by *slope*. This will be the basis for a **half-angle substitution** for evaluating certain integrals.

Let A be the point $(-1, 0)$, then for any other point P on the unit circle draw a line from A through P until it intersects the line $x = 1$, as shown in Figure 6.5.2 below:



(a) $0 \le \theta \le \frac{\pi}{2}$ **(b)** $\frac{\pi}{2} < \theta < \pi$

Figure 6.5.2 Half-angle substitution: $t = \tan\left(\frac{\theta}{2}\right) = \text{slope of } \overline{AP}$

From geometry you know that the inscribed angle that the line $\overline{AP}$ makes with the x-axis is half the measure of the central angle θ. So the slope of $\overline{AP}$ is the tangent of that angle: $\tan\frac{1}{2}\theta = \frac{t}{1} = t$, which is measured along the y-axis and can take any real value. Each point on the unit circle—except A—can be identified with that slope t.

Figure 6.5.2 shows only positive slopes—reflect the picture about the x-axis for negative slopes. The figure shows that

$$\sin \tfrac{1}{2}\theta = \frac{t}{\sqrt{1+t^2}} \qquad \text{and} \qquad \cos \tfrac{1}{2}\theta = \frac{1}{\sqrt{1+t^2}}$$

so that by the double-angle identities for sine and cosine,

$$\sin\theta = 2\sin\tfrac{1}{2}\theta\cos\tfrac{1}{2}\theta = 2\,\frac{t}{\sqrt{1+t^2}}\,\frac{1}{\sqrt{1+t^2}} = \frac{2t}{1+t^2}$$

and

$$\cos\theta = \cos^2\tfrac{1}{2}\theta - \sin^2\tfrac{1}{2}\theta = \frac{1}{1+t^2} - \frac{t^2}{1+t^2} = \frac{1-t^2}{1+t^2}\,.$$

Since $\theta = 2\tan^{-1} t$, then

$$d\theta = d\left(2\tan^{-1} t\right) = \frac{2\,dt}{1+t^2}\,.$$

Below is a summary of the substitution:

Half-angle substitution: The substitution $t = \tan\frac{1}{2}\theta$ yields:

$$\sin\theta = \frac{2t}{1+t^2} \qquad \cos\theta = \frac{1-t^2}{1+t^2} \qquad d\theta = \frac{2\,dt}{1+t^2}$$

The half-angle substitution thus turns rational functions of $\sin\theta$ and $\cos\theta$ into rational functions of t, which can be integrated using partial fractions or another method.

Example 6.29

Evaluate $\displaystyle\int \frac{d\theta}{1+\sin\theta+\cos\theta}$.

Solution: Using $t = \tan\frac{1}{2}\theta$, the denominator of the integrand is

$$1 + \sin\theta + \cos\theta = \frac{1+t^2}{1+t^2} + \frac{2t}{1+t^2} + \frac{1-t^2}{1+t^2} = \frac{2t+2}{1+t^2}$$

so that

$$\begin{aligned}\int \frac{d\theta}{1+\sin\theta+\cos\theta} &= \int \frac{\frac{2\,dt}{1+t^2}}{\frac{2t+2}{1+t^2}} = \int \frac{dt}{t+1} \\ &= \ln|t+1| + C \\ &= \ln\left|\tan\tfrac{1}{2}\theta + 1\right| + C\end{aligned}$$

Example 6.30

Evaluate $\displaystyle\int \frac{d\theta}{3\sin\theta + 4\cos\theta}$.

Solution: Using $t = \tan\frac{1}{2}\theta$, the integral becomes

$$\int \frac{d\theta}{3\sin\theta + 4\cos\theta} = \int \frac{\frac{2\,dt}{1+t^2}}{3\,\frac{2t}{1+t^2} + 4\,\frac{1-t^2}{1+t^2}} = \int \frac{-1}{2t^2-3t-2}\,dt$$

$$= \int \frac{-1}{(2t+1)(t-2)}\,dt = \int \left(\frac{A}{2t+1} + \frac{B}{t-2}\right)dt$$

where

$$\text{coefficient of } t: \quad A + 2B = 0 \quad\Rightarrow\quad A = -2B$$

$$\text{constant term}: \quad -2A + B = -1 \quad\Rightarrow\quad 4B + B = -1 \quad\Rightarrow\quad B = -\frac{1}{5} \text{ and } A = \frac{2}{5}$$

Thus,

$$\int \frac{d\theta}{3\sin\theta + 4\cos\theta} = \int \left(\frac{\frac{2}{5}}{2t+1} + \frac{-\frac{1}{5}}{t-2}\right)dt = \frac{1}{5}\ln|2t+1| - \frac{1}{5}\ln|t-2| + C$$

$$= \frac{1}{5}\ln\left|2\tan\tfrac{1}{2}\theta + 1\right| - \frac{1}{5}\ln\left|\tan\tfrac{1}{2}\theta - 2\right| + C$$

By the half-angle substitution $t = \tan\frac{1}{2}\theta$,

$$\frac{\sin\theta}{1+\cos\theta} = \frac{\dfrac{2t}{1+t^2}}{\dfrac{1+t^2}{1+t^2} + \dfrac{1-t^2}{1+t^2}} = \frac{\dfrac{2t}{1+t^2}}{\dfrac{2}{1+t^2}} = t$$

which yields the useful half-angle identities:[9]

$$\tan\tfrac{1}{2}\theta = \frac{\sin\theta}{1+\cos\theta} = \frac{1-\cos\theta}{\sin\theta} \qquad (6.18)$$

Example 6.31

Evaluate $\displaystyle\int \frac{\sin\theta}{1+\cos\theta}\,d\theta$.

Solution: Though you could use the half-angle substitution $t = \tan\frac{1}{2}\theta$, it is easier to use the half-angle identity (6.18) directly, since

$$\int \frac{\sin\theta}{1+\cos\theta}\,d\theta = \int \tan\tfrac{1}{2}\theta\;d\theta = 2\ln\left|\sec\tfrac{1}{2}\theta\right| + C$$

by formula (6.11) in Section 6.3.

[9] For a different derivation, see pp.79-80 in CORRAL, M., *Trigonometry*, `http://mecmath.net/trig/`, 2009.

Exercises

A

For Exercises 1-12, evaluate the given integral.

1. $\displaystyle\int \frac{1-2\cos\theta}{\sin\theta}\,d\theta$ **2.** $\displaystyle\int \frac{d\theta}{3-5\sin\theta}$ **3.** $\displaystyle\int \frac{d\theta}{2-\sin\theta}$ **4.** $\displaystyle\int \frac{d\theta}{4+\sin\theta}$

5. $\displaystyle\int \frac{\sin\theta}{2-\sin\theta}\,d\theta$ **6.** $\displaystyle\int \frac{d\theta}{5-3\cos\theta}$ **7.** $\displaystyle\int \frac{d\theta}{1+\sin\theta-\cos\theta}$ **8.** $\displaystyle\int \frac{d\theta}{1-\sin\theta+\cos\theta}$

9. $\displaystyle\int \frac{\cot\theta}{1+\sin\theta}\,d\theta$ **10.** $\displaystyle\int \frac{1-\cos\theta}{3\sin\theta}\,d\theta$ **11.** $\displaystyle\int_{-\infty}^{\infty} e^{-x^2/2}\,dx$ **12.** $\displaystyle\int_{-\infty}^{\infty} x^2 e^{-x^6}\,dx$

13. Consider the integral $\displaystyle\int \frac{\sin\theta}{1+\cos\theta}\,d\theta$ from Example 6.31.

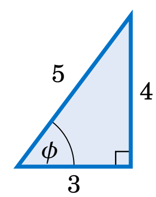
(a) Evaluate the integral using the substitution $u = 1+\cos\theta$.

(b) Evaluate the integral using the half-angle substitution $t = \tan\frac{1}{2}\theta$.

(c) Show that the answers from parts (a) and (b) are equivalent to the result from Example 6.31.

B

14. Evaluate the integral $\displaystyle\int \frac{d\theta}{3\sin\theta+4\cos\theta}$ from Example 6.30 by noting that



$$\begin{aligned}\int \frac{d\theta}{3\sin\theta+4\cos\theta} &= \int \frac{d\theta}{5\left(\frac{3}{5}\sin\theta+\frac{4}{5}\cos\theta\right)}\\ &= \int \frac{d\theta}{5\left(\cos\phi\,\sin\theta+\sin\phi\,\cos\theta\right)}\\ &= \int \frac{d\theta}{5\sin(\theta+\phi)} = \frac{1}{5}\int \csc(\theta+\phi)\,d\theta\end{aligned}$$

by the sine addition formula, where ϕ is the angle in the right triangle shown above. Complete the integration and show that your answer is equivalent to the result from Example 6.30.

15. Show directly from the definition of the Beta function that $B(x,y) = B(y,x)$ for all $x > 0$ and $y > 0$.

16. Show that the Beta function $B(x,y)$ can be written as

$$B(x,y) = \int_0^{\pi/2} 2\sin^{2x-1}(\theta)\,\cos^{2y-1}(\theta)\,d\theta \qquad \text{for all } x>0 \text{ and } y>0.$$

17. Use Exercise 16 and formula (6.16) to show that

$$\int_0^{\pi/2} \sin^m\theta\,\cos^n\theta\,d\theta = \frac{\Gamma\left(\frac{m+1}{2}\right)\Gamma\left(\frac{n+1}{2}\right)}{2\Gamma\left(\frac{m+n}{2}+1\right)} \qquad \text{for all } m>-1 \text{ and } n>-1.$$

18. Use Exercise 28 from Section 6.1, as well as Exercise 17 above, to show that for $m = 1, 2, 3, \ldots$,

$$\int_0^{\pi/2} \sin^{2m}\theta\,d\theta = \frac{\sqrt{\pi}\,\Gamma\left(m+\frac{1}{2}\right)}{2(m!)} \qquad \text{and} \qquad \int_0^{\pi/2} \sin^{2m+1}\theta\,d\theta = \frac{\sqrt{\pi}\,(m!)}{2\Gamma\left(m+\frac{3}{2}\right)}\,.$$

19. Show that $\int_0^\infty \frac{\ln x}{1+x^2}\,dx = 0.$

20. Show that $\int_0^\infty \frac{x^a}{a^x}\,dx = \frac{\Gamma(a+1)}{(\ln a)^{a+1}}$ for $a > 1$.

21. Use the result from Example 6.28 to show that

$$\frac{d^{1/2}}{dx^{1/2}}\left(\frac{d^{1/2}}{dx^{1/2}}(x)\right) = 1 = \frac{d}{dx}(x)\,.$$

22. Calculate $\frac{d^{1/2}}{dx^{1/2}}(c)$ for all constants c.

23. Calculate $\frac{d^{1/3}}{dx^{1/3}}(x)\,.$

24. Show that $\int_0^1 \frac{1}{\sqrt{1-x^n}}\,dx = \frac{1}{n}B\left(\frac{1}{n},\frac{1}{2}\right)$ for $n \ge 1$.

25. Show that the Gamma function $\Gamma(t)$ can be written as

$$\Gamma(t) = p^t \int_0^\infty u^{t-1}e^{-pu}\,du \quad \text{for all } t > 0 \text{ and } p > 0.$$

26. Show that the Gamma function $\Gamma(t)$ can be written as

$$\Gamma(t) = \int_0^1 \left(\ln\left(\frac{1}{u}\right)\right)^{t-1} du \quad \text{for all } t > 0.$$

27. Using the result from Exercise 27 in Section 6.1 that

$$\int e^{ax}\cos bx\,dx = \frac{e^{ax}(a\cos bx + b\sin bx)}{a^2+b^2}$$

for all constants a and $b \neq 0$, differentiate under the integral sign to show that for all $\alpha > 0$

$$\int_0^\infty x e^{-x}\sin\alpha x\,dx = \frac{2\alpha}{(1+\alpha^2)^2}\,.$$

C

28. Use the Leibniz rule and formula (6.8) from Section 6.3 to show that for all $a > 0$,

$$\int \frac{dx}{\sqrt{a^2+x^2}} = \ln\left|x+\sqrt{a^2+x^2}\right| + C\,.$$

29. Use Example 6.27 to show that the Beta function satisfies the relation

$$B(x,1-x) = \int_0^1 \frac{t^{-x} + t^{x-1}}{1+t}\,dt \quad \text{for all } 0 < x < 1.$$

(Hint: First use a substitution to show that $\int_0^\infty \frac{u^{x-1}}{1+u}\,du = \int_0^\infty \frac{t^{-x}}{1+t}\,dt$.*)*

30. Show that for all $a > -1$,

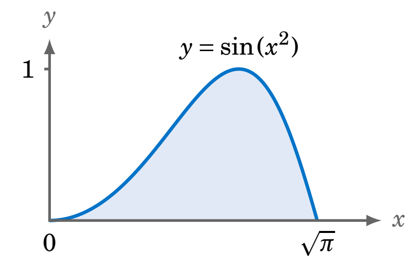
$$\int_0^{\pi/2} \frac{d\theta}{1 + a\sin^2\theta} = \frac{\pi}{2\sqrt{1+a}}\,.$$

6.6 Numerical Integration Methods

Section 5.2 showed how to obtain exact values for definite integrals of some simple functions (low-degree polynomials) by using areas of rectangles. For functions with no closed-form antiderivative, the rectangle method typically produces an approximate value of the definite integral—the more rectangles, the better the approximation.

For example, suppose you wanted to evaluate

$$\int_0^{\sqrt{\pi}} \sin(x^2)\,dx$$



with the rectangle method. This means finding the area of the shaded region in the figure on the right. Computers have eliminated the need to do these sorts of calculations by hand. Though the rectangle method is simple to implement in a traditional programming language (e.g. via a looping construct), there are easier ways in a *domain-specific language* (DSL) geared toward scientific computing. One such DSL is MATLAB®, or its free open-source clone Octave.[10]

Implementing the rectangle method *from scratch* in Octave is a one-liner. For example, suppose you divide the interval $[0,\sqrt{\pi}]$ into 100,000 (10^5) subintervals of equal length, producing 100,001 (1 + 1e5 in scientific notation) equally spaced points in $[0,\sqrt{\pi}]$ (including 0 and $\sqrt{\pi}$). First use the left endpoints of the 10^5 subintervals:

```
octave> sum(sin(linspace(0,sqrt(pi),1+1e5)(1:end-1).^2)*sqrt(pi)/1e5)
ans = 0.8948314693913354
```

Now use the right endpoints:

```
octave> sum(sin(linspace(0,sqrt(pi),1+1e5)(2:end).^2)*sqrt(pi)/1e5)
ans = 0.8948314693913354
```

Finally, use the midpoints of the subintervals:

```
octave> sum(sin((linspace(0,sqrt(pi),1+1e5)(1:end-1)+sqrt(pi)/2e5).^2)*
        sqrt(pi)/1e5)
ans = 0.8948314695305527
```

The true value of the integral up to 15 decimal places is 0.894831469484145, so all three approximations are accurate to 9 decimal places.

[10]Octave is freely available at `https://www.gnu.org/software/octave/`

The syntax in the above commands can be explained with some examples. The following command creates 4 equally spaced points in the interval $[1,7]$ (including $x = 1$ and $x = 7$), thus dividing $[1,7]$ into 3 subintervals each of length $(7-1)/3 = 2$:

```
octave> linspace(1,7,4)
ans =

  1   3   5   7
```

Get all but the last number in the above list:[11]

```
octave> linspace(1,7,4)(1:end-1)
ans =

  1   3   5
```

Now square each number in that list of numbers (the dot before the exponentiation operator ^ applies the squaring operation ^2 element-wise in the list):

```
octave> linspace(1,7,4)(1:end-1).^2
ans =

  1   9   25
```

Now take the sine of each of those squared numbers (measured in radians):

```
octave> sin(linspace(1,7,4)(1:end-1).^2)
ans =

  0.8414709848078965  0.4121184852417566  -0.132351750097773
```

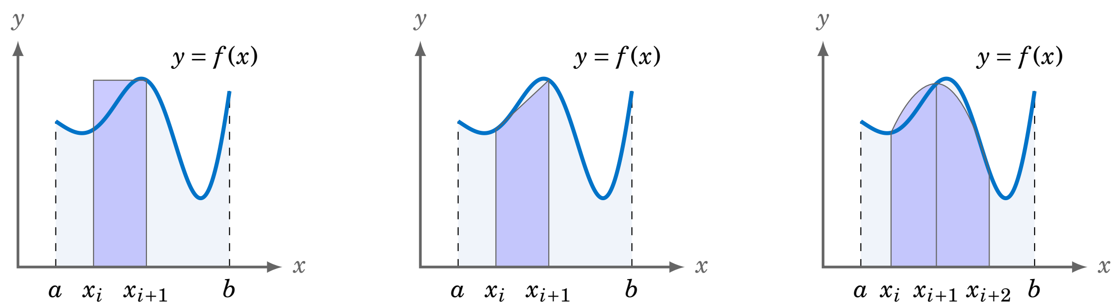
Now multiply each of those numbers (the heights of the rectangles) by the width (2) of each rectangle, then add up those areas:

```
octave> sum(sin(linspace(1,7,4)(1:end-1).^2)*2)
ans = 2.24247543990376
```

[11]MATLAB would require you to use two steps:

```
MATLAB>> x = linspace(1,7,4);
MATLAB>> x(1:end-1)
```

Before the advent of modern computing, the rectangle method was considered inefficient, and so alternative methods were created. Two such methods are the **trapezoid rule** and **Simpson's rule**. The idea behind both methods is to take advantage of a nonlinear function's changing slope by using nonrectangular regions. For the trapezoid rule those regions are trapezoids, while Simpson's rule uses quasi-rectangular regions whose top edges are parabolas, as shown in Figure 6.6.1:



(a) rectangle method **(b)** trapezoid rule **(c)** Simpson's rule

Figure 6.6.1 Comparison of numerical integration methods for $\int_a^b f(x)\,dx$

For a partition $P = \{a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b\}$ of an interval $[a,b]$ into $n \ge 1$ subintervals of equal width $h = (b-a)/n$, let $y_i = f(x_i)$ for $i = 0, 1, \ldots, n$. The trapezoid rule adds up the areas of trapezoids on each subinterval $[x_i, x_{i+1}]$, with the top edge being the line segment joining the points (x_i, y_i) and (x_{i+1}, y_{i+1}). The approximation formula is straightforward to derive, based on areas of trapezoids:

Trapezoid rule:

$$\int_a^b f(x)\,dx \approx \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n)$$

Simpson's rule depends on pairs of neighboring subintervals: $[x_0, x_1]$ and $[x_1, x_2]$, $[x_2, x_3]$ and $[x_3, x_4]$, ... , $[x_{n-2}, x_{n-1}]$ and $[x_{n-1}, x_n]$. Thus, $n \ge 2$ *must be even*. The top edge of the region over each pair $[x_i, x_{i+1}]$ and $[x_{i+1}, x_{i+2}]$ is the unique parabola joining the 3 points (x_i, y_i), (x_{i+1}, y_{i+1}), and (x_{i+2}, y_{i+2}). The approximation formula is then:[12]

Simpson's rule:

$$\int_a^b f(x)\,dx \approx \frac{h}{3}(y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \cdots + \; + 2y_{n-2} + 4y_{n-1} + y_n)$$

[12]For a full derivation of both formulas, see pp.144-149 in HORNBECK, R.W., *Numerical Methods*, New York: Quantum Publishers, Inc., 1975.

Example 6.32

Approximate the value of $\int_0^{\sqrt{\pi}} \sin(x^2)\,dx$ by using the trapezoid rule and Simpson's rule with $n = 10^5$ subintervals.

Solution: Since $x_0 = 0$ and $x_n = \sqrt{\pi}$, then $y_0 = \sin(x_0^2) = \sin 0 = 0$ and $y_n = \sin(x_n^2) = \sin\pi = 0$. Thus, y_0 and y_n contribute nothing to the summation formulas for both rules. In particular the trapezoid rule approximation becomes

$$\int_0^{\sqrt{\pi}} \sin(x^2)\,dx \approx \frac{h}{2}(0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + 0) = h \cdot \left(\sum_{k=1}^{n-1} y_k\right)$$

which is simple to implement in Octave:

```
octave> x = linspace(0,sqrt(pi),1+1e5);
octave> h = sqrt(pi)/1e5;
octave> h*sum(sin(x(2:end-1).^2))
ans = 0.8948314693913405
```

Likewise, the Simpson's rule approximation becomes

$$\int_0^{\sqrt{\pi}} \sin(x^2)\,dx \approx \frac{h}{3}(4y_1 + 2y_2 + 4y_3 + 2y_4 + \cdots + \ + 2y_{n-2} + 4y_{n-1})$$

which can be implemented easily by using Octave's powerful indexing features:

```
octave> (h/3)*(4*sum(sin(x(2:2:end-1).^2)) + 2*sum(sin(x(3:2:end-1).^2)))
ans = 0.8948314694841457
```

In the above command, the statement `x(3:2:end-1)` allows you to skip every other element in the list `x` after position 3, by moving up the list in increments of 2 positions all the way to the next-to-last position in the list (`end-1`). Similarly for `x(2:2:end-1)`, which starts at position 2 and then moves up in increments of 2.

Note that Simpson's rule gives essentially the true value in this case, and the value from the trapezoid rule is virtually the same as the value produced by the built-in `trapz` function in Octave/MATLAB:

```
octave> trapz(x,sin(x.^2))
ans = 0.8948314693913402
```

In general you are better off using these sorts of built-in functions instead of implementing your own.

Typically Simpson's rule is slightly more efficient than the trapezoid rule, which is slightly more efficient than the rectangle method. However, in the above examples all the approximations were accurate to at least 9 decimal places (equivalent to getting the distance between Detroit and Chicago correct within the thickness of a toothpick). The running time of each calculation was only a few thousandths of a second. Modern computing has generally made the efficiency differences negligible.

Notice that the approximations in the rectangle method, the trapezoid rule and Simpson's rule can all be written as linear combinations of function values $f(a_i)$ multiplied by "weights" w_i:

$$\int_a^b f(x)\,dx \;\approx\; \sum_{i=0}^{n} w_i\,f(a_i)$$

For example, the weights in Simpson's rule are $w_i = \frac{h}{3}$, $\frac{2h}{3}$, or $\frac{4h}{3}$, depending on the points a_i in the interval $[a,b]$. The method of **Gaussian quadrature** transforms an integral over any interval $[a,b]$ into an integral over the specific interval $[-1,1]$ and then uses a *standard* set of points in $[-1,1]$ and known weights for those points:[13]

Gaussian quadrature: Transform the integral $\int_a^b f(x)\,dx$ into an integral over $[-1,1]$ by means of the substitution $u = \frac{1}{b-a}(2x-a-b)$, so that $x = \frac{b-a}{2}u + \frac{a+b}{2}$ and $dx = \frac{b-a}{2}du$. Then

$$\int_a^b f(x)\,dx \;=\; \frac{b-a}{2}\int_{-1}^{1} g(u)\,du \;\approx\; \frac{b-a}{2}\sum_{i=1}^{n} w_i\,g(a_i)$$

where $g(u) = f\left(\frac{b-a}{2}u + \frac{a+b}{2}\right)$, with the points $a_1, \ldots, a_n$ and weights $w_1, \ldots, w_n$ given in Table 6.1 below for any choice of $2 \le n \le 10$ points in $[-1,1]$.

Table 6.1 Table of Gaussian quadrature points and weights

n	$a_1, \ldots, a_n$	$w_1, \ldots, w_n$
2	±0.577350	1
3	0	8/9
	±0.774597	5/9
4	±0.339981	0.652145
	±0.861136	0.347855
5	0	0.568889
	±0.538469	0.478629
	±0.906180	0.236927
6	±0.238619	0.467914
	±0.661209	0.360762
	±0.932470	0.171324
7	0	0.417959
	±0.405845	0.381830
	±0.741531	0.279705
	±0.949108	0.129485
8	±0.183435	0.362684
	±0.525532	0.313707
	±0.796666	0.222381
	±0.960290	0.101229
9	0	0.330239
	±0.324253	0.312347
	±0.613371	0.260611
	±0.836031	0.180648
	±0.968160	0.081274
10	±0.148874	0.295524
	±0.433395	0.269267
	±0.679410	0.219086
	±0.865063	0.149451
	±0.973907	0.066671

[13] The details are beyond the scope of this book. See Chapter 4 in RALSTON, A. AND P. RABINOWITZ, *A First Course in Numerical Analysis*, 2nd ed., New York: McGraw-Hill, Inc., 1978. See also Table 1 in STROUD, A.H. AND D. SECREST, *Gaussian Quadrature Formulas*, Englewood Cliffs, NJ: Prentice-Hall, Inc., 1966.

Example 6.33

Approximate the value of $\int_0^2 \frac{dx}{1+x^3}$ by using Gaussian quadrature with $n = 4$ points.

Solution: For $a = 0$ and $b = 2$, use the substitution $u = \frac{1}{b-a}(2x-a-b) = x-1$, so that $x = u+1$ and $dx = du$. Thus, $g(u) = f(u+1) = \frac{1}{1+(u+1)^3}$. Using $n = 4$ in Table 6.1, the points a_i and weights w_i are

$$\begin{aligned} a_1 &= -0.339981 & w_1 &= 0.652145 \\ a_2 &= 0.339981 & w_2 &= 0.652145 \\ a_3 &= -0.861136 & w_3 &= 0.347855 \\ a_4 &= 0.861136 & w_4 &= 0.347855 \end{aligned}$$

and so

$$\begin{aligned} \int_0^2 \frac{dx}{1+x^3} &= \frac{2-0}{2}\int_{-1}^1 g(u)\,du = \int_{-1}^1 \frac{du}{1+(u+1)^3} \approx \sum_{i=1}^4 w_i\,g(a_i) = \sum_{i=1}^4 w_i \cdot \frac{1}{1+(a_i+1)^3} \\ &\approx \frac{0.652145}{1+(-0.339981+1)^3} + \frac{0.652145}{1+(0.339981+1)^3} + \frac{0.347855}{1+(-0.861136+1)^3} + \frac{0.347855}{1+(0.861136+1)^3} \\ &\approx 1.091621 \end{aligned}$$

The true value of the integral to six decimal places is 1.090002.
Using more points (e.g. $n = 7$) is easy to implement in Octave, using element-wise operations on arrays:

```
octave> a = [ 0 -0.405845 0.405845 -0.741531 0.741531 -0.949108 0.949108 ];
octave> w = [ 0.417959 0.381830 0.381830 0.279705 0.279705 0.129485 0.129485 ];
octave> sum(w./(1 + (a+1).^3))
ans = 1.090016688064804
```

Gaussian quadrature can be applied to improper integrals. For example,

$$\int_0^\infty f(x)e^{-x}\,dx \approx \sum_{i=1}^n w_i\,f(a_i)$$

using the points a_i and weights w_i in Table 6.2[14] for $n = 3, 4$, or 5 points in $[0,\infty)$:

Table 6.2 Table of Gaussian quadrature points and weights for $\int_0^\infty f(x)e^{-x}\,dx$

n	$a_1, \ldots, a_n$	$w_1, \ldots, w_n$
3	0.415775	0.711093
	2.294280	0.278518
	6.289945	0.010389

n	$a_1, \ldots, a_n$	$w_1, \ldots, w_n$
4	0.322548	0.603154
	1.745761	0.357419
	4.536620	0.038888
	9.395071	0.000539

n	$a_1, \ldots, a_n$	$w_1, \ldots, w_n$
5	0.263560	0.521756
	1.413403	0.398667
	3.596426	0.075942
	7.085810	0.003612
	12.640801	0.000023

[14] See Table 6 in STROUD, A.H. AND D. SECREST, *Gaussian Quadrature Formulas*.

Example 6.34

Approximate the value of $\int_0^\infty x^5 e^{-x}\,dx$ by using Gaussian quadrature with $n = 3$ points in Table 6.2.

Solution: For $n = 3$, Table 6.2 gives $a_1 = 0.415775$, $a_2 = 2.294280$, $a_3 = 6.289945$, and $w_1 = 0.711093$, $w_2 = 0.278518$, $w_3 = 0.010389$. Then for $f(x) = x^5$,

$$\begin{aligned}\int_0^\infty x^5 e^{-x}\,dx &\approx \sum_{i=1}^n w_i f(a_i) = \sum_{i=1}^n w_i a_i^5 \\ &\approx 0.711093(0.415775)^5 + 0.278518(2.294280)^5 + 0.010389(6.289945)^5 \\ &\approx 119.9974709727211\end{aligned}$$

The true value is $\Gamma(6) = 5! = 120$.
Note: The points a_i in Table 6.2 are the roots of the *Laguerre polynomials* of degree n.

Exercises

A

1. A simple pendulum of length l swings through an angle of 90° on either side of the vertical with period P, given by

$$P = 4\sqrt{\frac{l}{g}}\int_0^{\pi/2}\frac{d\theta}{\sqrt{1 - 0.5\sin^2\theta}}$$

where $g = 9.8$ m/s^2 is the acceleration due to gravity. Use the rectangle method (with left endpoints), the trapezoid rule, and Simpson's rule to write P as a constant multiple of $\sqrt{l/g}$. Preferably, use a computer and $n = 10^5$ subintervals of equal width (or $n = 10$ subintervals if calculating by hand).

2. Repeat Exercise 1 using Gaussian quadrature with $n = 5$ points.

3. Approximate the value of $\int_0^{\sqrt{\pi}} \sin(x^2)\,dx$ by using Gaussian quadrature with $n = 7$ points.

4. Repeat Exercise 3 with $n = 9$ points.

5. Repeat Exercise 3 with $n = 10$ points.

6. The points a_i in Table 6.1 for Gaussian quadrature are the roots of the *Legendre polynomials* $P_n(x)$, with $P_0(x) = 1$, $P_1(x) = x$, and $P_n(x)$ defined for integers $n \ge 2$ by the recursion formula

$$nP_n(x) = (2n-1)xP_{n-1}(x) - (n-1)P_{n-2}(x)$$

(a) Write out $P_n(x)$ explicitly in standard polynomial form for $n = 2,3,4,5$.

(b) Verify that the roots of $P_n(x)$ match the n points $a_1, \ldots, a_n$ in Table 6.1 for $n = 2,3,4,5$.

(c) With no calculations, explain why $\int_{-1}^1 P_2(x)P_3(x)\,dx = \int_{-1}^1 P_3(x)P_4(x)\,dx = \int_{-1}^1 P_4(x)P_5(x)\,dx = 0$.

(d) For $n = 0,1,2,3$, verify that

$$\int_{-1}^1 P_n^2(x)\,dx = \frac{2}{2n+1}.$$

7. Repeat Example 6.34 with $n = 4$ points.

8. Repeat Example 6.34 with $n = 5$ points.

9. Use Table 6.2 to approximate the value of $\int_0^\infty \ln(1+x)e^{-x}\,dx$ with $n = 5$ points.

CHAPTER 7

Analytic Geometry and Plane Curves

7.1 Ellipses

If you were to ask a random person "What is a circle?" a typical response would be to kick the can down the road: "Something that's round." There is a simple definition:

A **circle** is the set of all points in a plane that are a fixed distance from a fixed point in that plane. The fixed point is the **center** of the circle, while the fixed distance from the center is the circle's **radius**.

Similarly, the question "What is an ellipse?" would likely be answered with "an oval," "something egg-shaped," or "a squished circle." A precise definition would be:

An **ellipse** is the set of all points in a plane such that the sum of the distances from each of those points to two fixed points in the plane is the same constant. The two fixed points are the **foci**—plural of **focus**—of the ellipse.

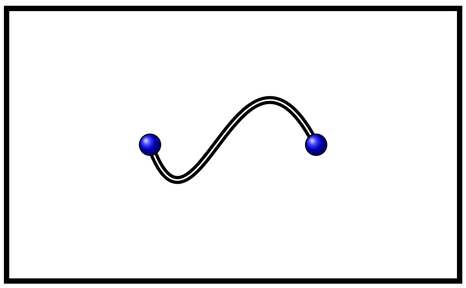
Figure 7.1.1 illustrates the above definitions, with a point P moving along each curve.

(a) Circle: center C, radius r = constant **(b)** Ellipse: foci F_1 and F_2, $d_1 + d_2$ = constant

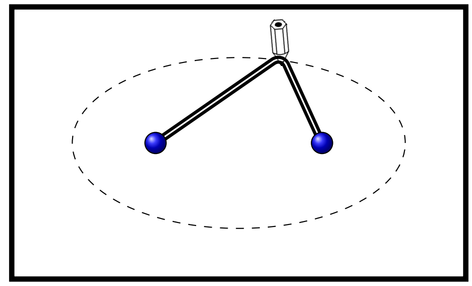
Figure 7.1.1 The circle and its "squished" sibling the ellipse

Along the ellipse the sum $d_1 + d_2$ of the distances from P to the foci remains constant.

The circle's definition makes it easy to imagine its shape, especially for anyone who has drawn a circle with a compass. The definition of the ellipse, on the other hand, might not immediately suggest an "oval" shape. Its shape becomes apparent when physically constructing an ellipse by hand, using only the definition. Stick two pins in a board and tie the ends of a piece of string to the pins, with the string long enough so that there is some slack (see Figure 7.1.2(a)). The pins will be the foci of the ellipse.



(a) Two pins and string on a board



(b) Pull the string taut with a pencil

Figure 7.1.2 Construction of an ellipse

Pull the string taut with a pencil whose point is touching the board, then move the pencil around as far as possible on all sides of the pins. The drawn figure will be an ellipse, as in Figure 7.1.2(b). The length of the string is the constant sum $d_1 + d_2$ of distances from points on the ellipse to the foci. The symmetry of the ellipse is obvious.

There is some terminology connected with ellipses. The **principal axis** is the line containing the foci, and the **center** is halfway between the foci, as in Figure 7.1.3:

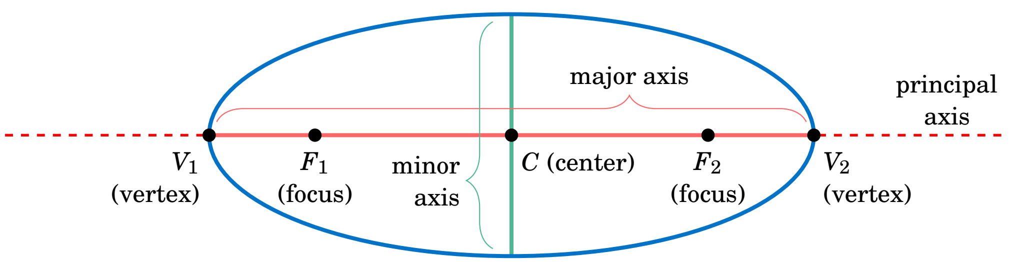
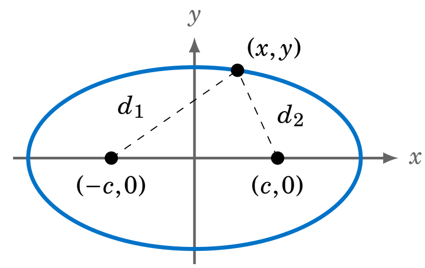


Figure 7.1.3 The parts of an ellipse

The **vertexes** are the points where the ellipse intersects the principal axis. The **major axis** is the chord joining the vertexes, and the **minor axis** is the chord through the center that is perpendicular to the major axis. The two **semi-major axes** are the halves of the major axis joining the center to the vertexes ($\overline{CV_1}$ and $\overline{CV_2}$ in Figure 7.1.3). Likewise the **semi-minor axes** are the two halves of the minor axis. A chord through the center is a **diameter**. Notice that a circle is the special case of an ellipse with identical foci (i.e. the foci and center are all the same point).

Ellipses appear in nature (e.g. the orbits of planets around the Sun) and in many applications. The ancient Greeks were able to derive many properties of the ellipse from its purely geometric definition.[1] Nowadays those properties are typically derived using methods from *analytic geometry*—the study of geometric objects in the context of coordinate systems.[2] You have already seen the equation of an ellipse in the xy-plane centered at the origin: $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, where $a > b > 0$, with the x-axis as the principal axis. The equation is straightforward to derive from the definition of the ellipse.

In the xy-plane, let the foci of an ellipse be the points $(\pm c, 0)$ for some $c > 0$, so that the center is the origin $(0,0)$ and the x-axis is the principal axis, as in the figure on the right. Denote by $2a$ the constant sum $d_1 + d_2$ of the distances from points (x, y) on the ellipse to the foci, with $a > 0$. Notice that $a > c$, since the distance $2c$ between the foci must be less than $d_1 + d_2 = 2a$. Then by the distance formula,



$$
\begin{aligned}
d_1 + d_2 &= 2a \\
\sqrt{(x+c)^2 + y^2} + \sqrt{(x-c)^2 + y^2} &= 2a \\
\left(\sqrt{(x-c)^2 + y^2}\right)^2 &= \left(2a - \sqrt{(x+c)^2 + y^2}\right)^2 \\
(x-c)^2 + \cancel{y^2} &= 4a^2 - 4a\sqrt{(x+c)^2 + y^2} + (x+c)^2 + \cancel{y^2} \\
4a\sqrt{(x+c)^2 + y^2} &= 4a^2 + (x+c)^2 - (x-c)^2 \\
\cancel{4}a\sqrt{(x+c)^2 + y^2} &= \cancel{4}a^2 + \cancel{4}xc \\
\sqrt{(x+c)^2 + y^2} &= a + \tfrac{c}{a}x \\
x^2 + \cancel{2cx} + c^2 + y^2 &= a^2 + \cancel{2cx} + \tfrac{c^2}{a^2}x^2 \\
\left(1 - \tfrac{c^2}{a^2}\right)x^2 + y^2 &= a^2 - c^2 \\
\frac{x^2}{a^2} + \frac{y^2}{a^2 - c^2} &= 1 \\
\frac{x^2}{a^2} + \frac{y^2}{b^2} &= 1 \quad \text{where } b^2 = a^2 - c^2 \text{ (and so } a > b > 0) \quad \checkmark
\end{aligned}
$$

[1] The word *ellipse* is in fact due to the Greek astronomer and geometer Apollonius of Perga (ca. 262-190 B.C.), which seems an improvement over the name “thyreos” that Euclid (ca. 360-300 B.C.) had given the shape.

[2] Pioneered by the French mathematician and philosopher René Descartes (1596-1650), for whom the *Cartesian* coordinate system is named. The proposition “I think, therefore I am” (*Cogito, ergo sum*) is due to Descartes.

The graph of the resulting ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ with $a > b > 0$ and foci at $(\pm c, 0)$ is shown in Figure 7.1.4. Since the the x-axis is principal axis then the vertexes are found by setting $y = 0$: $x = \pm a$. The vertexes are thus $(\pm a, 0)$, so the major axis goes from $(-a, 0)$ to $(a, 0)$ and has length $2a$ (i.e. the semi-major axis has length a). Similarly, setting $x = 0$ shows the minor axis goes from $(0, -b)$ to $(0, b)$, (i.e. the semi-minor axis has length b). Since $a > b > 0$ and $b^2 = a^2 - c^2$, then $c = \sqrt{a^2 - b^2}$. Thus, for *any* ellipse of the form $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ with $a > b > 0$, the foci will be at $(\pm c, 0)$, where $c = \sqrt{a^2 - b^2}$. The foci can be used to define an important geometric property of the ellipse:

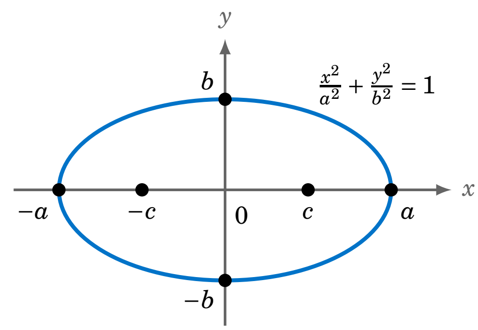
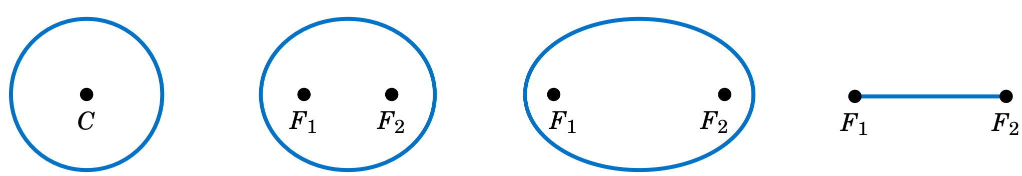


Figure 7.1.4

The **eccentricity** of an ellipse is the ratio of the distance between the foci to the length of the major axis. In the case of an ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ with $a > b > 0$, the eccentricity e is given by

$$e = \frac{c}{a} = \frac{\sqrt{a^2 - b^2}}{a} .$$

The eccentricity e is a measure of how "oval" an ellipse is, with $0 < e < 1$. The boundary case $e = 0$ is a circle, while $e = 1$ is a line segment; an ellipse is somewhere in between—the closer e gets to 1, the more "squished" the ellipse. See Figure 7.1.5.



(a) Circle: $e = 0$ **(b)** Ellipse: $e = 0.5$ **(c)** Ellipse: $e = 0.75$ **(d)** Segment: $e = 1$

Figure 7.1.5 Eccentricity e

Earth's elliptical orbit around the Sun is almost circular: the eccentricity is 0.017. Only the orbits of Venus and Neptune (both at 0.007) have a lower eccentricity among the nine planets in the solar system, while Pluto's (0.252) has the highest.

It is left as an exercise to show that the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ with $a > b > 0$ can be written in terms of the eccentricity e:

$$y^2 = (1 - e^2)(a^2 - x^2) \tag{7.1}$$

Example 7.1

Find the area inside the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.

Solution: By symmetry the area will be four times the area in the first quadrant. Solving for y in the equation of the ellipse gives

$$y^2 = b^2 - \frac{b^2x^2}{a^2} \quad \Rightarrow \quad y = b\sqrt{1 - \frac{x^2}{a^2}} = \frac{b}{a}\sqrt{a^2 - x^2}$$

for the upper hemisphere. Thus,

$$\begin{aligned}
\text{Area} &= 4\int_0^a y\,dx = \frac{4b}{a}\int_0^a \sqrt{a^2 - x^2}\,dx \\
&= \frac{4b}{a}\left(\frac{a^2}{2}\sin^{-1}\left(\frac{x}{a}\right) + \frac{1}{2}x\sqrt{a^2 - x^2}\,\Bigg|_0^a\right) \quad \text{(by formula (6.7) in Section 6.3)} \\
&= \frac{4b}{a}\left(\frac{a^2}{2}\frac{\pi}{2} + 0 - (0 + 0)\right) \\
&= \pi ab
\end{aligned}$$

A remarkable feature of the ellipse is the **reflection property**: *light shone from one focus to any point on the ellipse will reflect to the other focus*. Figure 7.1.6 shows light emanating from the focus F_1 and reflecting off a point P on the ellipse to the other focus F_2. Fermat's Principle from Example 4.4 in Section 4.1 showed that the incoming angle θ_1 (angle of incidence) of light from point A will equal the outgoing angle θ_2 (angle of reflection) to point B for light reflecting off a flat reflective surface at point P, as in Figure 7.1.7(a). Fermat's Principle also applies to curved surfaces—e.g. an ellipse—with the angles measured relative to the tangent line to the curve at the point of reflection, as in Figure 7.1.7(b).

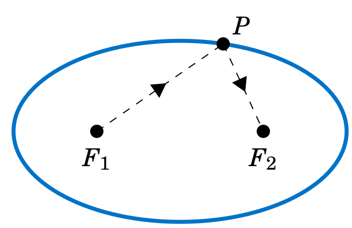


Figure 7.1.6

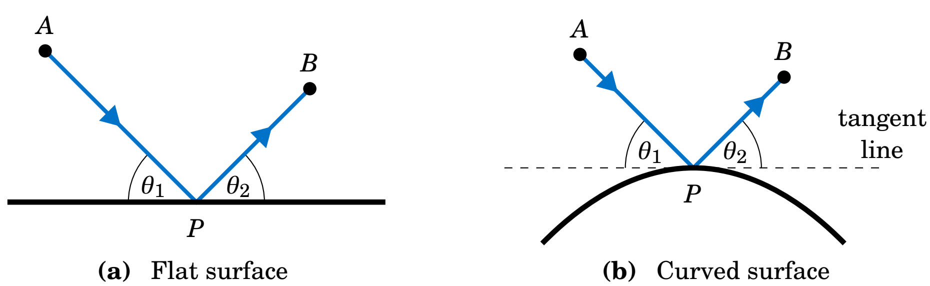
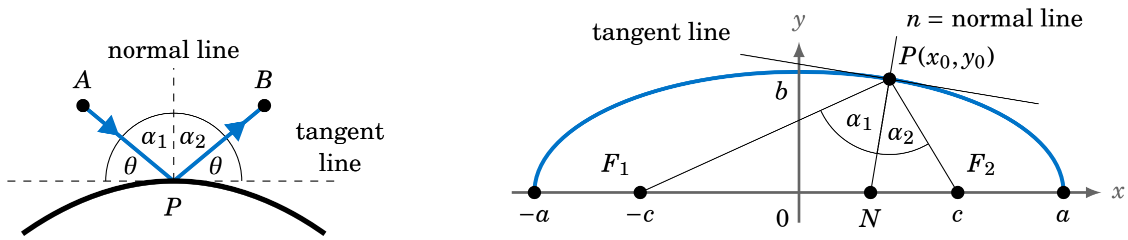


Figure 7.1.7 Fermat's Principle for reflection: $\theta_1 = \theta_2$

Notice that Fermat's Principle is equivalent to saying that the angles α_1 and α_2 that the light's path makes with the normal line through the point of reflection are equal, since each angle would equal $90° - \theta$, as in Figure 7.1.8(a):



(a) $\alpha_1 = \alpha_2$ **(b)** Ellipse: show that $\alpha_1 = \alpha_2$

Figure 7.1.8 Fermat's Principle with normal line

Thus, to prove the reflection property, it suffices to prove that the normal line n to the ellipse at P bisects the angle $\angle F_1PF_2$ in Figure 7.1.8(b)—this would make $\alpha_1 = \alpha_2$, so that the indicated path from F_1 to P to F_2 satisfies Fermat's Principle. First, let $P = (x_0, y_0)$ be a point on the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, with $a > b > 0$. Assume that P is not a vertex (i.e. $y_0 \ne 0$), otherwise the reflection property holds trivially. By Exercise 15 in Section 3.4, the equation of the tangent line to the ellipse at $P = (x_0, y_0)$ is

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} = 1\,, \tag{7.2}$$

so that its slope is $-\frac{b^2x_0}{a^2y_0}$. Hence, the negative reciprocal $\frac{a^2y_0}{b^2x_0}$ is the slope of the normal line n, whose equation—valid even when $x_0 = 0$ (i.e. when $y_0 = \pm b$)— is then

$$b^2x_0(y - y_0) = a^2y_0(x - x_0)\,. \tag{7.3}$$

Setting $y = 0$ and solving for x shows the x-intercept of n is at

$$x = \frac{(a^2 - b^2)x_0y_0}{a^2y_0} = \frac{c^2}{a^2}x_0 = e^2x_0$$

Let $N = (e^2x_0, 0)$ be that x-intercept, as in Figure 7.1.8(b). The distance F_1N from the focus $F_1 = (-c, 0) = (-ea, 0)$ to N is then

$$F_1N = e^2x_0 - (-ea) = e(a + ex_0)$$

while the distance F_2N from the focus $F_2 = (c, 0) = (ea, 0)$ to N is

$$F_2N = ea - e^2x_0 = e(a - ex_0)\,.$$

Therefore,

$$\frac{F_1N}{F_2N} = \frac{e(a + ex_0)}{e(a - ex_0)} = \frac{a + ex_0}{a - ex_0}\,.$$

By the distance formula, the distance F_1P from $F_1 = (-ea, 0)$ to $P = (x_0, y_0)$ is given by

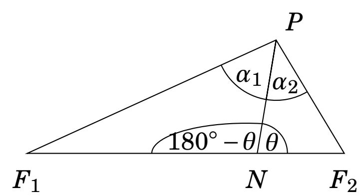
$$\begin{aligned}(F_1P)^2 &= (x_0+ea)^2 + y_0^2 \\ &= x_0^2 + 2eax_0 + e^2a^2 + (1-e^2)(a^2-x_0^2) \quad \text{(by formula (7.1))} \\ (F_1P)^2 &= a^2 + 2eax_0 + e^2x_0^2 = (a+ex_0)^2 \\ F_1P &= a+ex_0 \,.\end{aligned}$$

Similarly, the distance F_2P from $F_2 = (ea, 0)$ to $P = (x_0, y_0)$ is given by

$$\begin{aligned}(F_2P)^2 &= (x_0-ea)^2 + y_0^2 = x_0^2 - 2eax_0 + e^2a^2 + (1-e^2)(a^2-x_0^2) \\ (F_2P)^2 &= a^2 - 2eax_0 + e^2x_0^2 = (a-ex_0)^2 \\ F_2P &= a-ex_0 \,.\end{aligned}$$

Thus,

$$\frac{F_1P}{F_2P} = \frac{a+ex_0}{a-ex_0} = \frac{F_1N}{F_2N} \,,$$



which means that $\alpha_1 = \alpha_2$:[3] by the Law of Sines, and with $\theta = \angle F_2NP$ as in the figure on the right,

$$\frac{\sin\alpha_2}{F_2N} = \frac{\sin\theta}{F_2P} = \frac{\sin(180^\circ-\theta)}{F_2P} = \frac{\sin(180^\circ-\theta)}{F_1P}\cdot\frac{F_1P}{F_2P} = \frac{\sin\alpha_1}{F_1N}\cdot\frac{F_1N}{F_2N} = \frac{\sin\alpha_1}{F_2N}$$

and thus $\sin\alpha_2 = \sin\alpha_1$, so that $\alpha_2 = \alpha_1$ (since $0^\circ < \alpha_1, \alpha_2 < 90^\circ$). ✓

An ellipse of the form

$$\frac{x^2}{b^2} + \frac{y^2}{a^2} = 1$$

with $a > b > 0$ simply switches the roles of x and y in the previous examples: the principal axis is now the y-axis, the foci are at $(0, \pm c)$, where $c = \sqrt{a^2-b^2}$, and the vertexes are at $(0, \pm a)$, as in Figure 7.1.9.

Thus, the largest denominator on the left side of an equation of the form $\frac{x^2}{\square^2} + \frac{y^2}{\square^2} = 1$ tells you which axis is the principal axis. For example, the principal axis of the ellipse $\frac{x^2}{25} + \frac{y^2}{16} = 1$ is the x-axis (since 25 > 16), while the ellipse $\frac{x^2}{4} + \frac{y^2}{9} = 1$ has the y-axis as its principal axis (since 9 > 4).

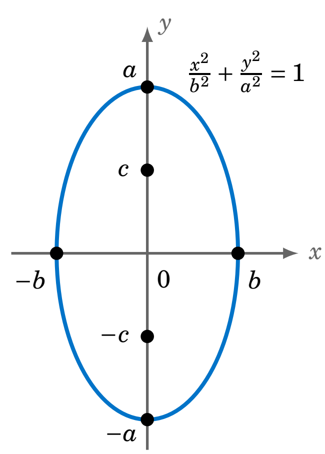


Figure 7.1.9

[3]This follows directly from Proposition 3 in Book VI of Euclid's *Elements*. See the purely geometric proof on pp.125-126 in EUCLID, *Elements*, (Thomas L. Heath translation), Santa Fe, NM: Green Lion Press, 2002.

Exercises

A

1. Construct an ellipse using the procedure shown in Figure 7.1.2. Place the two pins 7in apart and use a 10in piece of string.

For Exercises 2-6, sketch the graph of the given ellipse, indicate the major and minor axes and exact locations of the foci and vertexes, and find the eccentricity e.

2. $\dfrac{x^2}{25} + \dfrac{y^2}{16} = 1$ **3.** $\dfrac{x^2}{4} + \dfrac{y^2}{9} = 1$ **4.** $\dfrac{4x^2}{25} + \dfrac{y^2}{4} = 1$ **5.** $x^2 + 4y^2 = 1$ **6.** $25x^2 + 9y^2 = 225$

7. Show that for $a > b > 0$ the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ with eccentricity e can be written as $y^2 = (1-e^2)(a^2 - x^2)$.

8. Use Example 7.1 to show the area inside the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ with eccentricity e is $\pi a^2 \sqrt{1-e^2}$.

9. For all $a > b > 0$, find the points of intersection of the ellipses $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ and $\frac{x^2}{b^2} + \frac{y^2}{a^2} = 1$.

10. Show that the vertexes are the closest and farthest points on an ellipse to either focus.

B

11. Show that any line of slope m that is tangent to the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ must be of the form

$$y = mx \pm \sqrt{a^2m^2 + b^2}\,.$$

12. A 10ft ladder with a mark 3ft from the top rests against a wall. If the top of the ladder slides down the wall, with the foot of the ladder sliding away from the wall on the ground, as in Figure 7.1.10, then show the mark moves along part of an ellipse.

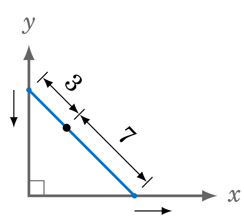


Figure 7.1.10 Exercise 12

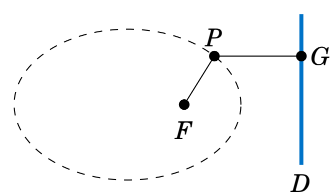


Figure 7.1.11 Exercise 13

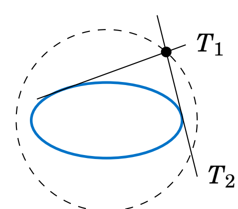


Figure 7.1.12 Exercise 14

13. Another definition of an ellipse is the set of points P for which the ratio of the distance from P to a fixed point F (a focus) to the distance from P to a fixed line D (the *directrix*) is a constant $e < 1$ (the eccentricity): $\frac{PF}{PG} = e$, as in Figure 7.1.11. Use this definition to show that the equation of an ellipse with focus $(c, 0)$ can be written as $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ for some $a > b > 0$. Find the equation of the directrix.

14. Show that the set of intersection points of all perpendicular tangent lines to an ellipse form a circle, as in Figure 7.1.12 (showing two such tangent lines $T_1 \perp T_2$).

15. A chord of an ellipse that passes through a focus and is perpendicular to the major axis is a *latus rectum*. Show that for the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ with $a > b > 0$ the length of each latus rectum is $\frac{2b^2}{a}$.

16. Suppose that the normal line at one end of a latus rectum of an ellipse passes through an end of the minor axis. Show that the eccentricity e is a root of the equation $e^4 + e^2 - 1 = 0$, then find e.

17. Show that the set of all midpoints of a family of parallel chords in an ellipse lie on a diameter. *(Hint: Use symmetry with chords of slope $m \neq 0$.)*

7.2 Parabolas

As with ellipses, you have seen parabolas (e.g. $y = x^2$) and some of their applications (e.g. projectile trajectories), but perhaps without knowing their purely geometric definition. The alternative definition of an ellipse described in Exercise 13 in Section 7.1 is, in fact, similar to the definition of the parabola:

A **parabola** is the set of all points in a plane that are equidistant from a fixed point (the **focus**) and a fixed line (the **directrix**).

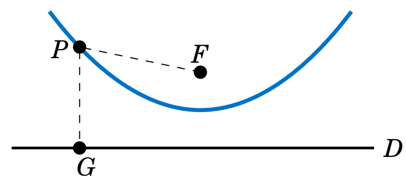


Figure 7.2.1 Parabola: $PF = PG$

Figure 7.2.1 illustrates the above definition, with a point P moving along the parabola so that the distance from P to the focus F equals the distance from P to the directrix D. Note that the point halfway between the focus and directrix must be on the parabola—that point is the **vertex**, which is the point on the parabola closest to the directrix. The **axis** of the parabola is the line that passes through the focus and is perpendicular to the directrix. Notice that the ratio $\frac{PF}{PG}$ equals 1, whereas that ratio for an ellipse—by the alternative definition—was the eccentricity $e < 1$. The **eccentricity** of the parabola, therefore, is always 1.[4]

To construct a parabola from the definition, cut a piece of string to have the same length AB as one side of a drafting triangle, as in Figure 7.2.2.

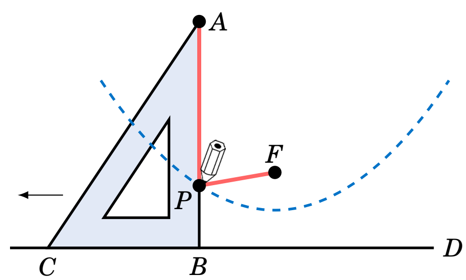
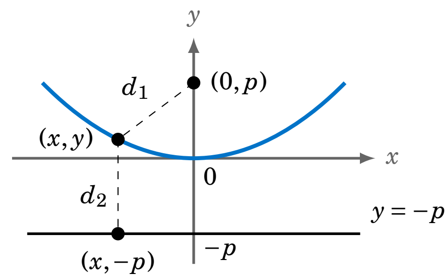


Figure 7.2.2 Construction of a parabola

Fasten one end of the string to the vertex A of the triangle and the other end to a pin somewhere between A and B—the pin will be the focus F of the parabola. Hold the string taut against the edge $\overline{AB}$ of the triangle at a point P on either side of the pin, then move the edge $\overline{BC}$ of the triangle along the directrix D. The drawn figure will be a parabola, as the lengths PF and PB will be equal (since the length of the string being $AB = AP + PF$ means $PF = PB$).

[4]The eccentricity e of the parabola being 1 means there is no second vertex, unlike the ellipse (where $e < 1$ *forced* the existence of two vertexes in the alternative definition).

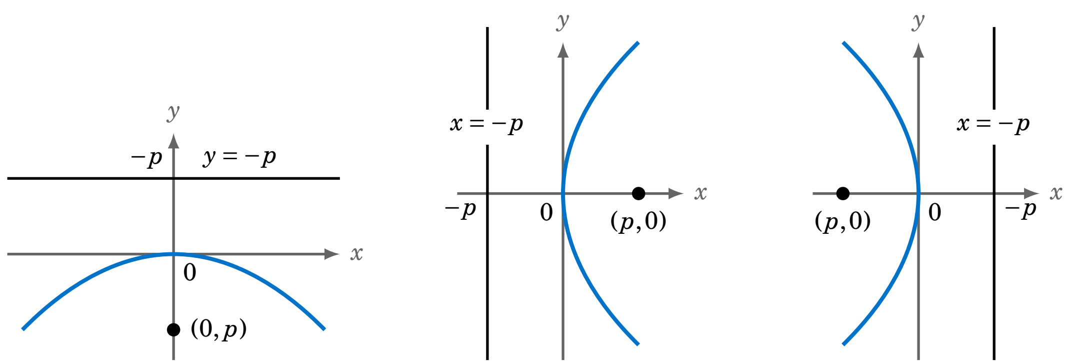
To derive the equation of a parabola in the xy-plane, start with the simple case of the focus on the y-axis at $(0,p)$, with $p>0$, and the line $y=-p$ as the directrix, as in the figure on the right. The vertex is then at the origin $(0,0)$. Pick a point (x,y) whose distances d_1 and d_2 from the focus $(0,p)$ and directrix $y=-p$, respectively, are equal. Then



$$\begin{aligned} d_1^2 &= d_2^2 \\ (x-0)^2 + (y-p)^2 &= (x-x)^2 + (y+p)^2 \\ x^2 + \cancel{y^2} - 2py + \cancel{p^2} &= \cancel{y^2} + 2py + \cancel{p^2} \\ x^2 &= 4py \end{aligned}$$

In other words, $y=\frac{1}{4p}x^2$, which is the more familiar form of a parabola. Thus, any curve of the form $y=ax^2$, with $a \neq 0$, is a parabola whose focus and directrix can be found by dividing 1 by $4a$: $p=\frac{1}{4a}$, so that the focus is at $\left(0,\frac{1}{4a}\right)$ and the directrix is the line $y=-\frac{1}{4a}$. For example, the parabola $y=x^2$ has its focus at $\left(0,\frac{1}{4}\right)$ and its directrix is the line $y=-\frac{1}{4}$.

When $p>0$ the parabola $4py=x^2$ extends upward; for $p<0$ it extends downward, as in Figure 7.2.3(a) below:



(a) $4py=x^2$: $p<0$ **(b)** $4px=y^2$: $p>0$ **(c)** $4px=y^2$: $p<0$

Figure 7.2.3 Parabolas with vertex at the origin

Switching the roles of x and y yields the parabola $4px=y^2$, with focus at $(p,0)$ and directrix $x=-p$. For $p>0$ this parabola extends rightward, while for $p<0$ it extends leftward. See Figure 7.2.3(b) and (c).

It is left as an exercise to show that in general a curve of the form $y=ax^2+bx+c$ is a parabola. Just like not every oval shape is an ellipse, not every "cupped" or "U" shape is a parabola (e.g. $y=x^4$).

The slope of the parabola $4py = x^2$ is $\frac{dy}{dx} = \frac{2x}{4p} = \frac{x}{2p}$, so that the equation of the tangent line to the parabola at a point (x_0, y_0) is:

$$\begin{aligned} y - y_0 &= \frac{x_0}{2p}(x - x_0) \\ 2p(y - y_0) &= x_0 x - x_0^2 \\ 2py - 2py_0 &= x_0 x - 4py_0 \\ 2p(y + y_0) &= x_0 x \end{aligned} \tag{7.4}$$

Likewise, switching the roles of x and y, the tangent line to the parabola $4px = y^2$ at a point (x_0, y_0) is:

$$2p(x + x_0) = y_0 y \tag{7.5}$$

Formula (7.5) simplifies the proof of the **reflection property** for parabolas: *light shone from the focus to any point on the parabola will reflect in a path parallel to the axis of the parabola*. Figure 7.2.4 shows light emanating from the focus $F = (p, 0)$ and reflecting off a point $P = (x_0, y_0)$ on the parabola $4px = y^2$. If that line of reflection is parallel to the x-axis—the axis of the parabola—then the tangent line to the parabola at (x_0, y_0) should make the same angle β with the line of reflection as it does with the x-axis. So extend the tangent line to intersect the x-axis and use formula (7.5) to find the x-intercept:

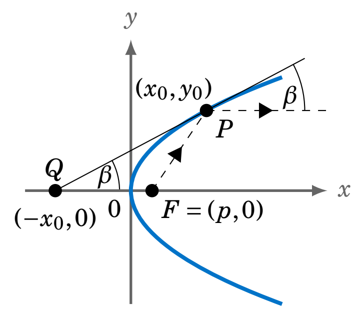


Figure 7.2.4 $4px = y^2$

$$2p(x + x_0) = y_0 y = y_0 \cdot 0 = 0 \quad \Rightarrow \quad x = -x_0$$

Let $Q = (-x_0, 0)$, so that the distance FQ equals $p + x_0$. The goal is to show that the angle of incidence $\angle FPQ$ equals the angle of reflection β. The **focal radius** $\overline{FP}$ has length

$$FP = \sqrt{(p - x_0)^2 + (0 - y_0)^2} = \sqrt{p^2 - 2px_0 + x_0^2 + 4px_0} = \sqrt{p^2 + 2px_0 + x_0^2} = p + x_0 \,.$$

Thus, $FQ = FP$ in the triangle $\triangle FPQ$, so that $\angle FPQ = \angle FQP = \beta$, i.e. the light's path does indeed satisfy Fermat's Principle for curved surfaces. ✓

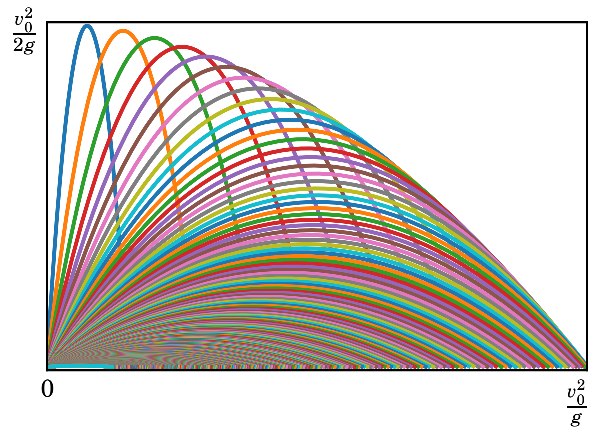
The parabola's reflection property shows up in some engineering applications, typically by revolving part of a parabola around its axis, producing a parabolic surface in three dimensions called a **paraboloid**. For example, it used to be common for vehicle headlights to use paraboloids for their inner reflective surface, with a bulb at the focus, so that—by the reflection property—the light would shine straight ahead in a solid beam. Many flashlights still operate on that principle. The reflection property also works in the opposite direction, which is why satellite dishes and radio telescopes are often wide paraboloids with a signal receiver at the focus, to maximize reception of *incoming* reflected signals.

Example 7.2

Suppose that an object is launched from the ground with an initial velocity v_0 and at varying angles with the ground. Show that the family of all the possible trajectories—which are parabolic—form a region whose boundary (called the **envelope** of the trajectories) is itself a parabola.

Solution: Recall from Example 4.3 in Section 4.1 that if the object is launched at an angle $0 < \theta < \frac{\pi}{2}$ with the ground, then the height y attained by the object as a function of the horizontal distance x that it travels is given by

$$y = -\frac{gx^2}{2v_0^2\cos^2\theta} + x\tan\theta\,.$$



The curve is a parabola, with the figure on the right showing these parabolic trajectories for 500 values of the angle θ. Clearly each parabola intersects at least one other.The maximum horizontal distance $\frac{v_0^2}{g}$ occurs only for $\theta = \frac{\pi}{4}$, as was shown in Example 4.3. The maximum vertical height $\frac{v_0^2}{2g}$ is attained when the object is launched straight up (i.e. $\theta = \frac{\pi}{2}$), as was shown in Exercise 18 in Section 5.1. By symmetry only the angles $0 < \theta \le \frac{\pi}{2}$ in the same vertical plane need be considered. So in the above figure, imagine if the trajectories for all possible angles were included, filling up a region that does appear to have a parabolic boundary. This will now be shown to be true.

First, it turns out that all the parabolas for $0 < \theta < \frac{\pi}{2}$ have the same directrix $y = \frac{v_0^2}{2g}$. To see why, recall from Exercise 11 in Section 4.1 that the maximum height reached by the object is $\frac{v_0^2\sin^2\theta}{2g}$, which is thus the y-coordinate of the vertex of the parabola. That vertex is midway between the focus and the directrix. The parabola is of the form $4py = x^2 + bx$, where b is a constant that does not affect the distance between the vertex and directrix[5], and $4p$ is a constant with $p < 0$ such that the directrix is $-p$ units above the vertex (since $p < 0$), just as in the case $4py = x^2$. The equation of the parabola then shows that

$$\frac{1}{4p} = -\frac{g}{2v_0^2\cos^2\theta} \quad\Rightarrow\quad p = -\frac{v_0^2\cos^2\theta}{2g}$$

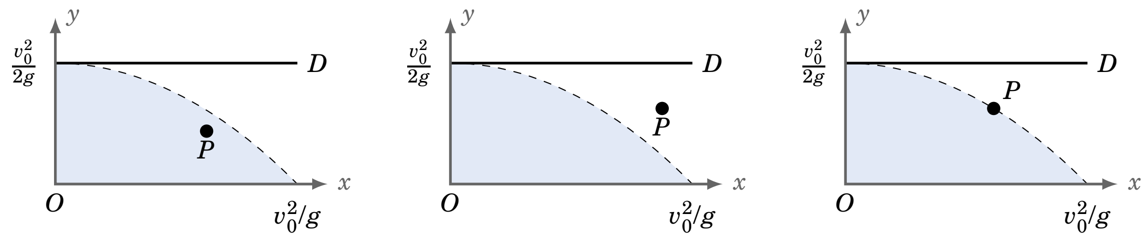
so that the directrix is at

$$\begin{aligned} y &= y\text{-coordinate of the vertex} + (-p) \\ &= \frac{v_0^2\sin^2\theta}{2g} + -\left(-\frac{v_0^2\cos^2\theta}{2g}\right) = \frac{v_0^2}{2g}(\sin^2\theta + \cos^2\theta) \\ y &= \frac{v_0^2}{2g} \end{aligned}$$

It is perhaps surprising that all the parabolic trajectories share the same directrix $y = \frac{v_0^2}{2g}$, which is independent of the angle θ. Note that the heights of each vertex $\left(\frac{v_0^2\sin^2\theta}{2g}\right)$ and focus $\left(\frac{v_0^2}{2g}(\sin^2\theta - \cos^2\theta)\right)$ *do* depend on θ. The common directrix is the key to the remainder of the proof.

[5]This will be discussed further in Section 7.4.

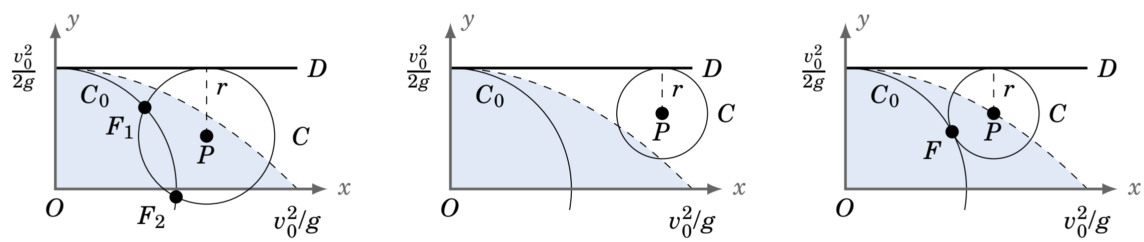
Now let P be a point in the first quadrant of the xy-plane below the common directrix $y = \frac{v_0^2}{2g}$, denoted by D. Then P can be either inside, outside, or on the envelope, as in Figure 7.2.5:



(a) Inside the envelope **(b)** Outside the envelope **(c)** On the envelope

Figure 7.2.5 Trajectory envelope and a point P

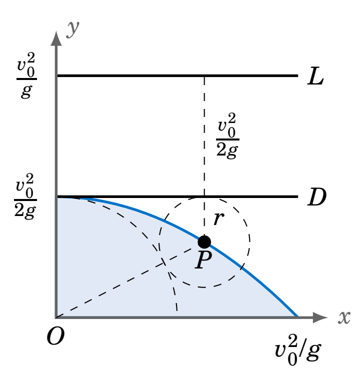
The origin $O = (0,0)$ is on each trajectory, so by definition of a parabola the foci for all the trajectories must be a distance $\frac{v_0^2}{2g}$ from O, i.e. the distance from O to D. That is, the foci of all the trajectories must lie on the circle C_0 of radius $\frac{v_0^2}{2g}$ centered at O. If P is any other point inside the envelope, so that it lies on at least one trajectory, then it must be a distance $r > 0$ below the line D. By definition of a parabola, P must be the same distance from the foci of any trajectories it belongs to. That is, the foci must be on a circle C of radius r centered at P and touching the directrix D, as in Figure 7.2.6:



(a) Two intersections F_1, F_2 **(b)** No intersections **(c)** One intersection F

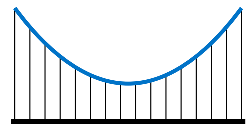
Figure 7.2.6 Circles C and C_0 intersect at foci of trajectories

In Figure 7.2.6(a) C and C_0 intersect at two points F_1 and F_2, so P belongs to two trajectories; P must then be *inside* the envelope. In Figure 7.2.6(b) C and C_0 do not intersect, so P must be *outside* the envelope (since it is not on a parabola with a focus on C_0). If C and C_0 intersect at only one point F, as in Figure 7.2.6(c), then P must be *on* the envelope. In that case, P is a distance $r + \frac{v_0^2}{2g}$ from O, which is also the distance from P to the line $y = \frac{v_0^2}{g}$ (denoted by L). Thus, by definition of a parabola, P is on a parabola with focus O and directrix L. The vertex is at $\left(0, \frac{v_0^2}{2g}\right)$. Therefore, the envelope is a parabola: the boundary of the shaded region in the figure on the right. ✓



In Example 7.2 all the trajectories were in the xy-plane only. Removing that restriction, so that trajectories in all vertical planes through the y-axis are possible, would result in a solid paraboloid consisting of all possible trajectories from the origin.

Parabolas also appear in suspension bridges: the suspension cables supporting a horizontal bridge (via vertical suspenders, as in the figure on the right) have to be parabolas if the weight of the bridge is uniformly distributed.[6]



Exercises

A

1. Construct a parabola using the procedure shown in Figure 7.2.2.

For Exercises 2-6, sketch the graph of the given parabola and indicate the exact locations of the focus, vertex, and directrix.

2. $8y = x^2$ **3.** $y = 8x^2$ **4.** $x = y^2$ **5.** $x = -3y^2$ **6.** $-1000y = x^2$

7. Find the points of intersection of the parabolas $4py = x^2$ and $4px = y^2$ when $p > 0$. What is the equation of the line through those points?

8. A vehicle headlight in the shape of a paraboloid is 3in deep and has an open edge with diameter 8in. Where should the center of the bulb be placed in order to be at the focus, measured in inches relative to the vertex?

9. The *latus rectum* of a parabola is the chord that passes through the focus and is parallel to the directrix. Find the length of the latus rectum for the parabola $4py = x^2$.

10. Show that the circle whose diameter is the latus rectum of a parabola touches the parabola's directrix at one point.

11. Find the points on the parabola $4px = y^2$ such that the focal radii to those points have the same length as the latus rectum.

12. From each end of the latus rectum of a parabola draw a line to the point where the directrix and axis intersect. Show that the two drawn lines are perpendicular.

B

13. Show that any point not on a parabola is on either zero or two tangent lines to the parabola.

14. Show that $y = mx - 2mp - m^3p$ is the normal line of slope m to the parabola $4px = y^2$.

15. From a point P on a parabola with vertex V let $\overline{PQ}$ be the line segment perpendicular to the axis at a point Q. Show that PQ^2 equals the product of QV and the length of the latus rectum.

16. Show that the curve $y = ax^2 + bx + c$ is a parabola for $a \neq 0$, using only the definition of a parabola. Find the focus, vertex and directrix.

17. Show that the set of all midpoints of a family of parallel chords in a parabola lie on a line parallel to the parabola's axis.

[6]See pp.159-161 in SMITH, C.E., *Applied Mechanics: Statics*, New York: John Wiley & Sons, Inc., 1976.

7.3 Hyperbolas

In the previous two sections you have seen curves with eccentricity $e = 0$ (circles), $0 < e < 1$ (ellipses) and $e = 1$ (parabolas). The remaining case is $e > 1$: the **hyperbola**, whose definition is similar to the second definition of the ellipse.

A **hyperbola** is the set of all points in a plane such that the ratio of the distance from a fixed point (a **focus**) to the distance from a fixed line (a **directrix**) is a constant $e > 1$, called the **eccentricity** of the hyperbola.

It will be shown in Section 7.4 that the curve $y = \frac{1}{x}$ is a hyperbola, which has two branches (see Figure 7.3.1). In general a hyperbola resembles a "wider" or less "cupped" parabola, and it has two symmetric branches (and hence two foci and two directrices) as well as two asymptotes.

The ratio of distances referred to in the definition of the hyperbola also appears in the second definition of the ellipse (where the ratio is smaller than 1) and in the definition of the parabola (where the ratio equals 1). In all three cases that ratio is the eccentricity. See Figure 7.3.2(a) for the comparisons.

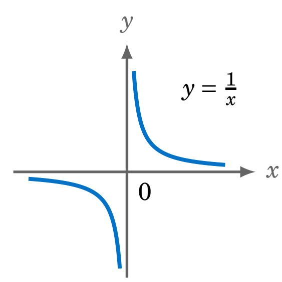
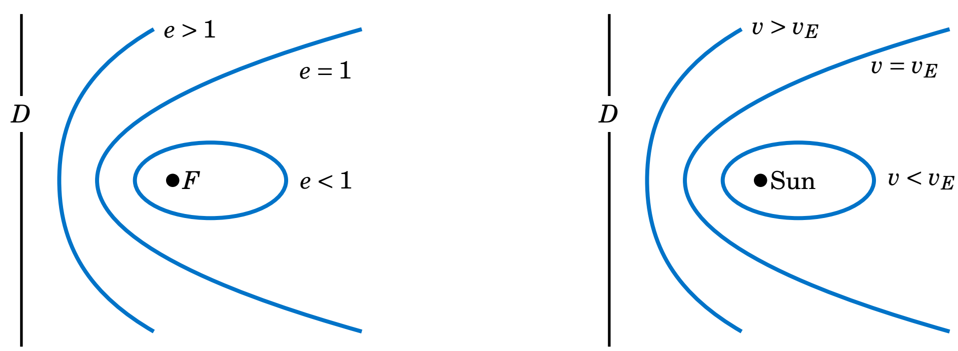


Figure 7.3.1



(a) Eccentricity e **(b)** Orbit velocity v and escape velocity v_E

Figure 7.3.2 Hyperbola, parabola and ellipse with focus F and directrix D

There is an analogue for Figure 7.3.2(a) in terms of orbits. For example, an object approaching the Sun must meet or exceed the **escape velocity** to overcome the Sun's gravitational pull and avoid returning to orbit. Figure 7.3.2(b) shows the three possible trajectories—hyperbola, parabola and ellipse—in terms of the object's velocity v and the escape velocity v_E. Note the apparent correlation between the eccentricity of the object's path and its speed as a fraction of the escape velocity (i.e. $\frac{v}{v_E}$).

Figure 7.3.3 illustrates the definition of a hyperbola, consisting of points P whose distance PF from the focus F exceeds the distance PG to the directrix D in a way so that the ratio $\frac{PF}{PG}$ is always the same constant $e > 1$ (the eccentricity).

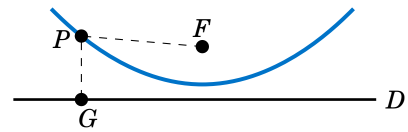


Figure 7.3.3 Hyperbola: $\frac{PF}{PG} = e > 1$

To derive the equation of a hyperbola with eccentricity $e > 1$, assume the focus is on the x-axis at $(ea, 0)$, with $a > 0$, and the line $x = \frac{a}{e}$ is the directrix, as in Figure 7.3.4. Pick a point (x, y) whose distances d_1 and d_2 from the focus $(ea, 0)$ and directrix $x = \frac{a}{e}$, respectively, satisfy the condition for a hyperbola: $\frac{d_1}{d_2} = e > 1$. Then

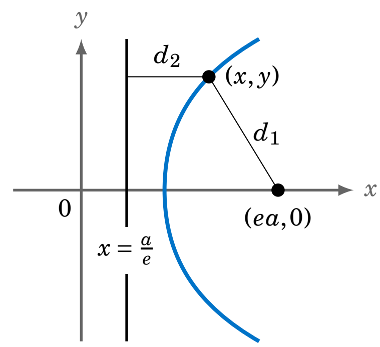


Figure 7.3.4

$$
\begin{aligned}
d_1^2 &= e^2 d_2^2 \\
(x - ea)^2 + (y - 0)^2 &= e^2 \left(\left(x - \frac{a}{e} \right)^2 + (y - y)^2 \right) \\
x^2 - \cancel{2eax} + e^2 a^2 + y^2 &= e^2 x^2 - \cancel{2eax} + a^2 \\
(e^2 - 1)x^2 - y^2 &= (e^2 - 1)a^2 \\
\frac{x^2}{a^2} - \frac{y^2}{b^2} &= 1
\end{aligned}
$$

where $b^2 = (e^2 - 1)a^2 > 0$. The hyperbola thus has two branches ($x = \pm \frac{a}{b} \sqrt{y^2 + b^2}$), as in Figure 7.3.5. Let $c = ea$, so that $c > a$ and $b^2 = c^2 - a^2$. By symmetry the hyperbola has two foci $(\pm c, 0)$ and two directrices $x = \pm \frac{a^2}{c}$, and the lines $y = \pm \frac{b}{a} x$ are asymptotes. The **vertexes** are the points on the hyperbola closest to the directrices.

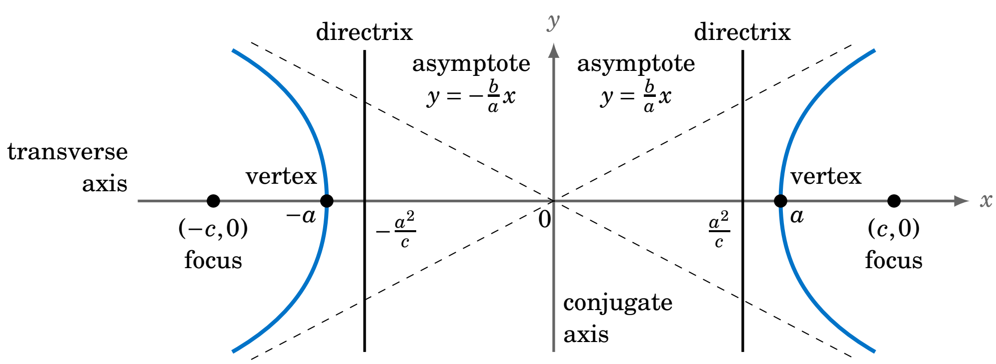


Figure 7.3.5 Parts of the hyperbola $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$, with $b^2 = c^2 - a^2$

The **center** is the point midway between the foci. The **transverse axis** and **conjugate axis** are the perpendicular lines through the foci and center, respectively.

In Figure 7.3.5 the vertexes are $(\pm a, 0)$, the x-axis is the transverse axis, the center is the origin $(0,0)$, and the conjugate axis is the y-axis. Note that the existence of *two* foci and directrices—when the definition of the hyperbola mentioned only *a* focus and *a* directrix—is simply a consequence of the symmetry about *both* axes imposed by the equation $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$. A parabola, in comparison, is symmetric about only *one* axis.

To see why the lines $y = \pm\frac{b}{a}x$ are asymptotes, consider the upper half $y = \frac{b}{a}\sqrt{x^2 - a^2}$ of both branches of the hyperbola. The difference between the line $y = \frac{b}{a}x$ and the upper right branch approaches zero as x approaches infinity:

$$\begin{aligned}\lim_{x\to\infty}\left(\frac{b}{a}x - \frac{b}{a}\sqrt{x^2-a^2}\right) &= \frac{b}{a}\lim_{x\to\infty}\left(x - \sqrt{x^2-a^2}\right)\cdot\frac{x + \sqrt{x^2-a^2}}{x + \sqrt{x^2-a^2}}\\ &= \frac{b}{a}\lim_{x\to\infty}\frac{x^2 - (x^2-a^2)}{x + \sqrt{x^2-a^2}}\\ &= \lim_{x\to\infty}\frac{ab}{x + \sqrt{x^2-a^2}} = 0\end{aligned}$$

Thus the line $y = \frac{b}{a}x$ is an (oblique) asymptote for the upper half $y = \frac{b}{a}\sqrt{x^2-a^2}$ of the right branch in the first quadrant of the xy-plane. So by symmetry the lines $y = \pm\frac{b}{a}x$ are asymptotes for both branches, i.e. for the entire hyperbola.

Conversely, given a hyperbola in the form $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$, let $c^2 = a^2 + b^2$ to get the foci $(\pm c, 0)$ and the directrices $x = \pm\frac{a^2}{c}$, while the eccentricity is $e = \frac{c}{a}$.

Example 7.3

For the hyperbola $x^2 - y^2 = 1$ find the vertexes, foci, directrices, asymptotes and eccentricity.

Solution: Here $a = b = 1$, so that $c^2 = a^2 + b^2 = 2$, i.e. $c = \sqrt{2}$. The vertexes are thus $(\pm 1, 0)$, the asymptotes are $y = \pm x$, the foci are $(\pm\sqrt{2}, 0)$, the directrices are $x = \pm\frac{1}{\sqrt{2}}$, and the eccentricity is $e = \sqrt{2}$.

Switching the roles of x and y produces the rotated hyperbola $\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1$, shown in Figure 7.3.6. The transverse axis is the y-axis, the conjugate axis is the x-axis, the vertexes are at $(0, \pm a)$, and the foci are at $(0, \pm c)$, where $c^2 = a^2 + b^2$. The directrices are $y = \pm\frac{a^2}{c}$, the asymptotes are $y = \pm\frac{a}{b}x$, and the eccentricity is $e = \frac{c}{a}$.

For example, the hyperbola $y^2 - x^2 = 1$ has $a = b = 1$, so that $c = \sqrt{2}$. The vertexes are $(0, \pm 1)$, the asymptotes are $y = \pm x$, the foci are $(0, \pm\sqrt{2})$, the directrices are $y = \pm\frac{1}{\sqrt{2}}$, and the eccentricity is $e = \sqrt{2}$. The hyperbola $y^2 - x^2 = 1$ is just the hyperbola $x^2 - y^2 = 1$ rotated 90°.

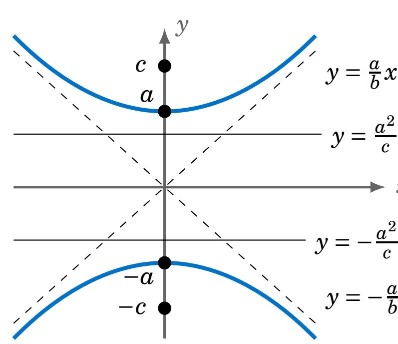


Figure 7.3.6 $\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1$

There is another way to define a hyperbola, in terms of two foci:

A **hyperbola** is the set of all points in a plane such that the absolute value of the difference of the distances from two fixed points (the **foci**) is a positive constant.

Figure 7.3.7 illustrates the above definition with foci F_1 and F_2. The difference $d_1 - d_2$ of the distances d_1 and d_2 can be positive or negative depending on which branch the point (x, y) is on, which is why the absolute value $|d_1 - d_2|$ is used.

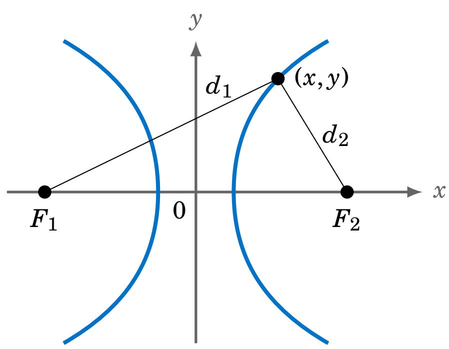


Figure 7.3.7 Hyperbola: $|d_1 - d_2| = \text{constant} > 0$

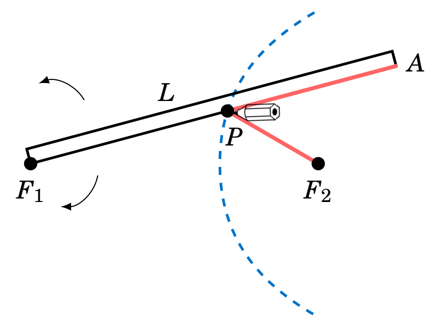


Figure 7.3.8 Construction of a hyperbola

It is left as an exercise to show that this second definition yields the same equation of the hyperbola as from the first definition. The second definition is often used as the primary definition in many textbooks, perhaps because it provides a simple way to construct a hyperbola by hand. Figure 7.3.8 shows the procedure for foci F_1 and F_2: at the focus F_1 fasten one end of a ruler of length L, and at the other end A of the ruler fasten one end of a string of length $L - d$ for some number $0 < d < F_1F_2$. Fasten the other end of the string to the focus F_2 and hold the string taut with a pencil against the ruler at a point P, while rotating the ruler about F_1. The drawn figure will be one branch of a hyperbola, since the difference $PF_1 - PF_2$ will always be the positive constant d:

$$PF_1 - PF_2 = (L - AP) - PF_2 = L - (AP + PF_2) = L - (L - d) = d$$

Reverse the roles of F_1 and F_2 to draw the other branch of the hyperbola.

By Exercise 16 in Section 3.4, the tangent line to the hyperbola $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ at a point (x_0, y_0) is

$$\frac{xx_0}{a^2} - \frac{yy_0}{b^2} = 1\,, \tag{7.6}$$

so that its slope is $\frac{b^2 x_0}{a^2 y_0}$ when $y_0 \neq 0$. Note that by the above equation, when $y_0 = 0$ (so that $x_0 = \pm a$) the two tangent lines are the vertical lines $x = \pm a$.

That slope will be used in proving the **reflection property** for the hyperbola: *Light shone from one focus will reflect off the hyperbola in the opposite direction from the other focus.* Figure 7.3.9 shows the light's path from focus F_2 as it reflects at the point P along the line through P and the other focus F_1.

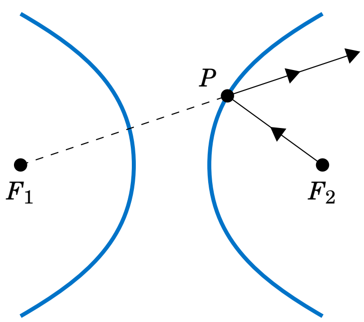


Figure 7.3.9 Reflection property

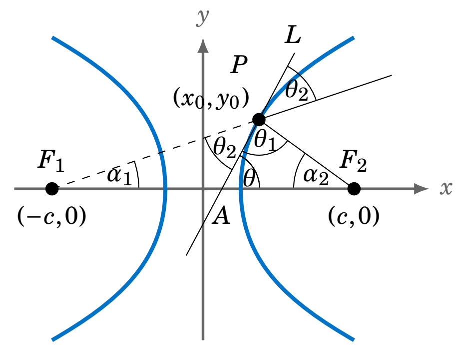


Figure 7.3.10 Hyperbola $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ with foci $(\pm c, 0)$

By Fermat's Principle for curved surfaces, the reflection property is equivalent to saying the tangent line L to the hyperbola at the point $P = (x_0, y_0)$ bisects the angle $\angle F_1PF_2$, i.e. $\theta_1 = \theta_2$ as in Figure 7.3.10. The reflection property holds trivially when $y_0 = 0$ (the light reflects straight back along the x-axis), so to show that $\theta_1 = \theta_2$ assume $y_0 \neq 0$. By symmetry only $x_0 > 0$ need be considered. For the case $x_0 \neq c$, let A be the x-intercept of the tangent line L. Then by Figure 7.3.10, since the sum of the angles in the triangle $\triangle F_1PA$ equals 180°,

$$\alpha_1 + \theta_2 + (180^\circ - \theta) = 180^\circ \quad \Rightarrow \quad \theta_2 = \theta - \alpha_1 \quad \Rightarrow \quad \tan\theta_2 \;=\; \tan(\theta - \alpha_1)\,.$$

Thus, since $\tan\theta$ is the slope of the tangent line L (i.e. $\frac{b^2x_0}{a^2y_0}$), and $\tan\alpha_1 = \frac{y_0}{x_0+c}$, then by the subtraction formula for the tangent function

$$\tan\theta_2 \;=\; \frac{\tan\theta \,-\, \tan\alpha_1}{1 \,+\, \tan\theta\,\tan\alpha_1} \;=\; \frac{\frac{b^2x_0}{a^2y_0} - \frac{y_0}{x_0+c}}{1 + \frac{b^2x_0}{a^2y_0}\cdot\frac{y_0}{x_0+c}} \;=\; \frac{\frac{b^2x_0^2 - a^2y_0^2 + b^2x_0c}{a^2y_0(x_0+c)}}{\frac{(a^2+b^2)x_0y_0 + a^2y_0c}{a^2y_0(x_0+c)}}$$

$$= \frac{a^2b^2 + b^2x_0c}{c^2x_0y_0 + a^2y_0c} \quad \left(\text{since } c^2 = a^2 + b^2, \text{ and } \frac{x_0^2}{a^2} - \frac{y_0^2}{b^2} = 1 \;\Rightarrow\; b^2x_0^2 - a^2y_0^2 = a^2b^2\right)$$

$$= \frac{b^2(a^2 + x_0c)}{cy_0(a^2 + x_0c)} = \frac{b^2}{cy_0}$$

Similarly, since the sum of the angles in the triangle $\triangle F_2PA$ equals 180°,

$$\alpha_2 + \theta_1 + \theta = 180^\circ \quad \Rightarrow \quad \theta_1 = 180^\circ - (\theta + \alpha_2) \quad \Rightarrow \quad \tan\theta_1 \;=\; -\tan(\theta + \alpha_2)\,.$$

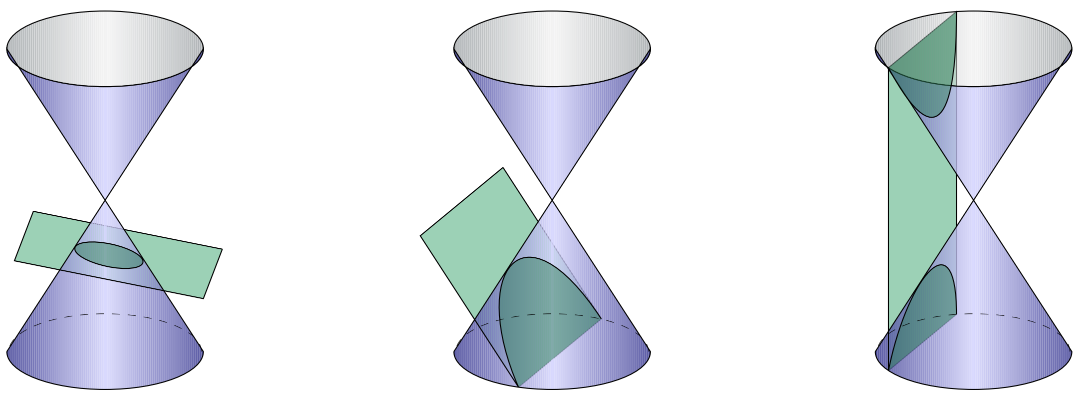
Thus, since $\tan\alpha_2 = -\tan(180° - \alpha_2) = -\frac{y_0}{x-c}$,

$$\tan\theta_1 = -\frac{\tan\theta + \tan\alpha_2}{1 - \tan\theta\,\tan\alpha_2} = -\frac{\frac{b^2x_0}{a^2y_0} + \frac{-y_0}{x_0 - c}}{1 - \frac{b^2x_0}{a^2y_0}\cdot\frac{-y_0}{x_0 - c}} = -\frac{\frac{b^2x_0^2 - a^2y_0^2 - b^2x_0c}{a^2y_0(x_0 - c)}}{\frac{(a^2 + b^2)x_0y_0 - a^2y_0c}{a^2y_0(x_0 - c)}}$$

$$= -\frac{a^2b^2 - b^2x_0c}{c^2x_0y_0 - a^2y_0c} = \frac{b^2(a^2 - x_0c)}{cy_0(a^2 - x_0c)} = \frac{b^2}{cy_0} = \tan\theta_2\ ,$$

i.e. $\theta_1 = \theta_2$ (since $0° < \theta_1, \theta_2 < 90°$). ✓ Note: The case $x_0 = c$ is left as an exercise.

Ellipses, parabolas and hyperbolas are sometimes called **conic sections**, due to being formed by intersections of planes with a double circular cone of unlimited extent:



(a) Ellipse **(b)** Parabola **(c)** Hyperbola

Figure 7.3.11 Conic sections

Each double cone in Figure 7.3.11 has two **nappes**—a cone extending upward and one extending downward. When a plane intersects only one nappe in a closed noncircular curve, as in Figure 7.3.11(a), that curve is an ellipse. A plane that is parallel to a line on one nappe, as in Figure 7.3.11(b), intersects only that nappe in a parabola. The intersection of a plane with both nappes, as in Figure 7.3.11(c), is a hyperbola.

To prove that the ellipse, parabola and hyperbola really are represented by the indicated conic sections, first a minor result is needed from three-dimensional geometry: tangent line segments to a sphere from the same point have equal lengths, as in Figure 7.3.12. Since the right triangles $\triangle QOP_1$ and $\triangle QOP_2$ share the same hypotenuse $\overline{QO}$ and have legs $\overline{OP_1}$ and $\overline{OP_2}$ of equal length (the radius of the sphere), the result $QP_1 = QP_2$ follows by the Pythagorean Theorem.

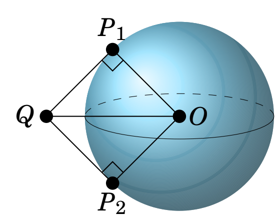


Figure 7.3.12

For the case of a right circular double cone (i.e. the base of each nappe is a circle in a plane perpendicular to the axis of the cone[7]), let β be the complement of the angle that the cone makes with its axis, as in Figure 7.3.13. So β is the angle the cone makes with any circular base of the cone. This constant angle β, with $0° < \beta < 90°$, is an intrinsic property of the cone. Let P_c be a plane that intersects the lower nappe of the cone in a curve C, such that P_c makes an angle α with any base circle of the cone. By symmetry, only the angles $0° < \alpha \le 90°$ need be considered.[8]

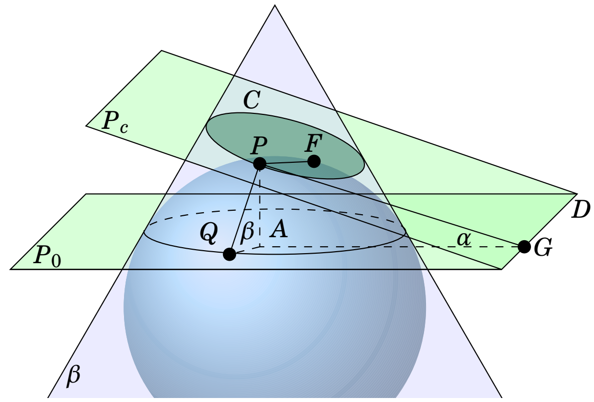


Figure 7.3.13

Inscribe a sphere in the cone so that it touches P_c at a point F, as in Figure 7.3.13 (P_c is the *tangent plane* to the sphere at F). Let P_0 be the plane through the circle where the inscribed sphere touches the cone, and let D be the line of intersection of the planes P_c and P_0. It will be shown that F and D are the focus and directrix, respectively, of the curve C.

Let P be any point on the curve C, then let Q be the point on the plane P_0 that lies on a line through P and the cone's vertex. Drop a perpendicular line segment from P to the point A in the plane P_0. From A draw a perpendicular line segment to the point G on the line D. Then as Figure 7.3.13 shows, $\triangle QAP$ and $\triangle PAG$ are right triangles, with

$$\sin\alpha = \frac{PA}{PG} \quad \text{and} \quad \sin\beta = \frac{PA}{PQ}\,.$$

However, since $\overline{PF}$ and $\overline{PQ}$ are both tangent line segments to the inscribed sphere from the same point P, the result proved earlier shows that $PQ = PF$. Thus,

$$PG\sin\alpha = PA = PQ\sin\beta = PF\sin\beta \quad \Rightarrow \quad \frac{PF}{PG} = \frac{\sin\alpha}{\sin\beta}\,.$$

Let $e = \frac{PF}{PG}$. Then e is the same constant $\frac{\sin\alpha}{\sin\beta}$ for any point P on the curve C. Thus, by definition, e is the eccentricity of the curve C with focus F and directrix D. If $0° < \alpha < \beta$ then $0° < \sin\alpha < \sin\beta$ so that $0 < \frac{\sin\alpha}{\sin\beta} < 1$, which means that $0 < e < 1$, and hence C is an ellipse (by the second definition of an ellipse). Likewise, if $\alpha = \beta$ then $e = 1$, so that C is a parabola. Finally, if $\beta < \alpha \le 90°$ then $e > 1$, so that C is a hyperbola (and will intersect both nappes of the cone). Thus, the ellipse, parabola and hyperbola truly are conic sections. ✓

[7]The proof can be extended to oblique double cones. See §364 in SALMON, G.S., *A Treatise on Conic Sections*, London: Longmans, Green and Co., 1929.

[8]The case where $\alpha = 0°$ results in a circle, which is typically not considered a conic section.

Exercises

A

1. Construct a hyperbola using the procedure shown in Figure 7.3.8. Place the two focus pins 7in apart and use a 9in piece of string attached to a 12in ruler.

For Exercises 2-6, sketch the graph of the given hyperbola, indicate the exact locations of the foci and vertexes, indicate each directrix and asymptote, and find the eccentricity e.

2. $\frac{x^2}{16} - \frac{y^2}{9} = 1$ **3.** $\frac{x^2}{8} - \frac{y^2}{15} = 1$ **4.** $\frac{4x^2}{25} - \frac{y^2}{4} = 1$ **5.** $x^2 - 4y^2 = 1$ **6.** $25y^2 - 9x^2 = 225$

7. Find the equation of the hyperbola with foci $(\pm 5,0)$ and vertexes $(\pm 3,0)$.

B

8. In the second definition of the hyperbola on p.219, let $2a$ be the constant absolute value of the difference of distances from points on the hyperbola to the foci, for some $a > 0$. Let the foci be at $(\pm c,0)$ for some $c > 0$. Show that this second definition yields a hyperbola having an equation of the form $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$, with $b > 0$.

9. Prove the reflection property for the hyperbola (see pp.220-221) when $x_0 = c$ for the point $P = (x_0, y_0)$ on the hyperbola with foci at $(\pm c,0)$.

10. Show that a straight line parallel to an asymptote of a hyperbola intersects the hyperbola at exactly one point.

11. A *latus rectum* (plural: latus recta) of a hyperbola is a chord through either focus perpendicular to the transverse axis. Show that the latus recta of the hyperbola $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ have length $\frac{2b^2}{a}$.

12. Show that the segment of an asymptote of a hyperbola between the two directrices has the same length as the line segment between the vertexes.

13. Show that the segment of a tangent line to a hyperbola between the hyperbola's asymptotes has its midpoint at the point of tangency.

14. The *focal radii* of a hyperbola are the line segments from the foci to points on the hyperbola. Show that the lengths of the focal radii to points (x,y) on the hyperbola $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ are $a + ex$ and $ex - a$, where e is the eccentricity.

15. Show that a tangent line to a hyperbola together with the hyperbola's asymptotes bounds a triangle of constant area (i.e. the area is independent of the point of tangency on the hyperbola).

16. Show that the product of the perpendicular distances from any point on the hyperbola $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ to the asymptotes $y = \pm \frac{b}{a}x$ is a constant.

17. A person at a point $P = (x,y)$ hears the crack of a rifle located at the point $F_1 = (-1000,0)$ and the sound of the fired bullet hitting its target located at the point $F_2 = (1000,0)$ at the same time. The bullet's speed is 2000 ft/sec and the speed of sound is 1100 ft/sec. Find an equation relating x and y.

18. Show that the set of all midpoints of a family of parallel chords either in one branch or between the two branches of a hyperbola lie on a line through the center of the hyperbola.

19. Prove a second reflection property of hyperbolas: a light shone between the two branches and directed toward one focus will reflect toward the other focus.

7.4 Translations and Rotations

For convenience the ellipses, parabolas and hyperbolas in the previous sections were centered at the origin and had their foci on one of the coordinate axes. In general the center and foci of those curves can be moved anywhere by means of **coordinate transformations**.

You have probably seen in earlier courses how the graph of a function $y = f(x)$ can be shifted horizontally by an amount h and vertically by an amount k by replacing x and y by $x-h$ and $y-k$, respectively:

$$y - k = f(x-h)$$

This coordinate transformation is called **translation**, and can be applied to any curve in the xy-plane. The origin $O = (0,0)$ is shifted to the point $O' = (h,k)$, which serves as the origin of the $x'y'$-plane,[9] as in Figure 7.4.1. Let $P = (x, y)$ be a point in the xy-plane. Consider P as a point $P' = (x', y')$ in the $x'y'$-plane, so that relative to the origin O', the x' and y' coordinates of P' are:

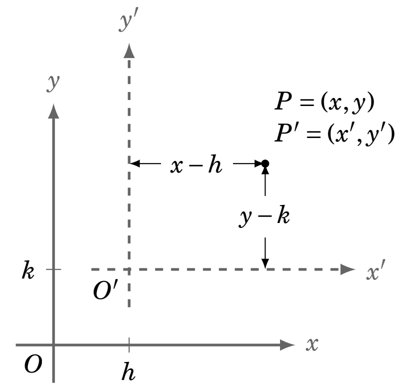


Figure 7.4.1 Translation

$$x' = x - h$$
$$y' = y - k$$

Using these translation equations (i.e. the substitutions $x \mapsto x-h$ and $y \mapsto y-k$), the graphs of the conic sections can be translated to any center (h,k):

Ellipse: For $a > b > 0$, an equation of the form

$$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1$$

describes an ellipse with center (h,k), vertexes $(h \pm a, k)$, and foci $(h \pm c, k)$, where $c^2 = a^2 - b^2$. The eccentricity is $e = \frac{c}{a}$, and the principal axis is the line $y = k$.

Likewise, an equation of the form

$$\frac{(y-k)^2}{a^2} + \frac{(x-h)^2}{b^2} = 1$$

describes an ellipse with center (h,k), vertexes $(h, k \pm a)$, and foci $(h, k \pm c)$, where $c^2 = a^2 - b^2$. The eccentricity is $e = \frac{c}{a}$, and the principal axis is the line $x = h$.

[9] The prime symbol ($'$) does not indicate differentiation—it acts merely to distinguish the new axes.

Parabola: For $p \neq 0$, an equation of the form

$$(x-h)^2 = 4p(y-k)$$

describes a parabola with vertex (h,k) and focus $(h,k+p)$. The directrix is the line $y = k-p$ and the axis is the line $x = h$.
Likewise, an equation of the form

$$(y-k)^2 = 4p(x-h)$$

describes a parabola with vertex (h,k) and focus $(h+p,k)$. The directrix is the line $x = h-p$ and the axis is the line $y = k$

Hyperbola: For $a \neq 0$ and $b \neq 0$, an equation of the form

$$\frac{(x-h)^2}{a^2} - \frac{(y-k)^2}{b^2} = 1$$

describes a hyperbola with center (h,k), vertexes $(h \pm a, k)$, and foci $(h \pm c, k)$, where $c^2 = a^2 + b^2$. The eccentricity is $e = \frac{c}{a}$, the directrices are the lines $x = h \pm \frac{a^2}{c}$, the asymptotes are the lines $y - k = \pm\frac{b}{a}(x-h)$, and the transverse axis is the line $y = k$.
Likewise, an equation of the form

$$\frac{(y-k)^2}{a^2} - \frac{(x-h)^2}{b^2} = 1$$

describes a hyperbola with center (h,k), vertexes $(h, k \pm a)$, and foci $(h, k \pm c)$, where $c^2 = a^2 + b^2$. The eccentricity is $e = \frac{c}{a}$, the directrices are the lines $y = k \pm \frac{a^2}{c}$, the asymptotes are the lines $y - k = \pm\frac{a}{b}(x-h)$, and the transverse axis is the line $x = h$.

Example 7.4

Find the vertexes and foci of the ellipse $\frac{(x-2)^2}{4} + (y-1)^2 = 1$.

Solution: The translation coordinates are $(h,k) = (2,1)$. Also, $a = 2$ and $b = 1$. Thus, the vertexes are $(h \pm a, k) = (2 \pm 2, 1) = (0,1)$ and $(4,1)$. Since $c^2 = a^2 - b^2 = 3$ then $c = \sqrt{3}$, so the foci are $(h \pm c, k) = (2 \pm \sqrt{3}, 1)$, as shown in Figure 7.4.2.

Note that the ellipse $\frac{(x-2)^2}{4} + (y-1)^2 = 1$ in the xy-plane is the ellipse $\frac{x'^2}{4} + y'^2 = 1$ in the $x'y'$-plane, for the translation equations $x' = x-2$ and $y' = y-1$ (i.e. the x'-axis is the line $y = 1$ and the y'-axis is the line $x = 2$).

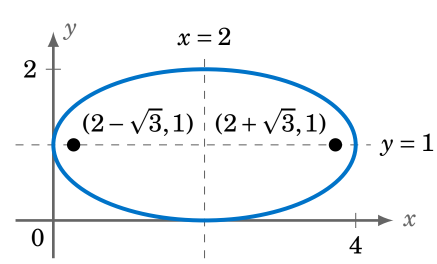


Figure 7.4.2 Ellipse $\frac{(x-2)^2}{4} + (y-1)^2 = 1$

Example 7.5

For $a \neq 0$ and constants b and c, find the vertex, focus and directrix of the parabola $y = ax^2 + bx + c$.

Solution: The idea here is to write $y = ax^2 + bx + c$ in the form $(x-h)^2 = 4p(y-k)$ for some h, k, and p, by completing the square:

$$\begin{aligned} ax^2 + bx + c &= y \\ a\left(x^2 + \frac{b}{a}x\right) &= y - c \\ a\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2}\right) &= y - c + a\frac{b^2}{4a^2} \\ \left(x + \frac{b}{2a}\right)^2 &= \frac{1}{a}\left(y + \frac{b^2 - 4ac}{4a}\right) \end{aligned}$$

So $h = -\frac{b}{2a}$, $k = \frac{4ac-b^2}{4a}$, and $4p = \frac{1}{a}$ means $p = \frac{1}{4a}$. Thus, the vertex is $(h,k) = \left(-\frac{b}{2a}, \frac{4ac-b^2}{4a}\right)$, the focus is $(h, k+p) = \left(-\frac{b}{2a}, \frac{4ac-b^2+1}{4a}\right)$, and the directrix is the line $y = k - p = \frac{4ac-b^2-1}{4a}$.

Rotation is another common coordinate transformation. Consider the case of rotating the xy-plane about the origin by an angle θ, as in Figure 7.4.3. The origin of the resulting $x'y'$-plane is unchanged from the origin $O = (0,0)$ of the xy-plane. To find the rotation equations for x' and y', let $P = (x,y)$ be a point in the xy-plane away from the origin. From trigonometry you know that for $r = \sqrt{x^2+y^2} \neq 0$ and the angle $0° \le \alpha < 360°$ measured from the positive x-axis to OP,

$$x = r\cos\alpha \quad \text{and} \quad y = r\sin\alpha\,.$$

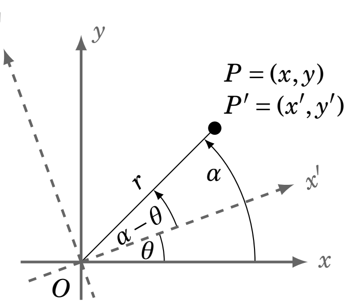


Figure 7.4.3 Rotation

Considering P as a point $P' = (x', y')$ in the $x'y'$-plane, OP' makes an angle $\alpha - \theta$ with the positive x'-axis, so that by the sine and cosine subtraction identities,

$$x' = r\cos(\alpha-\theta) = r\cos\alpha\cos\theta + r\sin\alpha\sin\theta = x\cos\theta + y\sin\theta \tag{7.7}$$

and

$$y' = r\sin(\alpha-\theta) = r\sin\alpha\cos\theta - r\cos\alpha\sin\theta = y\cos\theta - x\sin\theta\,. \tag{7.8}$$

Similar to the translation substitutions, the above rotation equations allow any curve in the xy-plane to be rotated:

To rotate a curve in the xy-plane about the origin by an angle θ, make the following substitutions:

$$x \mapsto x\cos\theta + y\sin\theta \quad \text{and} \quad y \mapsto -x\sin\theta + y\cos\theta \tag{7.9}$$

Example 7.6

Find the equation of the ellipse $\frac{x^2}{4} + y^2 = 1$ when rotated 45° counterclockwise about the origin. Simplify the equation.

Solution: For $\theta = 45°$ the substitutions are:

$$x \mapsto x\cos\theta + y\sin\theta = x\cos 45° + y\sin 45° = \frac{x+y}{\sqrt{2}}$$

$$y \mapsto -x\sin\theta + y\cos\theta = -x\sin 45° + y\cos 45° = \frac{-x+y}{\sqrt{2}}$$

The equation of the rotated ellipse (shown in Figure 7.4.4) is then:

$$\frac{1}{4}\left(\frac{x+y}{\sqrt{2}}\right)^2 + \left(\frac{-x+y}{\sqrt{2}}\right)^2 = 1$$

$$x^2 + 2xy + y^2 + 4(x^2 - 2xy + y^2) = 8$$

$$5x^2 - 6xy + 5y^2 - 8 = 0$$

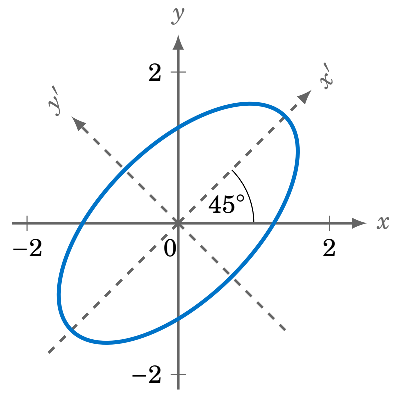


Figure 7.4.4

In the above example note the presence of the $6xy$ term in the equation of the rotated ellipse. In general if a conic section has a **second-degree equation** of the form

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \tag{7.10}$$

then $B \neq 0$ indicates rotation, and either $D \neq 0$ or $E \neq 0$ indicates translation.

Example 7.7

Find the value of a such that rotating the hyperbola $\frac{x^2}{a^2} - \frac{y^2}{a^2} = 1$ by 45° counterclockwise about the origin results in the curve $xy = 1$.

Solution: Since $\theta = 45°$ then as in Example 7.6 the substitutions are again:

$$x \mapsto \frac{x+y}{\sqrt{2}} \quad \text{and} \quad y \mapsto \frac{-x+y}{\sqrt{2}}$$

The equation of the rotated hyperbola is then:

$$\frac{1}{a^2}\left(\frac{x+y}{\sqrt{2}}\right)^2 - \frac{1}{a^2}\left(\frac{-x+y}{\sqrt{2}}\right)^2 = 1$$

$$\frac{x^2+2xy+y^2 - (x^2-2xy+y^2)}{2a^2} = 1$$

$$\frac{2xy}{a^2} = 1$$

Thus, when $a = \sqrt{2}$ the rotated hyperbola has the equation $xy = 1$, which shows that the curve $y = \frac{1}{x}$ is a hyperbola. In general any curve of the form $Bxy = 1$ is a hyperbola for $B \neq 0$.

The following result can be used for determining the type of conic section described by a second-degree equation:[10]

The graph of $Ax^2+Bxy+Cy^2+Dx+Ey+F=0$ (with A, B, C not all zero) describes a curve whose type is based on the sign of B^2-4AC:

(a) $B^2-4AC<0$: an ellipse (or a circle, point, or no curve)

(b) $B^2-4AC=0$: a parabola (or a line, two parallel lines, or no curve)

(c) $B^2-4AC>0$: a hyperbola (or two intersecting lines)

If $B\neq 0$ and the curve is a conic section, then the rotation angle θ is given by:

(a) $\theta=45°$ if $A=C$

(b) $\tan 2\theta=\frac{B}{A-C}$ if $A\neq C$, with $0°<\theta<90°$

Recall that the rotation equations (7.7) and (7.8) for x' and y' in terms of x and y provided substitutions that allowed the second-degree equation of the rotated conic section to be found. Conversely, to transform a second-degree equation in x and y into a "standard" conic section equation in terms of x' and y' (to simplify sketching the graph), the "reverse" rotation equations for x and y in terms of x' and y' are needed:

$$x = x'\cos\theta - y'\sin\theta \tag{7.11}$$

$$y = x'\sin\theta + y'\cos\theta \tag{7.12}$$

Example 7.8

Determine the type of curve whose equation is $5x^2+4xy+8y^2-36=0$, and sketch its graph.

Solution: Since $A=5$, $B=4$, and $C=8$, then $B^2-4AC=-144<0$, so the curve is an ellipse if it is a conic section. Since $B\neq 0$ and $A\neq C$, the rotation angle θ would be given by $\tan 2\theta=\frac{B}{A-C}=\frac{4}{-3}$, with $0°<\theta<90°$. Then 2θ is in the second quadrant, so that $\sin 2\theta=\frac{4}{5}$ and $\cos 2\theta=\frac{-3}{5}$. Use a half-angle identity to find θ:

$$\tan\theta = \frac{\sin 2\theta}{1+\cos 2\theta} = \frac{\frac{4}{5}}{1+\frac{-3}{5}} = 2$$

Hence, $\sin\theta=\frac{2}{\sqrt{5}}$ and $\cos\theta=\frac{1}{\sqrt{5}}$. Now use equations (7.11) and (7.12) to get expressions for x and y in terms of x' and y':

$$x = x'\cos\theta - y'\sin\theta = \frac{x'-2y'}{\sqrt{5}}$$

$$y = x'\sin\theta + y'\cos\theta = \frac{2x'+y'}{\sqrt{5}}$$

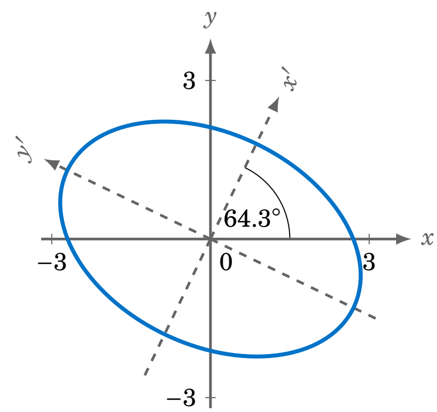
[10] For a proof see Section 6.8 in PROTTER, M.H. AND C.B. MORREY, *Analytic Geometry*, 2nd ed., Reading, MA: Addison-Wesley Publishing Company, Inc., 1975.

Substitute those expressions into $5x^2 + 4xy + 8y^2 - 36 = 0$:

$$5\left(\frac{x' - 2y'}{\sqrt{5}}\right)^2 + 4\left(\frac{x' - 2y'}{\sqrt{5}}\right)\left(\frac{2x' + y'}{\sqrt{5}}\right) + 8\left(\frac{2x' + y'}{\sqrt{5}}\right)^2 - 36 = 0$$

$$45x'^2 + 20y'^2 = 180$$

$$\frac{x'^2}{4} + \frac{y'^2}{9} = 1$$



So the curve's equation in the $x'y'$-plane is $\frac{x'^2}{4} + \frac{y'^2}{9} = 1$. This is just the ellipse $\frac{x^2}{4} + \frac{y^2}{9} = 1$ rotated counterclockwise by the angle $\theta = \tan^{-1} 2 \approx 63.4°$, as shown in the figure on the right.

Exercises

A

For Exercises 1-3, sketch the graph of the given ellipse with the exact locations of the foci and vertexes.

1. $\dfrac{(x-3)^2}{25} + \dfrac{(y-2)^2}{16} = 1$ **2.** $\dfrac{(x+1)^2}{9} + \dfrac{(y+3)^2}{4} = 1$ **3.** $4x^2 + y^2 + 24x - 2y + 21 = 0$

For Exercises 4-7, sketch the graph of the given parabola with the exact locations of the focus, vertex and directrix.

4. $4(x-1)^2 = 3(y+2)$ **5.** $y = 4x^2 + 24x + 21$ **6.** $3(y+2)^2 = 5(x-2)$ **7.** $4x^2 - 4x + 4y - 5 = 0$

For Exercises 8-10, sketch the graph of the given hyperbola with the exact locations of the foci, vertexes, directrices, and asymptotes.

8. $\dfrac{(x-3)^2}{25} - \dfrac{(y-2)^2}{16} = 1$ **9.** $\dfrac{(x+1)^2}{9} - \dfrac{(y+3)^2}{4} = 1$ **10.** $4x^2 - y^2 + 24x - 2y + 21 = 0$

11. Find the foci, vertexes, directrices and asymptotes for the hyperbola $y = \frac{1}{x}$ in Example 7.7.

12. Find the equation of the parabola $4y = x^2$ when rotated 60° counterclockwise about the origin.

13. Prove equations (7.11) and (7.12). *(Hint: Use equations (7.7) and (7.8).)*

B

14. Determine the type of curve whose equation is $8x^2 - 12xy + 17y^2 - 20 = 0$, and sketch its graph.

15. Determine the type of curve whose equation is $7x^2 + 6xy - y^2 - 32 = 0$, and sketch its graph.

16. Let $A'x'^2 + B'x'y' + C'y'^2 + D'x' + E'y' + F' = 0$ be the equation obtained by substituting the "reverse" translation equations $x = x' + h$ and $y = y' + k$ into equation (7.10). Show that $A = A'$, $B = B'$ and $C = C'$ (i.e. A, B and C are *invariant* under translation).

17. Similar to Exercise 16, substitute the "reverse" rotation equations (7.11) and (7.12) into equation (7.10) to show that $A + C$ and $B^2 - 4AC$ are invariant under rotation.

18. Sketch the graph of the conic section whose equation is $x^2 - 2xy + y^2 + 4x + 4y + 10 = 0$. *(Hint: First handle the rotation (since $A = C$ use $\theta = 45°$ even though $B^2 - 4AC = 0$), then the translation (by completing the square).)*

7.5 Hyperbolic Functions

In some textbooks you might see the sine and cosine functions called *circular functions*, since any point on the unit circle $x^2 + y^2 = 1$ can be defined in terms of those functions (see Figure 7.5.1). Those definitions motivate a similar idea for the **unit hyperbola** $x^2 - y^2 = 1$, whose points can be defined in terms of **hyperbolic functions**.

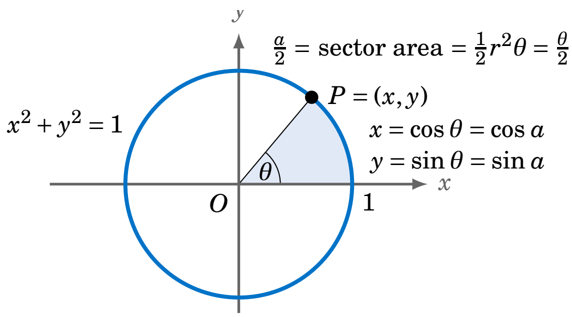


Figure 7.5.1 Circular

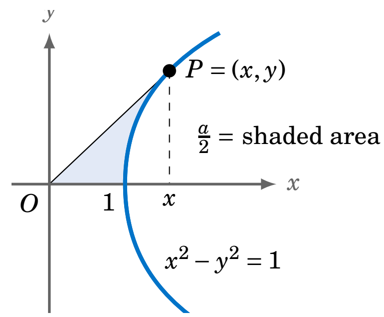


Figure 7.5.2 Hyperbolic

For a point $P = (x, y)$ on the unit hyperbola $x^2 - y^2 = 1$, the **hyperbolic angle** a is twice the area of the shaded **hyperbolic sector** in Figure 7.5.2).[11] The area $\frac{a}{2}$ thus equals the area of the right triangle with hypotenuse OP and legs of length x and y (so that the triangle's area is $\frac{1}{2}xy$) minus the area under the hyperbola over the interval $[1, x]$. So since the upper half of the hyperbola $x^2 - y^2 = 1$ is the function $y = \sqrt{x^2 - 1}$,

$$\begin{aligned}
\frac{a}{2} &= \text{(area of triangle)} - \text{(area under the hyperbola from 1 to } x\text{)} \\
&= \frac{1}{2}xy - \int_1^x \sqrt{u^2 - 1}\,du \\
\frac{a}{2} &= \frac{1}{2}x\sqrt{x^2 - 1} - \left(\frac{1}{2}x\sqrt{x^2 - 1} - \frac{1}{2}\ln\left(x + \sqrt{x^2 - 1}\right)\right) \quad \text{(by formula (6.9))} \\
a &= \ln\left(x + \sqrt{x^2 - 1}\right) \\
e^a &= x + \sqrt{x^2 - 1} \\
(e^a - x)^2 &= \left(\sqrt{x^2 - 1}\right)^2 \\
e^{2a} - 2xe^a + \cancel{x^2} &= \cancel{x^2} - 1 \\
x &= \frac{e^{2a} + 1}{2e^a} = \frac{e^a + e^{-a}}{2} = \cosh a
\end{aligned}$$

where $\cosh a$ is the **hyperbolic cosine** of a.

[11] The reason for using twice the area is merely to obtain a "cleaner" final result involving a instead of $2a$.

The y-coordinate of P can then be found:

$$y = \sqrt{x^2-1} = \sqrt{\left(\frac{e^a+e^{-a}}{2}\right)^2 - 1} = \sqrt{\frac{e^{2a}+2+e^{-2a}}{4} - \frac{4}{4}}$$

$$= \sqrt{\frac{e^{2a}-2+e^{-2a}}{4}} = \sqrt{\left(\frac{e^a-e^{-a}}{2}\right)^2}, \text{ so since } a \ge 0$$

$$y = \frac{e^a-e^{-a}}{2} = \sinh a$$

where $\sinh a$ is the **hyperbolic sine** of a. All six hyperbolic functions can now be defined in general, analogous to the trigonometric (circular) functions:

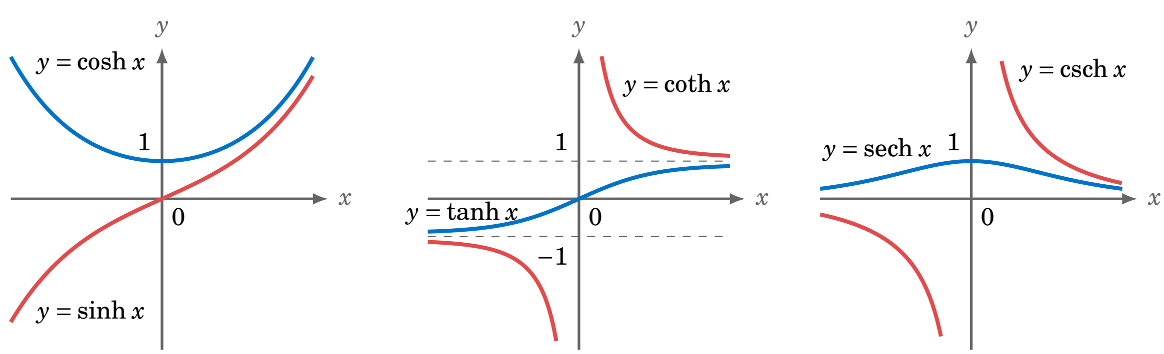
The **hyperbolic sine**, **hyperbolic cosine**, **hyperbolic tangent**, **hyperbolic cotangent**, **hyperbolic secant** and **hyperbolic cosecant**, denoted by sinh, cosh, tanh, coth, sech and csch, respectively, are:

$$\sinh x = \frac{e^x-e^{-x}}{2} \quad \text{for all } x \qquad \cosh x = \frac{e^x+e^{-x}}{2} \quad \text{for all } x$$

$$\tanh x = \frac{\sinh x}{\cosh x} \quad \text{for all } x \qquad \coth x = \frac{1}{\tanh x} \quad \text{for all } x \ne 0$$

$$\operatorname{sech} x = \frac{1}{\cosh x} \quad \text{for all } x \qquad \operatorname{csch} x = \frac{1}{\sinh x} \quad \text{for all } x \ne 0$$

The graphs of the hyperbolic functions are shown below:



(a) sinh x and cosh x **(b)** tanh x and coth x **(c)** sech x and csch x

Figure 7.5.3 Graphs of the six hyperbolic functions

The graph of $y = \cosh x$ in Figure 7.5.3(a) might look familiar: a **catenary**—a uniform cable hanging from two fixed points—has the shape of a hyperbolic cosine function.

The hyperbolic functions satisfy the following identities:

$$\cosh^2 x - \sinh^2 x = 1 \qquad \tanh^2 x + \operatorname{sech}^2 x = 1 \qquad \coth^2 x - \operatorname{csch}^2 x = 1$$

$$\sinh(-x) = -\sinh x \qquad \cosh(-x) = \cosh x \qquad \tanh(-x) = -\tanh x$$

$$\operatorname{sech}(-x) = \operatorname{sech} x \qquad \operatorname{csch}(-x) = -\operatorname{csch} x \qquad \coth(-x) = -\coth x$$

$$\sinh(u \pm v) = \sinh u \cosh v \pm \cosh u \sinh v \qquad \sinh 2x = 2 \sinh x \cosh x$$

$$\cosh(u \pm v) = \cosh u \cosh v \pm \sinh u \sinh v \qquad \cosh 2x = \cosh^2 x + \sinh^2 x$$

$$\tanh(u \pm v) = \frac{\tanh u \pm \tanh v}{1 \pm \tanh u \tanh v} \qquad \tanh 2x = \frac{2 \tanh x}{1 + \tanh^2 x}$$

The identity $\cosh^2 x - \sinh^2 x = 1$ was proved when deriving the coordinates of points on the unit hyperbola $x^2 - y^2 = 1$ in terms of the hyperbolic angle (since such a point $(x, y) = (\cosh a, \sinh a)$ must satisfy $x^2 - y^2 = 1$). The addition identities can be proved similarly using hyperbolic angles (i.e. areas).[12] However, it is simpler to use the definitions of sinh and cosh in terms of exponential functions. For example:

$$\begin{aligned}
\sinh u \cosh v + \cosh u \sinh v &= \frac{e^u - e^{-u}}{2} \cdot \frac{e^v + e^{-v}}{2} + \frac{e^u + e^{-u}}{2} \cdot \frac{e^v - e^{-v}}{2} \\
&= \frac{e^{u+v} + e^{u-v} - e^{-u+v} - e^{-u-v} + e^{u+v} - e^{u-v} + e^{-u+v} - e^{-u-v}}{4} \\
&= \frac{2e^{u+v} - 2e^{-u-v}}{4} \\
&= \frac{e^{u+v} - e^{-(u+v)}}{2} \\
&= \sinh(u+v) \quad \checkmark
\end{aligned}$$

The identity for $\sinh 2x$ is then easy to prove by letting $u = v = x$ in the above identity:

$$\sinh 2x = \sinh(x+x) = \sinh x \cosh x + \cosh x \sinh x = 2 \sinh x \cosh x \quad \checkmark$$

Note that the identities $\cosh(-x) = \cosh x$ and $\sinh(-x) = -\sinh x$ mean that cosh is an even function and sinh is an odd function. Those two functions thus (sort of) serve as even and odd versions of the exponential function (which is neither even nor odd). Both $\cosh x$ and $\sinh x$ grow exponentially (the e^{-x} term for both functions becomes negligible as $x \to \infty$), while $\sinh x$ decreases exponentially to $-\infty$ as $x \to -\infty$.

[12] See pp.25-29 in SHERVATOV, V.G., *Hyperbolic Functions*, Boston: D.C. Heath and Company, 1963.

The derivatives of the hyperbolic functions and their integral equivalents are:

$$\frac{d}{dx}(\sinh x) = \cosh x \qquad \frac{d}{dx}(\operatorname{csch} x) = -\operatorname{csch} x \coth x$$

$$\frac{d}{dx}(\cosh x) = \sinh x \qquad \frac{d}{dx}(\operatorname{sech} x) = -\operatorname{sech} x \tanh x$$

$$\frac{d}{dx}(\tanh x) = \operatorname{sech}^2 x \qquad \frac{d}{dx}(\coth x) = -\operatorname{csch}^2 x$$

$$\int \cosh x\,dx = \sinh x + C \qquad \int \operatorname{csch} x \coth x\,dx = -\operatorname{csch} x + C$$

$$\int \sinh x\,dx = \cosh x + C \qquad \int \operatorname{sech} x \tanh x\,dx = -\operatorname{sech} x + C$$

$$\int \operatorname{sech}^2 x\,dx = \tanh x + C \qquad \int \operatorname{csch}^2 x\,dx = -\coth x + C$$

For example, by definition of $\cosh x$:

$$\frac{d}{dx}(\cosh x) = \frac{d}{dx}\left(\frac{e^x + e^{-x}}{2}\right) = \frac{e^x - e^{-x}}{2} = \sinh x \quad \checkmark$$

Example 7.9

Find the derivative of $y = \sinh x^3$.

Solution: By the Chain Rule, $\dfrac{dy}{dx} = 3x^2 \cosh x^3$.

Example 7.10

Evaluate $\int \tanh x\,dx$.

Solution: Use the definition of $\tanh x$ and the substitutions $u = \cosh x$, $du = \sinh x\,dx$:

$$\int \tanh x\,dx = \int \frac{\sinh x}{\cosh x}\,dx = \int \frac{du}{u} = \ln|u| + C = \ln(\cosh x) + C$$

For any constant $a > 0$, both $y = \cosh at$ and $y = \sinh at$ satisfy the differential equation

$$y''(t) = a^2 y(t),$$

which models rectilinear motion of a particle under a repulsive force proportional to the displacement. This is just one of the many reasons why hyperbolic functions appear in so many physical applications.

Example 7.11

In the classical theory of paramagnetism, the total number n of molecules in a gas subject to a magnetic field of strength H is

$$n = 2\pi N \int_0^{\pi} e^{\frac{\mu H}{kT}\cos\theta} \sin\theta \, d\theta \,,$$

where N is the number of molecules per unit solid angle having zero potential energy, μ is the magnetic moment, k is Boltzmann's constant, and T is the temperature of the gas. Show that

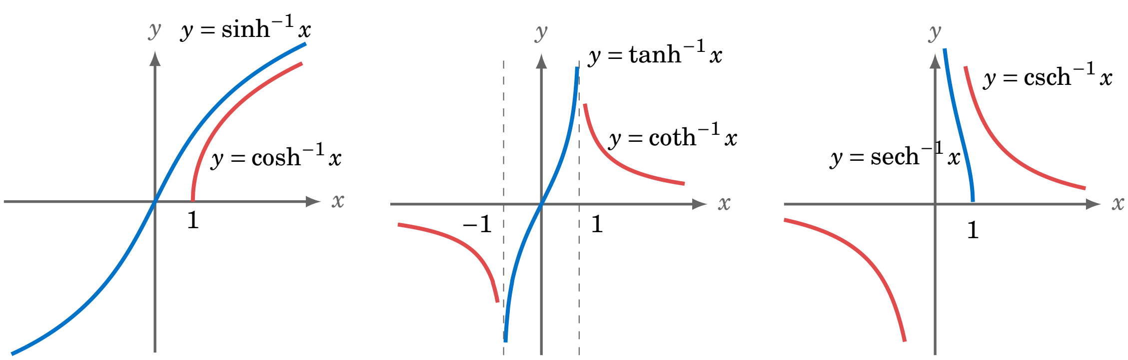
$$n = \frac{4\pi NkT}{\mu H} \sinh\left(\frac{\mu H}{kT}\right) .$$

Solution: Let $a = \frac{\mu H}{kT}$, and let $u = \cos\theta$ so that $du = -\sin\theta \, d\theta$:

$$n = -2\pi N \int_1^{-1} e^{au} \, du = 2\pi N \int_{-1}^{1} e^{au} \, du = \frac{2\pi N}{a} e^{au} \Big|_{-1}^{1} = \frac{2\pi N}{a}\left(2 \cdot \frac{e^a - e^{-a}}{2}\right)$$
$$= \frac{4\pi NkT}{\mu H} \sinh\left(\frac{\mu H}{kT}\right)$$

Since $\frac{d}{dx}(\sinh x) = \cosh x = \frac{e^x + e^{-x}}{2} > 0$ for all x, then $y = \sinh x$ is an increasing function and thus its inverse function $x = \sinh^{-1} y$ is defined. The remaining **inverse hyperbolic functions** can be defined similarly, with the following domains and ranges (switching the roles of x and y, as usual) plus their graphs:

function	$\sinh^{-1} x$	$\cosh^{-1} x$	$\tanh^{-1} x$	$\operatorname{csch}^{-1} x$	$\operatorname{sech}^{-1} x$	$\coth^{-1} x$
domain	all x	$x \ge 1$	$\lvert x\rvert < 1$	all $x \ne 0$	$0 < x \le 1$	$\lvert x\rvert > 1$
range	all y	$y \ge 0$	all y	all $y \ne 0$	$y \ge 0$	all $y \ne 0$



(a) $\sinh^{-1} x$ and $\cosh^{-1} x$ **(b)** $\tanh^{-1} x$ and $\coth^{-1}$ **(c)** $\operatorname{sech}^{-1} x$ and $\operatorname{csch}^{-1} x$

Figure 7.5.4 Graphs of the six inverse hyperbolic functions

The inverse hyperbolic functions can be expressed in terms of the natural logarithm:

$$\sinh^{-1} x = \ln\left(x + \sqrt{x^2+1}\right) \qquad \cosh^{-1} x = \ln\left(x + \sqrt{x^2-1}\right) \quad \text{for } x \ge 1$$

$$\tanh^{-1} x = \frac{1}{2}\ln\frac{1+x}{1-x} \quad \text{for } |x| < 1 \qquad \coth^{-1} x = \frac{1}{2}\ln\frac{x+1}{x-1} \quad \text{for } |x| > 1$$

$$\operatorname{sech}^{-1} x = \ln\frac{1+\sqrt{1-x^2}}{x} \quad \text{for } 0 < x \le 1 \qquad \operatorname{csch}^{-1} x = \ln\left(\frac{1}{x} + \sqrt{\frac{1}{x^2}+1}\right) \quad \text{for } x \ne 0$$

Notice that the above formula for $\cosh^{-1} x$ was actually proved at the beginning of this section. The formula for $\sinh^{-1} x$ follows from the definition of an inverse function:

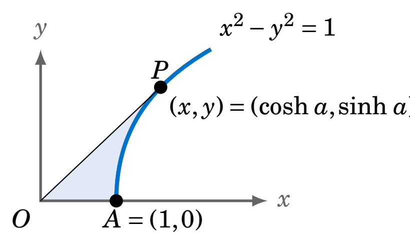
$$\begin{aligned}
y = \sinh^{-1} x \quad &\Rightarrow \quad x = \sinh y = \frac{e^y - e^{-y}}{2} \\
&\Rightarrow \quad e^y - 2x - e^{-y} = 0 \\
&\Rightarrow \quad e^{2y} - 2xe^y - 1 = 0 \\
&\Rightarrow \quad u^2 - 2xu - 1 = 0 \quad \text{for } u = e^y \\
&\Rightarrow \quad u = \frac{2x \pm \sqrt{4x^2 - 4(1)(-1)}}{2} = x \pm \sqrt{x^2+1} \\
&\Rightarrow \quad e^y = u = x + \sqrt{x^2+1} \quad \text{since } e^y > 0 \\
&\Rightarrow \quad y = \ln\left(x + \sqrt{x^2+1}\right) \quad \checkmark
\end{aligned}$$

The remaining formulas can be proved similarly.

Example 7.12

Show that for the hyperbolic angle a of a point $P = (x, y)$ on the unit hyperbola $x^2 - y^2 = 1$, the area $\frac{a}{2}$ of the hyperbolic sector $\sphericalangle OAP$ (the shaded region in the figure on the right) is

$$\frac{a}{2} = \frac{1}{2}\cosh^{-1} x\,.$$



$\frac{a}{2}$ = shaded area of $\sphericalangle OAP$

Solution: It was shown earlier that the area of $\sphericalangle OAP$ is

$$\frac{a}{2} = \frac{1}{2}\ln\left(x + \sqrt{x^2-1}\right)$$

so by the formula $\cosh^{-1} x = \ln\left(x + \sqrt{x^2-1}\right)$ the result follows.
This makes sense, since $x = \cosh a$ and so $\cosh^{-1} x = \cosh^{-1}(\cosh a) = a$, by definition of an inverse.

You can use either the general formula for the derivative of an inverse function or the above formulas to find the derivatives of the inverse hyperbolic functions:

$$\frac{d}{dx}(\sinh^{-1} x) = \frac{1}{\sqrt{x^2+1}} \qquad \frac{d}{dx}(\cosh^{-1} x) = \frac{1}{\sqrt{x^2-1}} \quad \text{for } x \ge 1$$

$$\frac{d}{dx}(\tanh^{-1} x) = \frac{1}{1-x^2} \quad \text{for } |x|<1 \qquad \frac{d}{dx}(\coth^{-1} x) = \frac{1}{1-x^2} \quad \text{for } |x|>1$$

$$\frac{d}{dx}(\operatorname{sech}^{-1} x) = \frac{-1}{x\sqrt{1-x^2}} \quad \text{for } 0<x\le 1 \qquad \frac{d}{dx}(\operatorname{csch}^{-1} x) = \frac{-1}{|x|\sqrt{1+x^2}} \quad \text{for } x \ne 0$$

For example, here is one way to find the derivative of $\tanh^{-1} x$:

$$\frac{d}{dx}(\tanh^{-1} x) = \frac{d}{dx}\left(\frac{1}{2}\ln\frac{1+x}{1-x}\right) = \frac{1}{2}\frac{d}{dx}(\ln(1+x) - \ln(1-x))$$

$$= \frac{1}{2}\left(\frac{1}{1+x} - \frac{-1}{1-x}\right) = = \frac{1-x+(1+x)}{2(1+x)(1-x)} = \frac{1}{1-x^2} \quad \checkmark$$

Exercises

A

1. Prove the identities for $\sinh(-x)$, $\cosh(-x)$, $\tanh(-x)$, $\coth(-x)$, $\operatorname{sech}(-x)$, and $\operatorname{csch}(-x)$ on p.232.

2. Prove the identities for $\cosh(u \pm v)$, $\tanh(u \pm v)$, and $\tanh 2x$ on p.232.

For Exercises 3-15 prove the given identity. Your proofs can use other identities.

3. $(\cosh x + \sinh x)^r = \cosh rx + \sinh rx$ for all r

4. $\sinh 3x = 3\sinh x + 4\sinh^3 x$

5. $\sinh A \cosh B = \frac{1}{2}(\sinh(A+B) + \sinh(A-B))$

6. $\tanh^2 x + \operatorname{sech}^2 x = 1$

7. $\sinh A \sinh B = \frac{1}{2}(\cosh(A+B) - \cosh(A-B))$

8. $\coth^2 x - \operatorname{csch}^2 x = 1$

9. $\cosh A \cosh B = \frac{1}{2}(\cosh(A+B) + \cosh(A-B))$

10. $\coth^{-1}\left(\frac{1}{x}\right) = \tanh^{-1} x$

11. $\cosh 2x = \cosh^2 x + \sinh^2 x = 2\cosh^2 x - 1 = 1 + 2\sinh^2 x$

12. $\sinh^2 \frac{x}{2} = \dfrac{\cosh x - 1}{2}$

13. $\cosh^2 \frac{x}{2} = \dfrac{\cosh x + 1}{2}$

14. $\tanh^2 \frac{x}{2} = \dfrac{\cosh x - 1}{\cosh x + 1}$

15. $\tanh \frac{x}{2} = \dfrac{\cosh x - 1}{\sinh x} = \dfrac{\sinh x}{\cosh x + 1}$

16. Prove the identities for $\tanh^{-1} x$, $\coth^{-1} x$, $\operatorname{sech}^{-1} x$, and $\operatorname{csch}^{-1} x$ on p.235.

17. Prove the derivative formulas for $\sinh x$, $\tanh x$, $\coth x$, $\operatorname{sech} x$, and $\operatorname{csch} x$ on p.233.

18. Prove the derivative formulas for $\sinh^{-1} x$, $\cosh^{-1} x$, $\coth^{-1} x$, $\operatorname{sech}^{-1} x$, and $\operatorname{csch}^{-1} x$ on p.236.

19. Show that $\cosh x = O(e^x)$ and $\sinh x = O(e^x)$.

20. Show that $\dfrac{d}{dx}(\tan^{-1}(\sinh x)) = \operatorname{sech} x$.

21. Verify that the curve $y = \tanh x$ has asymptotes $y = \pm 1$.

22. Show that $\sinh dx = dx$ and $\cosh dx = 1$ for any infinitesimal dx. *(Hint: Use $\frac{1}{1+dx} = 1 - dx$ from Exercise 3 in Section 1.3 along with $e^{dx} = 1 + dx$ from Exercise 29 in Section 2.3.)*

23. Use Exercise 22 and the addition formula for $\sinh x$ to show that $\frac{d}{dx}(\sinh x) = \cosh x$.

24. Sketch the graph of $f(x) = e^{-2x}\sinh x$. Find all local maxima and minima, inflection points, and vertical or horizontal asymptotes.

25. Denoting the speed of light by c, the *Lorentz transformations* for two inertial frames S and S' are

$$x' = \frac{x - vt}{\sqrt{1-(v/c)^2}} \quad \text{and} \quad t' = \frac{t - vx/c^2}{\sqrt{1-(v/c)^2}}$$

where S' moves with speed $v > 0$ parallel to the x-axis of S. Let $v/c = \tanh\xi$. Show that

$$x' = -ct\sinh\xi + x\cosh\xi \quad \text{and} \quad t' = t\cosh\xi - (x/c)\sinh\xi\ .$$

26. Evaluate the integral $I = \int_2^3 \frac{dx}{1-x^2}$ in two different ways:

(a) Use $\frac{d}{dx}(\coth^{-1}x) = \frac{1}{1-x^2}$ for $|x| > 1$ to show that $I = \coth^{-1}3 - \coth^{-1}2$.

(b) Use the substitution $u = \frac{1}{x}$ in the integral, then use $\frac{d}{dx}(\tanh^{-1}x) = \frac{1}{1-x^2}$ for $|x| < 1$ to show that $I = \tanh^{-1}\frac{1}{3} - \tanh^{-1}\frac{1}{2}$. Is this answer equivalent to the answer from part (a)? Explain.

B

27. The general solution of the differential equation $y'' = a^2 y$ is $y(t) = y_1(t) = c_1e^{at} + c_2e^{-at}$, where a is a positive constant, and c_1 and c_2 are arbitrary constants.

(a) Verify that $y(t) = y_2(t) = k_1\cosh at + k_2\sinh at$ is also a solution of $y'' = a^2 y$.

(b) Show that for any c_1 and c_2, $y_1(t) = c_1e^{at} + c_2e^{-at}$ can be written as $y_1(t) = k_1\cosh at + k_2\sinh at$ for some constants k_1 and k_2 in terms of c_1 and c_2.

28. Verify that for positive constants β, p, l and c the function

$$\eta(x) = \frac{\beta x}{p} - \frac{\beta c\sinh(px/c)}{p^2\cosh(pl/c)}$$

is a solution of the differential equation (related to water displacement in a canal of length $2l$)

$$\frac{d^2\eta}{dx^2} - \frac{p^2}{c^2}\eta = -\frac{\beta x p}{c^2}\ .$$

29. For any constant a and for $s > |a|$ the *Laplace transform* $\mathscr{L}(s)$ of the function $f(t) = \sinh at$ is

$$\mathscr{L}(s) = \int_0^\infty e^{-st}\sinh at\ dt\ .$$

Show that $\mathscr{L}(s) = \frac{a}{s^2-a^2}$.

C

30. In quantum mechanics the *scaling factor* e^{π/s_0} of the three-particle *Efimov trimer* is the solution $s = s_0 > 0$ of the equation

$$s\cosh\left(\tfrac{\pi s}{2}\right) = \frac{8\sinh\left(\tfrac{\pi s}{6}\right)}{\sqrt{3}}\ ,$$

Use a numerical method to approximate s_0, then calculate e^{π/s_0}.

31. Continuing Example 7.11, the *total magnetic moment* M is defined as

$$M = 2\pi N\mu \int_0^\pi e^{\frac{\mu H}{kT}\cos\theta} \sin\theta \cos\theta \, d\theta \,.$$

(a) Use the value of n from Example 7.11 to show that $\frac{M}{n\mu} = L(a)$, where $L(a) = \coth a - \frac{1}{a}$ is the *Langevin function* and $a = \frac{\mu H}{kT}$.

(b) Show that for $a = \frac{\mu H}{kT}$, $\mu_{\text{max}} > 0$, and $x = \frac{\mu_{\text{max}} H}{kT}$,

$$\int_0^{\mu_{\text{max}}} L(a)\, d\mu = \frac{1}{x} \ln\left(\frac{\sinh x}{x}\right) .$$

32. The age t_0 of the universe (in years) is given by

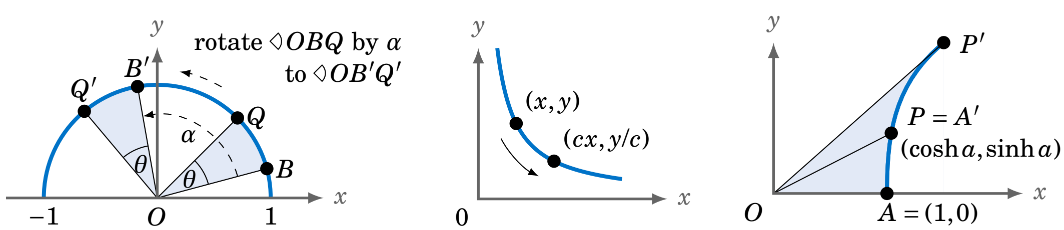
$$t_0 = \tau_0 \int_0^1 \frac{dx}{\sqrt{\frac{2q_0}{x} + 1 - 2q_0}} \,,$$

where $\tau_0 = 2 \times 10^{10}$ years is the *Hubble time* and $q_0 \geq 0$ is the *deceleration parameter*. For the cosmological model with $0 < q_0 < \frac{1}{2}$, use the substitution $x = \frac{2q_0}{1-2q_0}\sinh\theta$ to show that

$$t_0 = \tau_0 \left(\frac{1}{1-2q_0} - \frac{2q_0}{(1-2q_0)^{3/2}} \cosh^{-1}\left(\frac{1}{\sqrt{2q_0}}\right)\right) .$$

What fraction of τ_0 is t_0 (i.e. what is $\frac{t_0}{\tau_0}$) when $q_0 = \frac{1}{4}$? *(Hint: Use Exercise 11 or 12.)*

33. Circular rotations preserve the area of circular sectors (see Figure 7.5.5(a)). For any constant $c > 0$ the *hyperbolic rotation* $\phi : (x, y) \mapsto (cx, y/c)$ moves points along the hyperbola $xy = k$ (for $k > 0$), as shown in Figure 7.5.5(b) for $c > 1$. This hyperbolic rotation preserves the area of hyperbolic sectors.



(a) Unit circle $x^2 + y^2 = 1$ **(b)** Hyperbola $xy = k$ **(c)** Unit hyperbola $x^2 - y^2 = 1$

Figure 7.5.5 Circular and hyperbolic rotations

As an example of why this is true, let $a > 0$ and consider the unit hyperbola in Figure 7.5.5(c). Then:

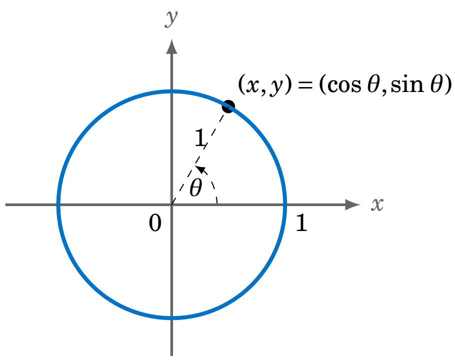
(a) Let $c > 0$. When $c \neq 1$ the mapping $\phi : (x, y) \mapsto (cx, y/c)$ does not move points along the unit hyperbola. Find the formula for ϕ on the unit hyperbola as follows: use the rotation equations from Section 7.4 to rotate the unit hyperbola 45° to a hyperbola of the form $xy = k$, apply ϕ to a generic point on that hyperbola, then rotate that hyperbola by −45° back to the unit hyperbola.

(b) Use part (a) to find the value of c such that ϕ maps $A = (1, 0)$ to $P = (\cosh a, \sinh a)$.

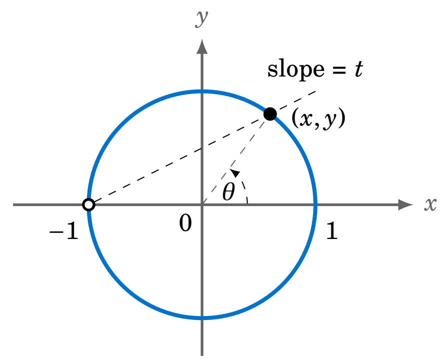
(c) Let P' be the point that ϕ maps P to, and let $A' = P$. Use part (b) to find the coordinates of P', then show that the hyperbolic sectors $\sphericalangle OAP$ and $\sphericalangle OA'P'$ have the same area $a/2$.

7.6 Parametric Equations

Recall that Section 6.5 presented two different ways to "identify" or represent points on the unit circle—by angle and by slope, as in Figure 7.6.1:



(a) Identify points by angle θ



(b) Identify points by slope t

Figure 7.6.1 Points on the unit circle $x^2 + y^2 = 1$

When identifying by the angle θ, all points (x, y) on the unit circle can be written as

$$x = \cos\theta \quad \text{and} \quad y = \sin\theta$$

for any angle θ. When identifying by the slope t of lines through the point $(-1, 0)$, recall from the derivation of the half-angle substitution that $\sin\theta = \frac{2t}{1+t^2}$ and $\cos\theta = \frac{1-t^2}{1+t^2}$. So all points (x, y) on the unit circle except $(-1, 0)$ can be written as

$$x = \frac{1-t^2}{1+t^2} \quad \text{and} \quad y = \frac{2t}{1+t^2}$$

for any slope t. These two distinct "identifications" are called **parametrizations** of the unit circle, with **parameter** θ in the first case and parameter t in the second.

In general, one way to describe a **plane curve** C (i.e. a curve in the xy-plane) is to write its x and y-coordinates as functions of a variable t:

$$x = x(t) \quad \text{and} \quad y = y(t)$$

These are **parametric equations** of C, which consists of all points (x, y) such that $x = x(t)$ and $y = y(t)$ for the parameter t in some interval I. The shorthand for this is:

$$C : x = x(t),\ y = y(t),\ t \text{ in } I$$

Notice the flexibility that parametric equations provide, since plane curves can take any shape, not limited to the graph of a single function $y = f(x)$. In fact, a curve $y = f(x)$ is the special case where the parametric equations are $x = t$ and $y = f(t)$. In physical settings the parameter t often denotes time, but it can represent anything and any symbol can be used in its place. A curve can have many parametrizations.

Example 7.13

Show that for any constants $\omega \neq 0$ and $r > 0$, and for t measured in radians,

$$x = h + r\cos\omega t \quad \text{and} \quad y = k + r\sin\omega t \quad \text{for } -\infty < t < \infty$$

is a parametrization of the circle $(x-h)^2+(y-k)^2 = r^2$ with center (h,k) and radius r.

Solution: Since ωt is similar to the angle θ in Figure 7.6.1(a), it suffices to show that $(x-h)^2+(y-k)^2 = r^2$:

$$(x-h)^2 + (y-k)^2 = r^2\cos^2\omega t + r^2\sin^2\omega t = r^2 \quad \checkmark$$

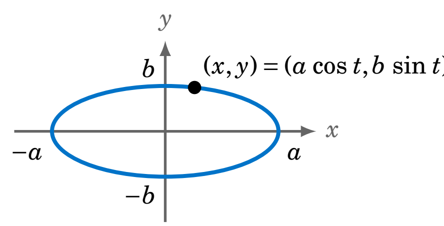
The constant ω determines how fast and in which direction the circle is traced as the parameter t varies. For example, for $\omega = 2$ the circle C is traced counterclockwise at twice the speed of the parametrization $C: x = h + r\cos t,\ y = k + r\sin t$. In all cases the circle is re-traced every $2\pi/\omega$ radians. For that reason the interval for t is often restricted to the interval $[0, 2\pi/\omega]$, so that the circle is traced only once.

Example 7.14

Show that for $a > 0$ and $b > 0$ the parametric equations

$$x = a\cos t \quad \text{and} \quad y = b\sin t$$

for $0 \le t \le 2\pi$ describe an ellipse.



Solution: Since

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{a^2\cos^2 t}{a^2} + \frac{b^2\sin^2 t}{b^2} = \cos^2 t + \sin^2 t = 1$$

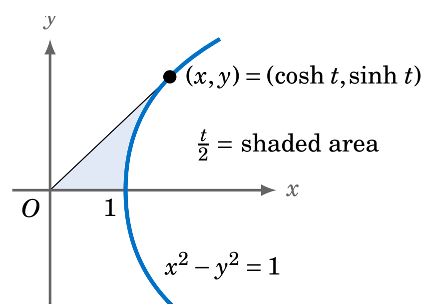
for all t, then the points $(x,y) = (a\cos t, b\sin t)$ lie on the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. It should be obvious that the entire ellipse is traced, but it is left as an exercise to show what the parameter t represents.

Example 7.15

Show that the parametric equations

$$x = \cosh t \quad \text{and} \quad y = \sinh t$$

for $-\infty < t < \infty$ describe one branch of a hyperbola.



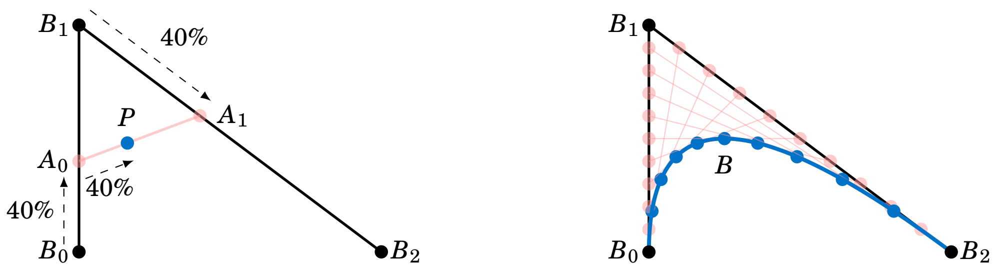
Solution: Since

$$x^2 - y^2 = \cosh^2 t - \sinh^2 t = 1$$

for all t, then the points $(x,y) = (\cosh t, \sinh t)$ lie on the unit hyperbola $x^2 - y^2 = 1$. This was in fact shown in Section 7.5, where t is half the area of the shaded region in the above figure for $t > 0$. For $t < 0$ the shaded region is reflected below the x-axis. Since $\cosh t \ge 1$ and $\sinh t$ can take any value, then the entire right branch of the hyperbola is traced as t varies. Likewise the left branch has parametric equations $x = -\cosh t$ and $y = \sinh t$. The hyperbola is thus parametrized by area (or negative area for $t < 0$). In general, for $a > 0$ and $b > 0$ the hyperbola $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ has parametric equations $x = \pm a\cosh t$ and $y = b\sinh t$ for all t.

Example 7.16

Bézier curves[13] are used in Computer Aided Design (CAD) to join the ends of an open polygonal path of noncollinear *control points* with a smooth curve that models the "shape" of the path. The curve is created via *repeated linear interpolation*, illustrated in Figure 7.6.2 and described below for $n = 3$ points:



(a) $t = 0.4 = 40\%$ **(b)** Many values of t in $[0,1]$

Figure 7.6.2 Bézier curve B with 3 control points B_0, B_1, B_2

For three control points B_0, B_1, B_2, pick $0 \le t \le 1$. Say $t = 0.4$, which you can think of as a percentage: $0.4 = 40\%$. Let A_0 and A_1 be the points 40% of the way from B_0 to B_1 and from B_1 to B_2, respectively, as in Figure 7.6.2(a). Then the point P that is 40% of the way from A_0 to A_1 is on the Bézier curve B joining B_0 and B_2. Do this for every t in $[0,1]$ to fill out the curve B, as in Figure 7.6.2(b).

It can be shown via *de Casteljau's algorithm* that the Bézier curve B for any three control points $B_0 = (x_0, y_0)$, $B_1 = (x_1, y_1)$ and $B_2 = (x_2, y_2)$ in the xy-plane has parametric equations

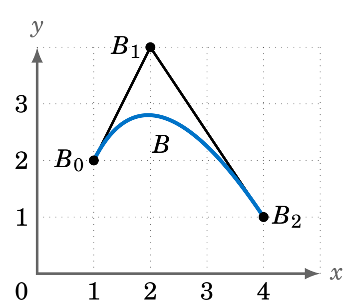
$$x \;=\; (1-t)^2 x_0 \;+\; 2t(1-t)x_1 \;+\; t^2 x_2 \quad \text{and} \quad y \;=\; (1-t)^2 y_0 \;+\; 2t(1-t)y_1 \;+\; t^2 y_2 \tag{7.13}$$

for $0 \le t \le 1$. Write out and simplify the explicit parametric equations for the Bézier curve B with control points $B_0 = (1,2)$, $B_1 = (2,4)$ and $B_2 = (4,1)$.

Solution: The parametric equations for $B_0 = (x_0, y_0) = (1,2)$, $B_1 = (x_1, y_1) = (2,4)$, $B_2 = (x_2, y_2) = (4,1)$ are:

$$\begin{aligned} x &= (1-t)^2(1) + 2t(1-t)(2) + t^2(4) = 1 - 2t + t^2 + 4t - 4t^2 + 4t^2 = t^2 + 2t + 1 \\ y &= (1-t)^2(2) + 2t(1-t)(4) + t^2(1) = 2 - 4t + 2t^2 + 8t - 8t^2 + t^2 = -5t^2 + 4t + 2 \end{aligned}$$

The Bézier curve $B : x = t^2 + 2t + 1,\ y = -5t^2 + 4t + 2,\ 0 \le t \le 1$ for B_0, B_1, B_2 is shown in the figure on the right. It is left as an exercise to show that this curve is part of a parabola. In general Bézier curves can be created for $n \ge 3$ control points in the plane, with the parametric equations being polynomials of degree $n - 1$ in the parameter t. In the exercises you will be guided in how to derive the parametric equations in the cases $n = 3$ and $n = 4$. Bézier curves can also be constructed for control points in three-dimensional space. A similar construct—a *Bézier surface*—is used in three dimensions to model the boundary of a polyhedron (i.e. a solid whose faces are polygons).



[13] Developed and popularized in the 1960s by two engineers, Pierre Bézier and Paul de Casteljau, for vehicle body modeling at the French automotive manufacturers Renault and Citroën, respectively.

A curve with parametric equations $x = x(t)$ and $y = y(t)$ might not be the graph of a single function $y = f(x)$, but the derivative $\frac{dy}{dx}$ can still be found by using the differentials of x and y as functions of t: $dy = y'(t)\,dt$ and $dx = x'(t)\,dt$, so that

$$\frac{dy}{dx} = \frac{y'(t)\,dt}{x'(t)\,dt} = \frac{y'(t)}{x'(t)} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} \tag{7.14}$$

when $\frac{dx}{dt} \neq 0$. The second derivative $\frac{d^2y}{dx^2}$ can then be found via the Chain Rule:

$$\frac{d}{dt}\left(\frac{dy}{dx}\right) = \left(\frac{d}{dx}\left(\frac{dy}{dx}\right)\right) \cdot \frac{dx}{dt} = \frac{d^2y}{dx^2} \cdot \frac{dx}{dt} \quad\Rightarrow\quad \frac{d^2y}{dx^2} = \frac{\frac{d}{dt}\left(\frac{dy}{dx}\right)}{\frac{dx}{dt}} \tag{7.15}$$

For t in $[a,b]$ with $x_1 = x(a)$ and $x_2 = x(b)$, the integral $\int_{x_1}^{x_2} y\,dx$ is given by:

$$\int_{x_1}^{x_2} y\; dx = \int_a^b y(t)x'(t)\; dt \tag{7.16}$$

Example 7.17

A *cycloid* is the path of a point P on a circle rolling along a straight line. Figure 7.6.3 shows the cycloid C traced by a circle of radius a rolling along the x-axis so that P touches the origin during the roll:

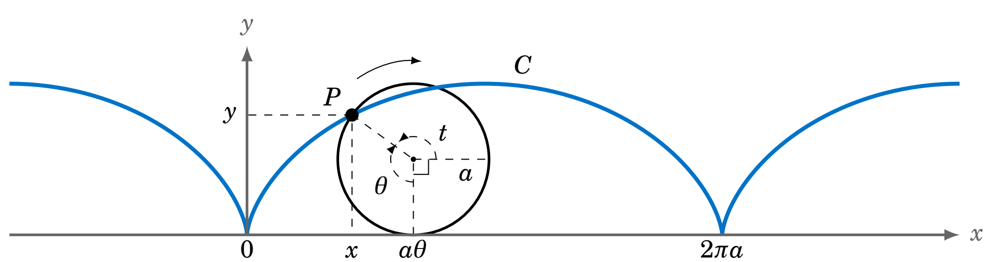


Figure 7.6.3 Cycloid C for a circle of radius a

Find parametric equations for C and find $\frac{dy}{dx}$.

Solution: For the angles t and θ—measured in radians—shown in Figure 7.6.3, $t + \theta + \pi/2 = 2\pi$, so $t = 3\pi/2 - \theta$. The point P touches the origin as the circle rolls, so the horizontal distance from the circle's center to the y-axis is the length of the circular arc with central angle θ, namely $a\theta$. So by the parametrization of the circle as in Example 7.13, but with center $(h,k) = (a\theta, a)$, radius $r = a$, and $\omega = 1$,

$$x = a\theta + a\cos t = a\left(\theta + \cos\left(\tfrac{3\pi}{2} - \theta\right)\right) = a(\theta - \sin\theta)$$
$$y = a + a\sin t = a\left(1 + \sin\left(\tfrac{3\pi}{2} - \theta\right)\right) = a(1 - \cos\theta)$$

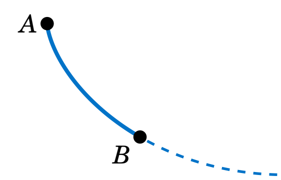
Thus, $C : x = a(\theta - \sin\theta)$, $y = a(1 - \cos\theta)$, $-\infty < \theta < \infty$ is a parametrization of the cycloid C.

As in formula (7.14) the derivative $\frac{dy}{dx}$ is given by:

$$\frac{dy}{dx} = \frac{y'(\theta)}{x'(\theta)} = \frac{a\sin\theta}{a(1-\cos\theta)} = \frac{\sin\theta}{1-\cos\theta} = \cot\tfrac{1}{2}\theta$$

Thus, $\frac{dy}{dx}$ is undefined when $\cos\theta = 1$, namely, when $\theta = 2\pi k$ for all integers k, i.e. when $x = a(\theta - \sin\theta) = a(2\pi k - \sin 2\pi k) = 2\pi k a$. Notice from Figure 7.6.3 that the cycloid has cusps at those values of x.

A cycloid appears in the solution of the famous *brachistochrone problem*:[14] find the plane curve joining two points A and B—where B is at a lower height than A but not directly under it—along which an object slides frictionless under the force of gravity alone from A to B in the shortest time. It turns out that the optimal path is not a straight line, but part of an inverted (upside-down) cycloid with a cusp at A, as in the figure on the right.[15]



Exercises

A

1. For $a > b > 0$, in the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ inscribe a circle of radius b centered at the origin, called the *minor auxiliary circle*. This circle is parametrized by $x = b\cos t$ and $y = b\sin t$ for $0 \le t \le 2\pi$, with the angle t shown in Figure 7.6.4. From each point on that circle draw a horizontal line segment to the point P on the ellipse in the same quadrant, as shown. Show that $P = (a\cos t, b\sin t)$.
Note: The angle t is called the *eccentric angle* of the ellipse, and it is the parameter for the parametrization of the ellipse in Example 7.14.

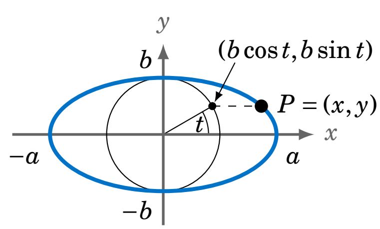


Figure 7.6.4 Exercise 1

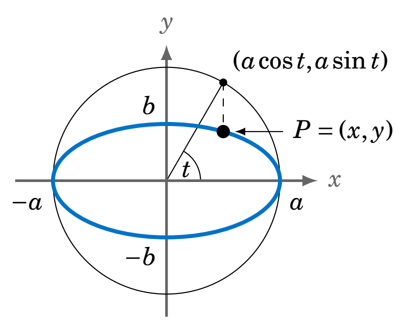


Figure 7.6.5 Exercise 2

2. Similar to Exercise 1, circumscribe a circle of radius a (called the *major auxiliary circle*) around the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. This circle is parametrized by $x = a\cos t$ and $y = a\sin t$ for $0 \le t \le 2\pi$, with the eccentric angle t shown in Figure 7.6.5. From each point on that circle draw a vertical line segment to the point P on the ellipse in the same quadrant, as shown. Show again that $P = (a\cos t, b\sin t)$.

3. Show that the cycloid in Example 7.17 has global maxima at $x = (2k+1)\pi a$ for all integers k, and that the cycloid is always concave down.

4. Show that the area under the cycloid in Example 7.17 over the interval $[0, 2\pi a]$ is $3\pi a^2$.

[14] First solved in 1696 by the Swiss physicist and mathematician Johann Bernoulli (1667-1748).
[15] See pp.60-62 in CLEGG, J.C., *Calculus of Variations*, Edinburgh: Oliver & Boyd, Ltd., 1968. For Bernoulli's proof see pp.644-655 in SMITH, D.E., *A Source Book in Mathematics*, New York: Dover Publications, Inc., 1959.

5. The parametrization $C: x = \frac{1-t^2}{1+t^2},\ y = \frac{2t}{1+t^2},\ -\infty < t < \infty$ of the unit circle C shown earlier makes the unit circle a *rational curve*, since x and y are rational functions of the parameter t. Is the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ a rational curve? Justify your answer.

6. Let $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$ be distinct points in the xy-plane. For t in $[0,1]$ define

$$x(t) = (1-t)x_0 + tx_1 \quad \text{and} \quad y(t) = (1-t)y_0 + ty_1$$

and let $P(t) = (x(t), y(t))$.

(a) Show that $C: x = x(t),\ y = y(t),\ 0 \le t \le 1$ is a parametrization of the line segment from P_0 to P_1.

(b) Show that the parameter t is the proportion of the length $P_0P(t)$ to the length P_0P_1: $\frac{P_0P(t)}{P_0P_1} = t$.

7. In an *ellipsograph* a rod of length a has a peg at one end and another peg a distance b from the other end, so that the rod can slide along two thin perpendicular rails, with the pegs in the rails as in Figure 7.6.6. Show that a marker at the endpoint P traces an ellipse as the rod moves. *(Hint: Treat the rails as the x and y-axes. Find parametric equations, with a different angle than in Exercise 1.)*

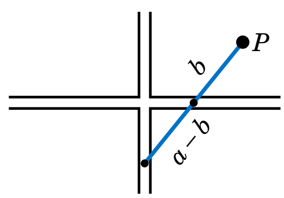


Figure 7.6.6 Exercise 7

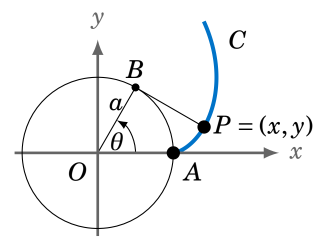


Figure 7.6.7 Exercise 8

B

8. The end P of a thread is held taut as it is unwound from a circle of radius a starting from a point A, as in Figure 7.6.7. Show that the path C that the end traces—called the *involute* of the circle—has parametric equations $x = a(\cos\theta + \theta\sin\theta),\ y = a(\sin\theta - \theta\cos\theta)$.

9. Recall that a parabola of the form $y = ax^2 + bx + c$ has a constant second derivative $\frac{d^2y}{dx^2} = 2a$. Consider the Bézier curve B in Example 7.16.

(a) Find $\frac{d^2y}{dx^2}$ for B. Is it a constant?

(b) Show that B is part of a parabola. Does this contradict part (a)? Explain. *(Hint: Show the curve has a second-degree equation of the form (7.10).)*

(c) Find the point where the curve B has a global maximum.

10. Use Exercise 6 to derive the formulas (7.13) for the parametric equations of the general Bézier curve for $n = 3$ control points.

11. To form the Bézier curve for $n = 4$ control points $B_0 = (x_0, y_0)$, $B_1 = (x_1, y_1)$, $B_2 = (x_2, y_2)$, $B_3 = (x_3, y_3)$, for $0 \le t \le 1$ use the three points that are $100t\%$ of the way along the line segments $\overline{B_0B_1}$, $\overline{B_1B_2}$, $\overline{B_2B_3}$ as the control points in the $n = 3$ case. Show that the resulting parametrization is:

$$x = (1-t)^3x_0 + 3t(1-t)^2x_1 + 3t^2(1-t)x_2 + t^3x_3 \quad \text{and} \quad y = (1-t)^3y_0 + 3t(1-t)^2y_1 + 3t^2(1-t)y_2 + t^3y_3$$

12. Each point on a plane curve lies on some line through the origin. Use that fact to show that the equation $y^2 = x^2 + x^3$ defines a rational curve (see Exercise 5). *(Hint: For all real t, find the intersections of the lines $y = tx$ with the curve. Consider also the special case of the line $x = 0$.)*

13. Sketch the graph of the curve $C: x = 2t - 4t^3,\ y = t^3 - 3t^4,\ -\infty < t < \infty$.

7.7 Polar Coordinates

Suppose that you wanted to write the equation of a spiral, like the one in Figure 7.7.1. The curve is clearly not the graph of a function $y = f(x)$ in Cartesian coordinates, as it violates the vertical line test. However, this spiral is simple to express using **polar coordinates**.[16] Recall that any point P distinct from the origin (denoted by O) in the xy-plane is a distance $r > 0$ from the origin, and the ray $\overrightarrow{OP}$ makes an angle θ with the positive x-axis, as in Figure 7.7.2. Imagine $\overrightarrow{OP}$ swings around the "pole" at the origin.

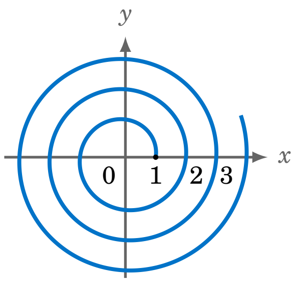


Figure 7.7.1 Spiral

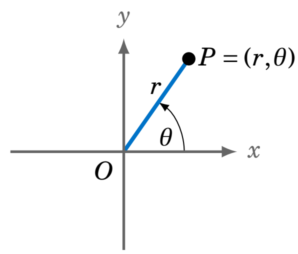


Figure 7.7.2 Polar coordinates (r,θ)

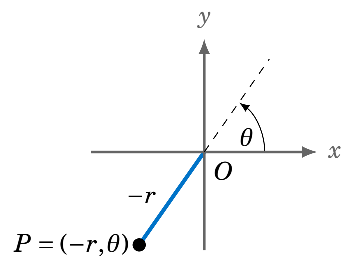


Figure 7.7.3 Negative r: $(-r,\theta)$

The pair (r,θ) contains the polar coordinates of P, and the positive x-axis is called the **polar axis** of this coordinate system. For the angle θ measured in radians, $(r,\theta) = (r,\theta+2\pi k)$ for all integers k. Thus, the polar coordinates of a point are not unique. By convention r can be negative, by defining $(-r,\theta) = (r,\theta+\pi)$ for any angle θ: the ray $\overrightarrow{OP}$ is drawn in the opposite direction from the angle θ, as in Figure 7.7.3. When $r = 0$, the point $(r,\theta) = (0,\theta)$ is the origin O, regardless of the value of θ.

Example 7.18

Express the spiral in Figure 7.7.1 in polar coordinates, such that the distance between any two points separated by 2π radians is always 1.

Solution: The goal is to find some equation involving r and θ (measured in radians) that describes the spiral. The distance between any two points separated by 2π radians is always 1, for example:

$$\begin{aligned} \theta &= 0 &\Rightarrow\quad r &= 1\\ \theta &= 2\pi &\Rightarrow\quad r &= 2\\ \theta &= 4\pi &\Rightarrow\quad r &= 3\\ &&\cdots&\\ \theta &= 2\pi k &\Rightarrow\quad r &= 1+k \end{aligned}$$

for all integers $k \ge 0$. In fact $r = 1+k$ when $\theta = 2\pi k$ for all real $k \ge 0$, by the assumption about the distance. So solving for k in terms of θ yields the polar equation $r = 1 + \frac{\theta}{2\pi}$ for all $\theta \ge 0$.

[16]Created by the Flemish mathematician Grégoire de Saint-Vincent (1584-1667) and Italian mathematician Bonaventura Cavalieri (1598-1647) in the 17$^{\text{th}}$ century, later used by Newton in his *Method of Fluxions* (1671).

You might be familiar with graphing paper, for plotting points or functions given in Cartesian coordinates. Such paper consists of a rectangular grid, where the horizontal and vertical lines represent where x and y, respectively, are constants, at regular intervals. Similar graphing paper exists for polar coordinates, as in Figure 7.7.4.

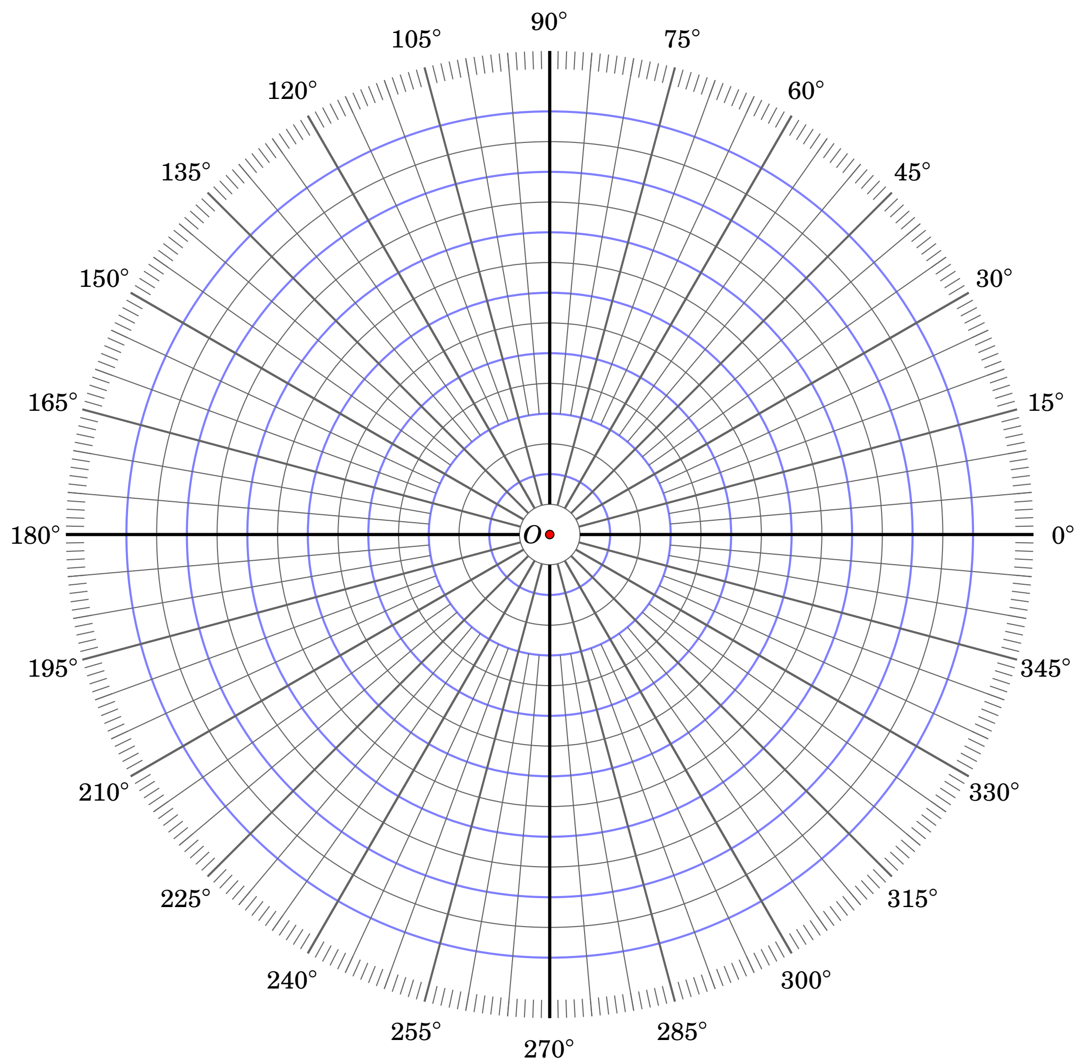


Figure 7.7.4 Polar coordinate graphing paper

This polar grid is radial, not rectangular. The concentric circles around the origin O are where r is constant (e.g. $r = 1$, $r = 2$), while the lines through the origin are where θ is constant, at regular intervals for each. The angle θ shown here is in degrees, though radians are often preferred for their "unitless" nature.

In general, polar coordinates are useful in describing plane curves that exhibit symmetry about the origin (though there are other situations), which arise in many physical applications.

Figure 7.7.5 shows how to convert between polar coordinates and Cartesian coordinates. Namely, for a point with polar coordinates (r,θ) and Cartesian coordinates (x,y):

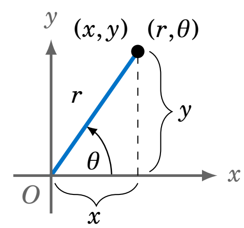


Figure 7.7.5

Polar to Cartesian:

$$x = r\cos\theta \qquad y = r\sin\theta \tag{7.17}$$

Cartesian to Polar:

$$r = \pm\sqrt{x^2 + y^2} \qquad \tan\theta = \frac{y}{x} \text{ if } x \neq 0 \tag{7.18}$$

In formula (7.18), if $x = 0$ then $\theta = \pi/2$ or $\theta = 3\pi/2$. If $x \neq 0$ and $y \neq 0$ then the two possible solutions for θ in the equation $\tan\theta = \frac{y}{x}$ are in opposite quadrants (for $0 \le \theta < 2\pi$). If the angle θ is in the same quadrant as the point (x,y), then $r = \sqrt{x^2+y^2}$ (i.e. r is positive); otherwise $r = -\sqrt{x^2+y^2}$ (i.e. r is negative).

Example 7.19

Write the equation of the unit circle $x^2 + y^2 = 1$ in polar coordinates.

Solution: By formula (7.18), $r^2 = x^2 + y^2 = 1$, so in polar coordinates the equation is simply $r = 1$. In general the circle $x^2 + y^2 = a^2$ of radius $a > 0$ has the simpler expression $r = a$ in polar coordinates.

Example 7.20

Write the equation $x^2 + (y-4)^2 = 16$ in polar coordinates.

Solution: This is the equation of a circle of radius 4 centered at the point $(0,4)$. Expand the equation:

$$\begin{aligned}
x^2 + (y-4)^2 &= 16 \\
x^2 + y^2 - 8y + 16 &= 16 \\
x^2 + y^2 &= 8y \\
r^2 &= 8r\sin\theta \quad \text{(by formulas (7.17) and (7.18))} \\
r &= 8\sin\theta
\end{aligned}$$

Is it valid to cancel r from both sides in the last step? Yes. The point $(0,0)$ is on the circle, so canceling r does not eliminate $r = 0$ as a potential solution of the equation (e.g. $\theta = 0$ would make $r = 8\sin\theta = 0$). Thus, the polar equation is $r = 8\sin\theta$.

Notice that this polar equation is actually less intuitive than its Cartesian equivalent. From the equation $x^2 + (y-4)^2 = 16$ it is easy to identify the curve as a circle and read off its radius and center; these properties are not so obvious from the polar equation. Since the circle's center is not the origin, there is no symmetry about the origin, which is when polar coordinates are often better suited.

Derivatives in Polar Coordinates

Suppose that the polar coordinates (r,θ) for a plane curve are related by a function: $r = r(\theta)$. Then by formula (7.17), $x = r(\theta)\cos\theta$ and $y = r(\theta)\sin\theta$ are now parametric equations for the curve in the parameter θ. Thus, by the Product Rule and formulas (7.14) and (7.15) from Section 7.6, with $dx = x'(\theta)\,d\theta$ and $dy = y'(\theta)\,d\theta$:

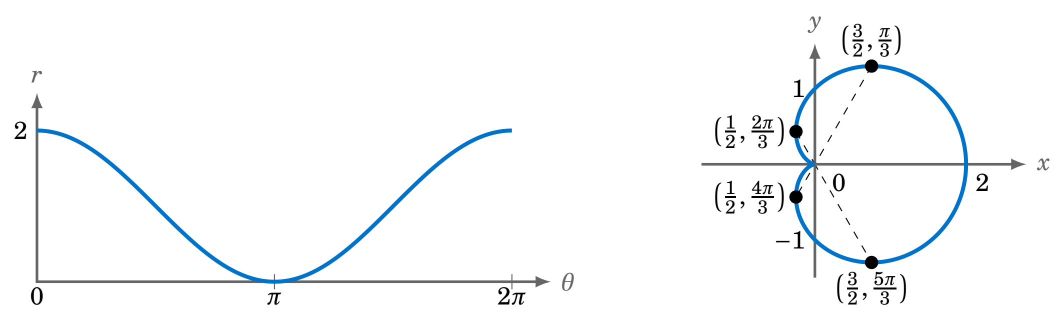
For a plane curve with polar equation $r = r(\theta)$,

$$\frac{dy}{dx} = \frac{r'(\theta)\sin\theta + r(\theta)\cos\theta}{r'(\theta)\cos\theta - r(\theta)\sin\theta} \quad\text{and}\quad \frac{d^2y}{dx^2} = \frac{\dfrac{d}{d\theta}\left(\dfrac{dy}{dx}\right)}{r'(\theta)\cos\theta - r(\theta)\sin\theta}\,. \tag{7.19}$$

Example 7.21

Sketch the graph of $r = 1+\cos\theta$.

Solution: First sketch the graph treating (r,θ) as Cartesian coordinates, for $0 \le \theta \le 2\pi$ as in Figure 7.7.6(a). Then use that graph to trace out a rough graph in polar coordinates, as in Figure 7.7.6(b).[17]



(a) θr-plane **(b)** Polar coordinates (r,θ)

Figure 7.7.6 Graph of $r = 1+\cos\theta$

To find the maxima, minima, and inflection points it is still necessary to find $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$. It is left as an exercise to use formula (7.19) and double-angle identities to show that

$$\frac{dy}{dx} = -\frac{\cos\theta+\cos 2\theta}{\sin\theta+\sin 2\theta} \quad\text{and}\quad \frac{d^2y}{dx^2} = -\frac{3(1+\cos\theta)}{(\sin\theta+\sin 2\theta)^3}$$

and that for θ in $[0,2\pi]$, $\frac{dy}{dx} = 0$ only when $\theta = \frac{\pi}{3}$ and $\frac{5\pi}{3}$, while $\frac{dy}{dx}$ is undefined for $\theta = 0$, $\frac{2\pi}{3}$, π, $\frac{4\pi}{3}$, and 2π. Since $\left.\frac{d^2y}{dx^2}\right|_{\theta=\frac{\pi}{3}} < 0$ and $\left.\frac{d^2y}{dx^2}\right|_{\theta=\frac{5\pi}{3}} > 0$ then the curve has a local maximum when $\theta = \frac{\pi}{3}$ and a local minimum when $\theta = \frac{5\pi}{3}$. It can also be shown that $\frac{d^2y}{dx^2}$ changes sign around $\theta = 0$, $\frac{2\pi}{3}$, π, $\frac{4\pi}{3}$, so that the inflection points occur at those values of θ, as shown (with the correct concavity) in Figure 7.7.6(b).

[17]There are far too many interesting plane curves to cover here. For an extensive collection, see Lawrence, J.D., *A Catalog of Special Plane Curves*, New York: Dover Publications, Inc., 1972. See also Seggern, D.H. von, *CRC Handbook of Mathematical Curves and Surfaces*, Boca Raton, FL: CRC Press, Inc., 1990.

Integration in Polar Coordinates

In some cases polar coordinates can simplify evaluation of a definite integral or finding an area. To determine the polar form of a definite integral, suppose that r is a function of θ: $r = f(\theta)$. A polar region swept out by $r = f(\theta)$ between $\theta = \alpha$ and $\theta = \beta$ would look like the shaded region in Figure 7.7.7 with area A:

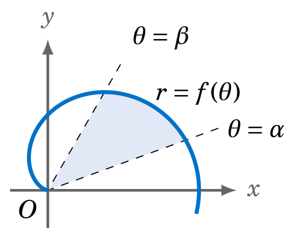


Figure 7.7.7 Area A

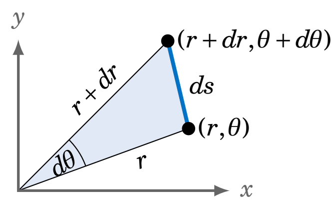


Figure 7.7.8 Area dA

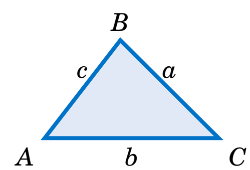


Figure 7.7.9 Area $= \frac{1}{2}bc\sin A$

A typical infinitesimal wedge of that region—shown in Figure 7.7.8—is produced by an infinitesimal change $d\theta$ in the angle θ, which results in an infinitesimal change dr in r. By the Microstraightness Property, the curve $r = f(\theta)$ is a straight line with infinitesimal length ds over that infinitesimal angle $d\theta$. The area dA of the wedge thus equals the area of the triangle with sides of lengths r and $r + dr$ with included angle $d\theta$ and a side of length ds. Since $\sin d\theta = d\theta$ and $dr = f'(\theta)\,d\theta$, then by the formula for the area of a triangle (as shown in Figure 7.7.9),[18]

$$\begin{aligned} dA &= \tfrac{1}{2}(r)(r+dr)\sin d\theta \\ &= \tfrac{1}{2}r^2 d\theta \;+\; \tfrac{1}{2}r\,(dr)\,d\theta \\ &= \tfrac{1}{2}r^2 d\theta \;+\; \tfrac{1}{2}r\,(f'(\theta)\,d\theta)\,d\theta \\ &= \tfrac{1}{2}r^2 d\theta \;+\; 0 \;=\; \tfrac{1}{2}r^2 d\theta \end{aligned}$$

since $(d\theta)^2 = 0$. The area A of the region is the sum of these infinitesimal areas dA:

For a plane curve with polar equation $r = f(\theta)$, the area A of the region swept out between $\theta = \alpha$ and $\theta = \beta$ is:

$$A \;=\; \int_{\theta=\alpha}^{\theta=\beta} dA \;=\; \int_{\alpha}^{\beta} \tfrac{1}{2}r^2\,d\theta \;=\; \int_{\alpha}^{\beta} \tfrac{1}{2}(f(\theta))^2\,d\theta \qquad (7.20)$$

If f is periodic then choose the angle interval $[\alpha, \beta]$ so that the area is swept out only once.

[18] The formula $\frac{1}{2}bc\sin A$ for the area of a triangle $\triangle ABC$ is derived in most trigonometry texts. For example, see p.54 in Corral, M., *Trigonometry*, `http://mecmath.net/trig/`, 2009.

Example 7.22

Use polar coordinates to show that the area of a circle of radius R is πR^2.

Solution: Let the origin be the center of the circle. Then $r = R$ is the polar equation of the circle, with $0 \le \theta \le 2\pi$ sweeping out exactly one full circle. The area A inside the circle is then

$$A = \int_0^{2\pi} \tfrac{1}{2} r^2 \, d\theta = \int_0^{2\pi} \tfrac{1}{2} R^2 \, d\theta = \tfrac{1}{2} R^2 \theta \,\Big|_0^{2\pi} = \tfrac{1}{2} R^2 (2\pi - 0) = \pi R^2 \ . \quad \checkmark$$

Note the simplicity of this integral compared to the trigonometric substitution required when using Cartesian coordinates, as in Section 6.3. Notice also that using a larger interval—say $[0,4\pi]$—for θ would result in an incorrect area ($4\pi R^2$) even though the region inside the curve is the same.

Example 7.23

Find the area inside the curve $r = 1 + \cos\theta$.

Solution: Choose $0 \le \theta \le 2\pi$ as in Example 7.21, so that the area A is:

$$\begin{aligned} A &= \int_0^{2\pi} \tfrac{1}{2} r^2 \, d\theta = \int_0^{2\pi} \tfrac{1}{2}(1+\cos\theta)^2 \, d\theta = \int_0^{2\pi} \left(\frac{1}{2} + \cos\theta + \frac{\cos^2\theta}{2} \right) d\theta \\ &= \int_0^{2\pi} \left(\frac{1}{2} + \cos\theta + \frac{1+\cos 2\theta}{4} \right) d\theta = \int_0^{2\pi} \left(\frac{3}{4} + \cos\theta + \frac{\cos 2\theta}{4} \right) d\theta \\ &= \frac{3}{4}\theta + \sin\theta + \frac{1}{8}\sin 2\theta \,\Big|_0^{2\pi} = \frac{3\pi}{2} \end{aligned}$$

Exercises

A

For Exercises 1-8 write the given equation in polar coordinates.

1. $(x-3)^2 + y^2 = 9$ **2.** $y = x$ **3.** $x^2 - y^2 = 1$ **4.** $3x^2 + 4y^2 - 6x = 9$

5. $y = -x$ **6.** $y = x + 1$ **7.** $y = x^2$ **8.** $y = x^3$

9. Write the polar equation $r^2 = 4\cos 2\theta$ in Cartesian coordinates.

10. Find the tangent line to $r = \cos 2\theta$ at $\left(\frac{1}{2}, \frac{\pi}{6}\right)$. **11.** Find the tangent line to $r = 8\sin^2\theta$ at $\left(2, \frac{5\pi}{6}\right)$.

12. Recall the curve $r = 1 + \cos\theta$ from Example 7.21.

(a) Verify that for this curve,

$$\frac{dy}{dx} = -\frac{\cos\theta + \cos 2\theta}{\sin\theta + \sin 2\theta} \quad \text{and} \quad \frac{d^2y}{dx^2} = -\frac{3(1+\cos\theta)}{(\sin\theta + \sin 2\theta)^3} \ .$$

(b) Verify that for θ in $[0,2\pi]$, $\frac{dy}{dx} = 0$ only when $\theta = \frac{\pi}{3}$ and $\frac{5\pi}{3}$, and is undefined for $\theta = 0$, $\frac{2\pi}{3}$, π, $\frac{4\pi}{3}$, and 2π.

(c) Verify that $\left.\frac{d^2y}{dx^2}\right|_{\theta=\frac{\pi}{3}} < 0$ and $\left.\frac{d^2y}{dx^2}\right|_{\theta=\frac{5\pi}{3}} > 0$.

(d) Verify that $\frac{d^2y}{dx^2}$ changes sign around $\theta = 0$, $\frac{2\pi}{3}$, π, and $\frac{4\pi}{3}$.

For Exercises 13-15, sketch the graph of the given curve and indicate all local maxima and minima.

13. $r = 1 + \sin\theta$ **14.** $r = 1 - \cos\theta$ **15.** $r = \sin 2\theta$

16. Find the area inside $r = 1 + \sin\theta$. **17.** Find the area inside $r = \sin 2\theta$.

B

18. Sketch a rough graph of the meridian voltage component E_θ for a half-wavelength linear antenna:

$$E_\theta = r(\theta) = \frac{\cos(\frac{\pi}{2}\cos\theta)}{\sin\theta}$$

19. Show that the distance d between two points (r_1,θ_1) and (r_2,θ_2) in polar coordinates is

$$d = \sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos(\theta_1 - \theta_2)}\ .$$

20. For a point $P = (r,\theta)$ on the curve $r = r(\theta)$, let α be the angle that the tangent line through P makes with the positive x-axis. Let $\psi = \alpha - \theta$ be the angle between the tangent line and the line through P and the origin. Show that $\tan\psi = \frac{r(\theta)}{r'(\theta)}$. *(Hint: Use formulas (7.19) and (3.2).)*

21. For the parabola with focus at $(0,0)$, vertex at $(0,-p)$, and directrix $y = -2p$ (with $p > 0$), show that the polar equation is

$$r = \frac{2p}{1 - \sin\theta}\ .$$

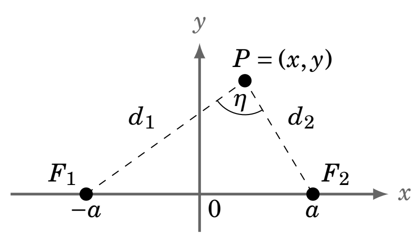
22. For the ellipse with foci $(0,0)$ and $(2c,0)$, vertexes $(c \pm a, 0)$, and eccentricity $e = \frac{c}{a}$ (with $0 < c < a$), show that the polar equation is

$$r = \frac{a(1 - e^2)}{1 - e\cos\theta}\ .$$

23. For the hyperbola with foci $(0,0)$ and $(2c,0)$, vertexes $(c \pm a, 0)$, and eccentricity $e = \frac{c}{a}$ (with $0 < a < c$), show that the polar equation is

$$r = \frac{a(e^2 - 1)}{1 + e\cos\theta}\ .$$

24. The *bipolar coordinates*[19] (ξ,η) of a point $P = (x,y)$ relative to two poles $F_1 = (-a,0)$ and $F_2 = (a,0)$ are given by $\xi = \ln\frac{d_1}{d_2}$ and $\eta = \angle F_1PF_2$, where $d_1 = F_1P$ and $d_2 = F_2P$, as in the figure on the right . Then $-\infty < \xi < \infty$ except at F_1 (which corresponds to $\xi = -\infty$) and F_2 ($\xi = \infty$), while $0 \le \eta \le \pi$ for points on or above the x-axis, and $\pi < \eta < 2\pi$ below the x-axis.



(a) Show that

$$x = \frac{a\sinh\xi}{\cosh\xi - \cos\eta} \quad\text{and}\quad y = \frac{a\sin\eta}{\cosh\xi - \cos\eta}\ .$$

(Hint: Use the distance formula, the Law of Cosines, and the Law of Sines.)

(b) Show that for constants $\tau \ne 0$ and $\sigma \ne 0$ or π, the curves $\xi = \tau$ and $\eta = \sigma$ are the respective circles

$$(x - a\coth\tau)^2 + y^2 = \frac{a^2}{\sinh^2\tau} \quad\text{and}\quad x^2 + (y - a\cot\sigma)^2 = \frac{a^2}{\sin^2\sigma}\ .$$

---

[19] Inspired by the lines of force and equipotential lines for an electric dipole. See pp.55-56 in STRATTON, J.A., *Electromagnetic Theory*, New York: McGraw-Hill Book Company, Inc., 1941.

CHAPTER 8

Applications of Integrals

8.1 Area Between Curves

The "area under a curve" was defined in Chapter 5 as the area below some curve $y = f(x)$ and above the x-axis over some interval. That was a special case of the **area between curves**, where in general one curve $y = f_1(x)$ is not necessarily always above another curve $y = f_2(x)$ over the entire interval, as in Figure 8.1.1 for an interval $[a, b]$.

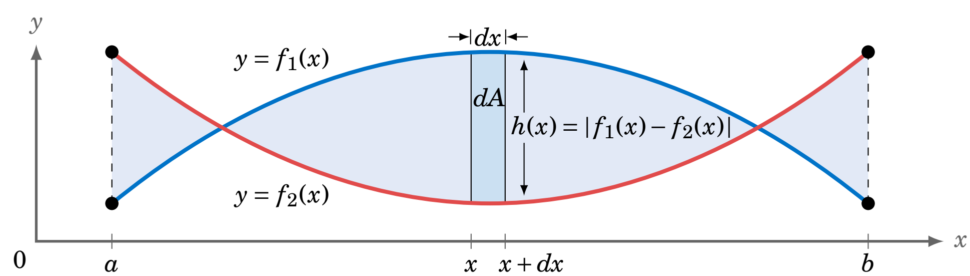


Figure 8.1.1 The area A between two curves $y = f_1(x)$ and $y = f_2(x)$ over $[a, b]$

The area A of the region between the curves in Figure 8.1.1 cannot be negative. Thus, a typical infinitesimal area element dA of the region is of the form $h(x)\,dx$, where the **height function** $h(x)$ is the nonnegative difference in the y-coordinates of the curves at each x in $[a, b]$: $h(x) = \left|f_1(x) - f_2(x)\right|$. Hence:

The area A between two curves $y = f_1(x)$ and $y = f_2(x)$ over an interval $[a, b]$ is:

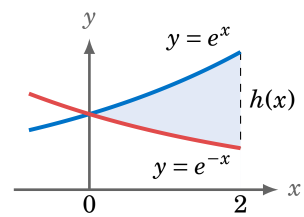
$$A = \int_a^b dA = \int_a^b \left|f_1(x) - f_2(x)\right| dx \tag{8.1}$$

The interval $[a, b]$ can be replaced by any interval—finite or infinite—over which the integral is defined. Neither curve is required to be above the x-axis.

Example 8.1

Find the area between $y = e^x$ and $y = e^{-x}$ over $[0,2]$.

Solution: Since $e^x \ge e^{-x}$ for x in $[0,2]$, the height function $h(x)$ for the region between the curves over $[0,2]$ is $h(x) = \left|e^x - e^{-x}\right| = e^x - e^{-x}$. The area A of the region is thus

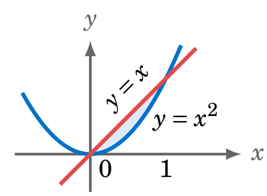


$$\begin{aligned} A &= \int_0^2 (e^x - e^{-x})\,dx \\ &= e^x + e^{-x}\,\Big|_0^2 = e^2 + e^{-2} - (1+1) \\ &= 2(\cosh 2 - 1)\,. \end{aligned}$$

Example 8.2

Find the area of the region bounded by $y = x^2$ and $y = x$.

Solution: A "bounded" region will always mean a region of finite area, as opposed to unbounded regions. The curves $y = x^2$ and $y = x$ intersect at $x = 0$ and $x = 1$, so the region the curves bound is the shaded region shown in the figure on the right. Since $x \ge x^2$ for $0 \le x \le 1$, then the height function for the region is $h(x) = |x^2 - x| = x - x^2$. The region's area A is then

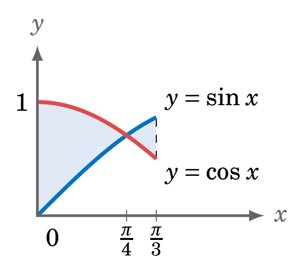


$$\begin{aligned} A &= \int_0^1 (x - x^2)\,dx \\ &= \frac{1}{2}x - \frac{1}{3}x^2\,\Big|_0^1 = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}\,. \end{aligned}$$

Example 8.3

Find the area of the region bounded by $y = \sin x$ and $y = \cos x$ over $[0, \pi/3]$.

Solution: As shown in the figure on the right, the curves intersect at $x = \frac{\pi}{4}$, and $\cos x \ge \sin x$ for $0 \le x \le \frac{\pi}{4}$, while $\sin x \ge \cos x$ for $\frac{\pi}{4} \le x \le \frac{\pi}{3}$. The area A of the region thus needs to be split into two integrals:



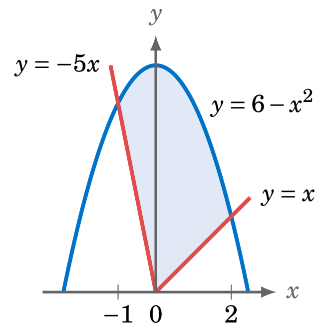
$$\begin{aligned} A &= \int_0^{\pi/3} |\sin x - \cos x|\,dx \\ &= \int_0^{\pi/4} (\cos x - \sin x)\,dx + \int_{\pi/4}^{\pi/3} (\sin x - \cos x)\,dx \\ &= \left(\sin x + \cos x\,\Big|_0^{\pi/4}\right) + \left(-\cos x - \sin x\,\Big|_{\pi/4}^{\pi/3}\right) \\ &= \left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\right) - (0-1) + \left(-\frac{1}{2} - \frac{\sqrt{3}}{2}\right) - \left(-\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}\right) \\ &= \frac{4\sqrt{2} - 3 - \sqrt{3}}{2} \end{aligned}$$

Formula (8.1) can be extended to find the area between any number of curves, by splitting the integral over subintervals with different height functions.

Example 8.4

Find the area of the region bounded by $y = 6 - x^2$, $y = x$ and $y = -5x$ above the x-axis.

Solution: As shown in the figure on the right, above the x-axis the curve $y = 6 - x^2$ intersects the line $y = x$ at $x = 2$ and intersects the line $y = -5x$ at $x = -1$. Since $6 - x^2 \ge -5x$ over $[-1,0]$ and $6 - x^2 \ge x$ over $[0,2]$, the area A of the shaded region needs to be split into two integrals:



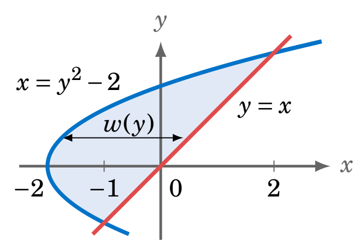
$$\begin{aligned} A &= \int_{-1}^{0} \left|6 - x^2 - (-5x)\right| \,dx \;+\; \int_{0}^{2} \left|6 - x^2 - x\right| \,dx \\ &= \int_{-1}^{0} (6 - x^2 + 5x) \,dx \;+\; \int_{0}^{2} (6 - x^2 - x) \,dx \\ &= \left(6x - \frac{1}{3}x^3 + \frac{5}{2}x^2 \,\Bigg|_{-1}^{0} \right) + \left(6x - \frac{1}{3}x^3 - \frac{1}{2}x^2 \,\Bigg|_{0}^{2} \right) \\ &= 0 - \left(-6 + \frac{1}{3} + \frac{5}{2} \right) + \left(12 - \frac{8}{3} - 2 \right) - 0 = \frac{21}{2} \end{aligned}$$

For some areas between curves it might be easier to switch the roles of x and y, so that instead of a vertical height function you would use a horizontal width function.

Example 8.5

Find the area of the region bounded by $x = y^2 - 2$ and $y = x$.

Solution: As shown in the figure on the right, the parabola $x = y^2 - 2$ intersects the line $y = x$ at $x = -1$ and $x = 2$. The region has different height functions $h(x)$ for $-2 \le x \le -1$ and $-1 \le x \le 2$, so that two integrals would be required for the area A. However, notice that the width function $w(y)$ has one definition over the entire region between the curves $x = y^2 - 2$ and $x = y$: $w(y) = \left|y - (y^2 - 2)\right| = y - (y^2 - 2)$. Thus, instead of integrating the vertical strips $dA = h(x)\,dx$ along the x-axis, integrate the horizontal strips $dA = w(y)\,dy$ along the y-axis, from $y = -1$ to $y = 2$:



$$\begin{aligned} A &= \int_{-1}^{2} w(y) \,dy \;=\; \int_{-1}^{2} \left|y - (y^2 - 2)\right| \,dy \\ &= \int_{-1}^{2} (y - (y^2 - 2)) \,dy \\ &= \frac{1}{2}y^2 - \frac{1}{3}y^3 + 2y \,\Bigg|_{-1}^{2} \\ &= \left(2 - \frac{8}{3} + 4 \right) - \left(\frac{1}{2} + \frac{1}{3} - 2 \right) = \frac{9}{2} \end{aligned}$$

The area between curves given by polar equations can be found similarly. For example, consider curves $r = r_1(\theta)$ and $r = r_2(\theta)$ with $r_1(\theta) \ge r_2(\theta)$ when $\alpha \le \theta \le \beta$ as in Figure 8.1.2. The area A of the region between the curves and those angles is simply the difference between the "outer" and "inner" areas, each given by formula (7.20):

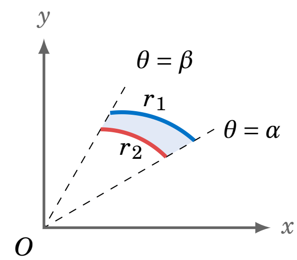


Figure 8.1.2

$$A = \int_\alpha^\beta \tfrac{1}{2} r_1^2 d\theta - \int_\alpha^\beta \tfrac{1}{2} r_2^2 d\theta = \int_\alpha^\beta \tfrac{1}{2}(r_1^2 - r_2^2)\, d\theta$$

In general, to include cases where the "outer" and "inner" curves switch positions, take the absolute value of the difference:

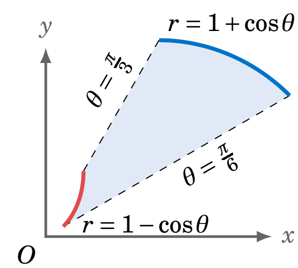
The area A between two polar curves $r = r_1(\theta)$ and $r = r_2(\theta)$ for $\alpha \le \theta \le \beta$ is:

$$A = \int_\alpha^\beta \tfrac{1}{2}\left|r_1^2 - r_2^2\right|\, d\theta \tag{8.2}$$

Example 8.6

Find the area between $r = 1 + \cos\theta$ and $r = 1 - \cos\theta$ for $\frac{\pi}{6} \le \theta \le \frac{\pi}{3}$.

Solution: Let $r_1(\theta) = 1 + \cos\theta$ and $r_2(\theta) = 1 - \cos\theta$. Since $r_1(\theta) > r_2(\theta)$ for $\frac{\pi}{6} \le \theta \le \frac{\pi}{3}$, the area A of the region (shown in the figure on the right) is



$$A = \int_{\pi/6}^{\pi/3} \tfrac{1}{2}\left|r_1^2 - r_2^2\right|\, d\theta = \int_{\pi/6}^{\pi/3} \tfrac{1}{2}\left((1+\cos\theta)^2 - (1-\cos\theta)^2\right)\, d\theta$$

$$= \int_{\pi/6}^{\pi/3} 2\cos\theta\, d\theta = 2\sin\theta \,\Big|_{\pi/6}^{\pi/3} = 2\left(\frac{\sqrt{3}}{2} - \frac{1}{2}\right) = \sqrt{3} - 1$$

Monte Carlo integration is a technique for approximating the area of a region by taking a large number of random points in a rectangle that encloses the region (see Figure 8.1.3). The idea is simple:

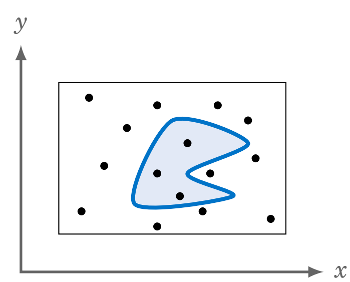


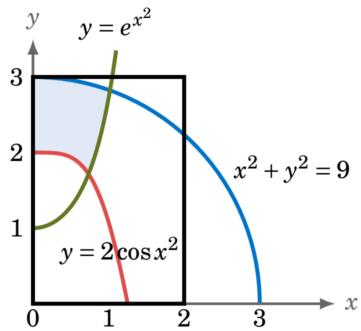
Figure 8.1.3

$$\frac{\text{\# of points in the region}}{\text{\# of points in the rectangle}} \approx \frac{\text{area of the region}}{\text{area of the rectangle}}$$

For example, if 20% of the random points in the rectangle fall inside the region, then—by randomness—you would expect the area of the region to be about 20% of the area of the rectangle. The more random points you take, the better the approximation. Since the area of the rectangle is known, as well as the number of random points inside the region and the rectangle, the area of the region is easy to approximate.

Example 8.7

Use Monte Carlo integration to approximate the area of the region in the first quadrant above the curves $y = e^{x^2}$ and $y = 2\cos x^2$, and inside the circle $x^2 + y^2 = 9$.



Solution: The region is the shaded area shown in the figure on the right, enclosed in a rectangle of width 2 and height 3. The area of the rectangle is 6, and a point (x, y) in the rectangle is inside the region if only if the following conditions are met:

$$y > e^{x^2} \quad \text{and} \quad y > 2\cos x^2 \quad \text{and} \quad x^2 + y^2 < 9$$

Notice that these conditions are all in the form of inequalities. The Monte Carlo integration is then simple to perform in Octave, using 10 million random points:

```
octave> N = 1e7;
octave> x = 2*rand(1,N);
octave> y = 3*rand(1,N);
octave> 6*(sum(y > exp(x.^2) & y > 2*cos(x.^2) & x.^2 + y.^2 < 9))/N
ans = 0.94612
```

The actual value—accurate to 5 decimal places—is 0.94606.

The `rand(1,N)` command returns an array of `N` random numbers between 0 and 1. So the statement `x = 2*rand(1,N)` stores `N` random numbers between 0 and 2 in an array for the x-coordinates, and the statement `y = 3*rand(1,N)` stores `N` random numbers between 0 and 3 in an array for the y-coordinates. The statement `y > exp(x.^2)` returns a value of 1 if the condition $y > e^{x^2}$ is met, 0 otherwise. Similarly for the statements `y > 2*cos(x.^2)` and `x.^2 + y.^2 < 9`. Joining those three statements with `&` symbols returns a value of 1 if all three conditions are met, 0 otherwise. The `sum` command thus counts how many of the `N` points are inside the region. Dividing that count by `N` gives the ratio of points inside the region, then multiplying by 6 (the area of the rectangle) gives the approximate area of the region.

Note that the size of the rectangle can affect the approximation—generally the larger the rectangle the more points must be used. Note also in this example that finding the area by using definite integrals would require numerical integration methods, since $f(x) = e^{x^2}$ and $f(x) = 2\cos x^2$ cannot be integrated in a closed form. In fact, even finding the points of intersection of the three curves—in order to split the integrals—would require a numerical root-finding method (e.g. Newton's method).

Exercises

A

For Exercises 1-6, find the area of the region bounded by the given curves.

1. $y = x^2$ and $y = 2x + 3$ **2.** $x = -y^2 + 2y$ and $x = 0$ **3.** $y = x^2 - 1$ and $y = x^3 - 1$

4. $y = x^4$ and $y = x$ **5.** $x = y^2$ and $x = y + 2$ **6.** $y = 4 - 4x^2$ and $y = 1 - x^2$

7. Find the area between $y = 4x - x^2$ and $y = x$ over $[0, 4]$.

8. Find the area between $y = \cosh x$ and $y = \sinh x$ over $[0, \infty)$.

9. Find the area of the region defined by the inequalities $0 \le x \le y - x \le 1 - y \le 1$.

10. Find the area between $r = 1 + \cos\theta$ and $r = 2 + 2\cos\theta$.

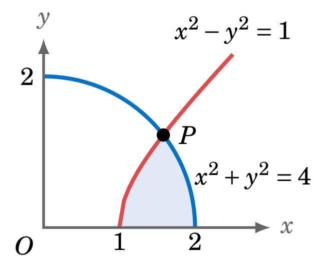
11. Find the area between $r = 1 + \cos\theta$ and $r = 2 + \cos\theta$.

B

12. Find the area of the region in the first quadrant between the unit hyperbola $x^2 - y^2 = 1$ and the circle $x^2 + y^2 = 4$ (i.e. the shaded region shown in the figure on the right) in two ways:

(a) Integration using formula (8.1).

(b) Draw a line segment from the origin O to the point of intersection P on the hyperbola, then use the areas of the resulting circular sector and hyperbolic sector, without resorting to integration. Does your answer agree with part (a)? Explain.



13. Find the area common to the four circles of radius 5 shown in Figure 8.1.4. *(Hint: Use symmetry.)*

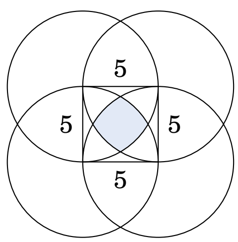
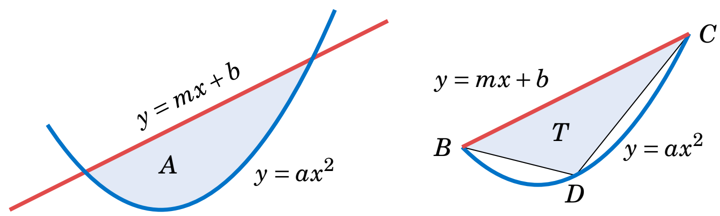


Figure 8.1.4 Exercise 13



(a) Area A **(b)** Area T

Figure 8.1.5 Exercise 14

14. Let A be the area of the region bounded by the parabola $y = ax^2$ and the line $y = mx + b$, where a, m, and b are positive constants (see Figure 8.1.5(a)). Let T be the area of the triangle $\triangle BCD$, where B and C are where the line intersects the parabola, and the point D on the parabola has the same x-coordinate as the midpoint of the line segment $\overline{BC}$ (see Figure 8.1.5(b)). Show that $A = \frac{4}{3}T$.

15. In Example 8.7 the region was enclosed in the rectangle $R = \{(x,y) : 0 \le x \le 2\ ,\ 0 \le y \le 3\}$. Use Monte Carlo integration to approximate the area of the region again, using a different enclosing rectangle R:

(a) $R = \{(x,y) : 0 \le x \le 2\ ,\ 1 \le y \le 3\}$

(b) $R = \{(x,y) : 0 \le x \le 3\ ,\ 0 \le y \le 3\}$

Are the results significantly different than before?

16. Approximate the area of the region bounded by $y = x^2$ and $y = \cos x$ in two different ways:

(a) Use Monte Carlo integration with 10 million points.

(b) Use a numerical root-finding method to find the points of intersection of the curves, then use those points in a numerical integration method to find the area.

C

17. A dog is chained to a fixed point at the circular base of a cylindrical silo. The silo's radius is 50 ft and the chain can wrap exactly halfway around the silo. How much total area can the dog roam, not counting the area inside the silo?

8.2 Average Value of a Function

According to Kepler's laws of planetary motion, a planet revolving around the Sun follows an elliptical orbit, with the Sun at one focus of the ellipse, as in Figure 8.2.1. The distance d between the planet and the Sun varies over the ellipse, reaching a minimum distance and a maximum distance (by the Extreme Value Theorem). How would you find the *average* distance between the planet and the Sun over one complete orbit? The idea is to generalize the usual notion of an average of numbers.

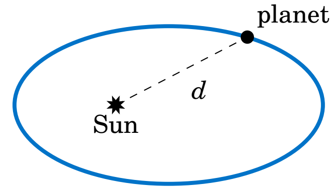


Figure 8.2.1

Recall that for n numbers $x_1, x_2, \ldots, x_n$ the *average*, denoted by $\bar{x}$, is simply the sum of the numbers divided by how many numbers there are, namely

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

In statistics $\bar{x}$ is called the *mean* of $x_1, x_2, \ldots, x_n$. This definition makes sense for a finite set of numbers, but in the case of a planet revolving around the Sun, there are an uncountably infinite number of distances between the planet and the Sun, making the above definition impossible to use. A way of taking a sum over an infinite continuum of values is needed instead. Such a method has already been encountered: the definite integral, which is merely a sum of a continuum of infinitesimal quantities.

To motivate the definition of the average value of a function f over a closed interval $[a,b]$, denoted by $\langle f \rangle$, consider a partition

$$P = \{a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b\}$$

that divides $[a,b]$ into n subintervals $[x_{i-1}, x_i]$ of equal length $\Delta x_i = x_i - x_{i-1} = (b-a)/n$, as in Figure 8.2.2. The n function values $f(x_1)$, $f(x_2)$, ... , $f(x_n)$ constitute only a finite subset of all the function values $f(x)$ over $[a,b]$, so their average would be an approximation of the true function average $\langle f \rangle$, namely:

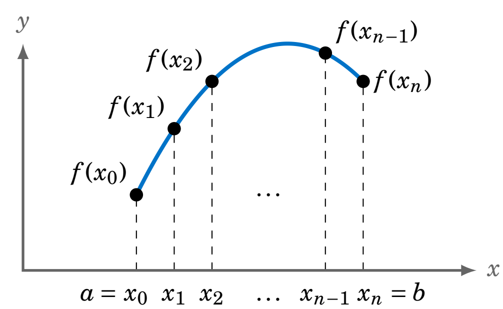


Figure 8.2.2 $\Delta x_i = x_i - x_{i-1} = (b-a)/n$

$$\langle f \rangle \approx \frac{f(x_1) + f(x_2) + \cdots + f(x_n)}{n} = \sum_{i=1}^{n} \frac{f(x_i)}{n}$$

By properties of summations, divide the entire sum by the constant $b-a$ and multiply each term in the sum by $b-a$ to get:

$$\langle f \rangle \approx \frac{1}{b-a}\sum_{i=1}^{n} f(x_i)\cdot\frac{b-a}{n} = \frac{1}{b-a}\sum_{i=1}^{n} f(x_i)\,\Delta x_i$$

Note that the last summation on the right is just a Riemann sum for the definite integral $\int_a^b f(x)\,dx$, with the points x_i^* chosen to be the right endpoints of the intervals $[x_{i-1},x_i]$ for $i=1$ to n. Thus, taking the limit of that sum as $n\to\infty$ (which means including more and more function values in the average) yields the following definition:

The **average value** $\langle f \rangle$ of a function f over a closed interval $[a,b]$ is:[1]

$$\langle f \rangle = \frac{1}{b-a}\int_a^b f(x)\,dx \tag{8.3}$$

Example 8.8

Find the average value of $f(x)=x^2$ over $[0,1]$.

Solution: By definition, with $a=0$ and $b=1$,

$$\langle f \rangle = \frac{1}{1-0}\int_0^1 f(x)\,dx = \int_0^1 x^2\,dx = \left.\frac{x^3}{3}\right|_0^1 = \frac{1^3}{3} - \frac{0^3}{3} = \frac{1}{3}$$

Note that this says that if you took all the numbers between 0 and 1 and squared them, then the average of those squares would be 1/3.

Example 8.9

Find the average value of $f(x)=x^2$ over $[-1,1]$.

Solution: By definition, with $a=-1$ and $b=1$,

$$\langle f \rangle = \frac{1}{1-(-1)}\int_{-1}^1 f(x)\,dx = \frac{1}{2}\int_{-1}^1 x^2\,dx = \left.\frac{x^3}{6}\right|_{-1}^1 = \frac{1^3}{6} - \frac{(-1)^3}{6} = \frac{1}{3}$$

Note that this is the same as the average over $[0,1]$, as shown in the previous example. This should make sense, since the function $f(x)=x^2$ is symmetric about the y-axis, so the values of $f(x)$ from $[-1,0]$ are the same as those from $[0,1]$. The values from $[-1,1]$ just duplicate the values from $[0,1]$ and hence do not change the average.

Example 8.10

Find the average value of $f(x)=\sin x$ over $[0,\pi]$.

Solution: By definition, with $a=0$ and $b=\pi$,

$$\langle f \rangle = \frac{1}{\pi-0}\int_0^\pi f(x)\,dx = \frac{1}{\pi}\int_0^\pi \sin x\,dx = -\frac{1}{\pi}\cos x\,\Big|_0^\pi = -\frac{1}{\pi}(\cos\pi - \cos 0) = -\frac{1}{\pi}(-1-1) = \frac{2}{\pi}$$

[1] In some statistics or mathematics texts you might see the notation $\bar{f}$ for the average value.

Example 8.11

Find the average distance from the ellipse $\frac{x^2}{25}+\frac{y^2}{9}=1$ to the point $(4,0)$.

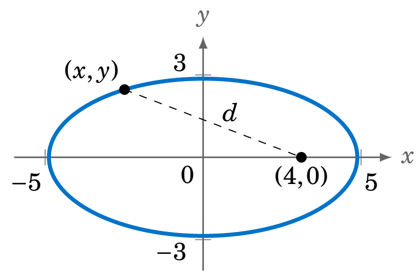


Figure 8.2.3 $\frac{x^2}{25}+\frac{y^2}{9}=1$

Solution: Let d represent the distance from any point (x,y) on the ellipse to the point $(4,0)$, as in Figure 8.2.3. If (x,y) is on the ellipse $\frac{x^2}{25}+\frac{y^2}{9}=1$ then $y^2=9(1-\frac{x^2}{25})=\frac{9}{25}(25-x^2)$. So by the distance formula, d is given by

$$\begin{aligned}
d^2 &= (x-4)^2 + (y-0)^2 = (x-4)^2 + y^2\\
&= (x-4)^2 + \frac{9}{25}(25-x^2)\\
&= \frac{25(x-4)^2 + 9(25-x)^2}{25} = \frac{25x^2 - 200x + 400 + 225 - 9x^2}{25}\\
&= \frac{16x^2 - 200x + 625}{25}\\
d^2 &= \frac{(4x-25)^2}{25} \quad \text{, and so taking square roots gives}\\
d &= \pm\frac{4x-25}{5} = -\frac{4x-25}{5} = \frac{25-4x}{5}
\end{aligned}$$

for $-5 \le x \le 5$, since $d=(4x-25)/5<0$ on $[-5,5]$ and the distance cannot be negative. Note that by symmetry of the ellipse about the x-axis, only the upper half of the ellipse is needed for the average distance, since the lower half just duplicates the distances. Hence, the average distance is

$$\langle d\rangle = \frac{1}{5-(-5)}\int_{-5}^{5}\frac{25-4x}{5}\,dx = \frac{1}{50}(25x-2x^2)\Big|_{-5}^{5} = \frac{1}{50}(125-50-(-125-50)) = 5\,.$$

Notice that the point $(4,0)$ is a focus of the ellipse $\frac{x^2}{25}+\frac{y^2}{9}=1$ (why?), which as it turns out makes the calculation of the average distance fairly simple.

What if you wanted the average value of a function f that is not easily integrable? One alternative to numerical integration techniques is the **Monte Carlo method**. The idea behind it is simple: go back to the usual definition of an average, by taking a large number N of random numbers $x_1, x_2, \ldots, x_N$ in $[a,b]$ and then using the approximation

$$\langle f\rangle \approx \frac{f(x_1)+f(x_2)+\cdots+f(x_N)}{N}\,.$$

This might seem like taking a step backward from calculus, and it is, but it is surprisingly useful, as well as simple to implement with a computer. In addition, it can be shown that as the number of random points in $[a,b]$ increases, the approximations converge to the actual average.

Example 8.12

The Monte Carlo method is easy to implement in Octave/MATLAB. Typically only a "one-liner" is needed, owing to Octave's *vectorization*—i.e. the ability to perform mathematical operations on entire arrays of objects all at once.

For example, recall from Example 8.8 that the average value of $f(x) = x^2$ over $[0,1]$ is $1/3 = 0.33333\ldots$. Approximate the average value using an array of 100 million (10^8) random numbers in $[0,1]$:

```
octave> mean(rand(1,1e8).^2)
ans = 0.3333292094741531
```

The dot in the `rand(1,1e8).^2` command applies the squaring operation (`^2`) to each of the 10^8 random numbers in the array returned by the `rand(1,1e8)` command. The aggregate `mean` function then calculates the array's mean. Trigonometric, exponential and other functions can be applied to arrays, with the function evaluating each array element individually. In general the command $(b-a)$`.*rand(1,N)+`a will return an array of `N` random numbers in the interval (a,b).

For example, the function $f(x) = \sin(x^2)$ cannot be integrated in a closed form, but its average value over $[\pi, 2\pi]$ can be approximated easily in Octave (actual average = -0.04154374531416104):

```
octave> mean(sin((pi.*rand(1,1e8)+pi).^2))
ans = -0.04153426177596753
```

Exercises

A

For Exercises 1-9, find the average value of the function $f(x)$ over the given interval.

1. $f(x) = 1$, over $[0,3]$

2. $f(x) = x$, over $[0,1]$

3. $f(x) = x^2$, over $[0,2]$

4. $f(x) = x^3$, over $[0,2]$

5. $f(x) = \sin 2x$, over $[0,\pi/2]$

6. $f(x) = e^x$, over $[-1,4]$

7. $f(x) = x^3$, over $[-1,1]$

8. $f(x) = \sin x$, over $[-\pi/2,\pi/2]$

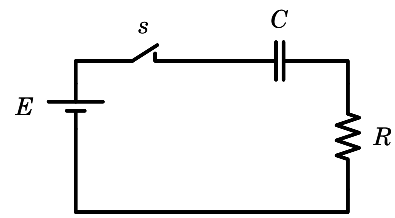
9. $f(x) = \dfrac{1}{x}$, over $[1,3]$

10. Electrical signals are commonly represented by a *periodic waveform* $x(t)$, which is a function of time t and has period T (i.e. T is the smallest positive number such that $x(t+T) = x(t)$ for all t). The *average power* of the waveform is defined as the average value of its square over a single period:

$$\left\langle x^2(t) \right\rangle = \frac{1}{T} \int_0^T x^2(t)\,dt\,.$$

(a) Find the average power of the waveform $x(t) = A\cos(\omega t + \phi)$, where $A > 0$ and $\omega > 0$ and ϕ are all constants.

(b) The *root mean square* of a waveform, abbreviated as *rms*, is the square root of the average power. Calculate the rms of the waveform from part (a). Write your answer in decimal form as a percentage of the amplitude A.



11. An electric circuit with a supplied voltage (electromotive force) E, a capacitor with capacitance C, and a resistor with resistance R, is shown in the picture on the right. When a switch s in the circuit is opened at time $t = 0$ the current I through the circuit begins to decrease exponentially as a function of time t (measured in seconds after the switch is opened), given by

$$I = \frac{E}{R} e^{-t/RC}$$

for $t \ge 0$.

(a) Sketch a rough graph of I as a function of t.

(b) Note that at time $t = 0$ the current is $I = \frac{E}{R}$ (measured in amperes), which is the familiar formula from Ohm's Law. That is the peak value of I. What is the current I at time $t = 5RC$? Write your answer in decimal form as a percentage of the peak current $\frac{E}{R}$ (e.g. $0.42\frac{E}{R}$, which would be 42% of the peak current).

(c) Find the average current in the circuit over the time interval $[0, 5RC]$. Write your answer in decimal form as a percentage of the peak current.

12. A spring with spring constant k and damping constant ν connects two point particles with mass m in a gravitational wave detector. A gravitational wave passes through the detector at time $t = 0$ and induces oscillation in the spring, with a period of $2\pi/\Omega$ and energy E at time $t \ge 0$ given by

$$E(t) = \frac{1}{4} mR^2 \left(\Omega^2 \sin^2(\Omega t + \phi) + \omega_0^2 \cos^2(\Omega t + \phi)\right) ,$$

where $\omega_0^2 = 2k/m$, $\phi = \tan^{-1}(2\nu\Omega/(m(\omega_0^2 - \Omega^2))$, and R is a constant.

(a) Show that the average energy $\langle E \rangle$ over one period $[0, 2\pi/\Omega]$ of oscillation is

$$\langle E \rangle = \frac{1}{8} mR^2 (\omega_0^2 + \Omega^2) .$$

(b) Suppose a large number of identical detectors of this type are uniformly distributed in a planar array at a density of σ detectors per unit area. The energy $E_\sigma(t)$ imparted to each detector at time $t \ge 0$ by a gravitational wave is

$$E_\sigma(t) = \nu\Omega^2 R^2 \sin^2(\Omega t + \phi) .$$

Show that the average energy $\langle E_\sigma \rangle$ over one period $[0, 2\pi/\Omega]$ of oscillation is

$$\langle E_\sigma \rangle = \frac{1}{2} \nu\Omega^2 R^2 .$$

B

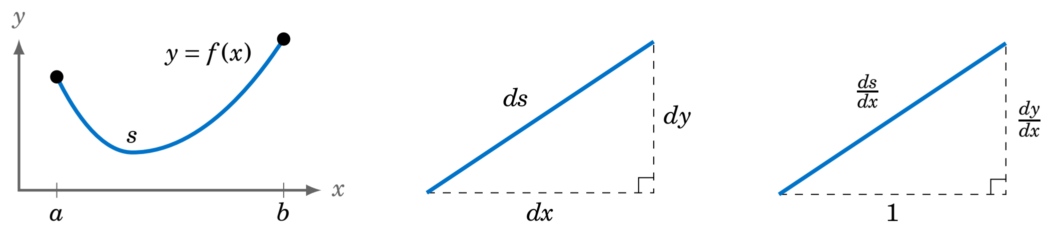
13. For the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, with $a > b > 0$, the foci are the points $(c, 0)$ and $(-c, 0)$, where $c = \sqrt{a^2 - b^2}$. Find the average distance from the ellipse to either of its foci in terms of the constants a, b, and c.

14. Write a computer program to use the Monte Carlo method with 1 million random points to approximate the average distance from the ellipse $\frac{x^2}{25} + \frac{y^2}{9} = 1$ to the point $(0, 0)$. Use symmetry to choose the smallest interval for the points. Could you have used formula (8.3) instead? Explain.

8.3 Arc Length and Curvature

Just like the area of a plane region can be found using calculus, so too can the length of a plane curve. Along the way the mystery mentioned in a footnote in Chapter 1 will finally be solved: what is the length of the hypotenuse of a right triangle with infinitesimal sides?

For a function $y = f(x)$ denote by s the length of the piece of that curve over an interval $[a,b]$, as in Figure 8.3.1 (a). Call s the curve's **arc length** over $[a,b]$.



(a) Arc length s **(b)** Infinitesimal triangle **(c)** Noninfinitesimal triangle

Figure 8.3.1 Arc length

By the Microstraightness Property, for $a \le x < b$ the curve is a straight line of length ds over the infinitesimal interval $[x, x+dx]$, as in Figure 8.3.1 (b), where $ds > 0$ is the infinitesimal change in s over that interval. Notice that you cannot simply apply the Pythagorean Theorem here, since that would make $ds = \sqrt{(dx)^2 + (dy)^2} = \sqrt{0+0} = 0$, which is false. The trick is to divide all sides of this infinitesimal right triangle by dx, which yields the similar—and noninfinitesimal—right triangle shown in Figure 8.3.1(c). The Pythagorean Theorem can then be applied to that triangle:

$$\frac{ds}{dx} = \sqrt{1^2 + \left(\frac{dy}{dx}\right)^2} \quad \Rightarrow \quad ds = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx$$

Summing up those infinitesimal lengths ds then yields the arc length s:

The arc length s of a curve $y = f(x)$ over $[a,b]$ is:

$$s = \int_a^b ds = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx \tag{8.4}$$

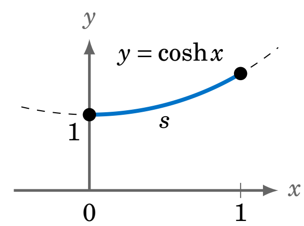
That such a formula exists is of course good news, but as you have probably guessed, the integral cannot be evaluated in a closed form except for a few functions.[2] In most cases numerical integration methods will be required.

[2]This will be the last "good news/bad news" scenario in this book. That's the good news.

Example 8.13

Find the arc length of the curve $y = \cosh x$ over $[0,1]$.

Solution: Since $\frac{dy}{dx} = \sinh x$, then the arc length s is:



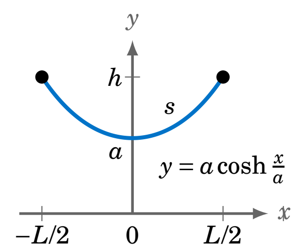
$$s = \int_0^1 \sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx = \int_0^1 \sqrt{1+\sinh^2 x}\,dx$$

$$= \int_0^1 \cosh x\,dx = \sinh x\,\Big|_0^1 = \sinh 1 - \sinh 0 = \sinh 1 \approx 1.1752$$

Example 8.14

A *catenary*—a hanging uniform cable whose ends are fastened at the same height h a distance L apart—has its lowest point—the *apex*—a distance $a > 0$ above the ground. It can be shown[3] that with the apex at $(0,a)$, the equation of the catenary is $y = a\cosh\frac{x}{a}$. Find the arc length of the catenary.

Solution: The figure on the right shows the catenary. By symmetry the total arc length s is twice the arc length over $[0,L/2]$:

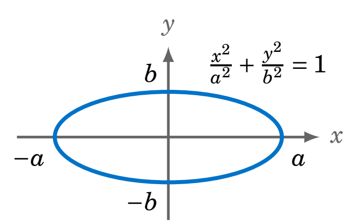


$$s = 2\int_0^{L/2} \sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx = 2\int_0^{L/2} \sqrt{1+\sinh^2\tfrac{x}{a}}\,dx$$

$$= 2\int_0^{L/2} \cosh\tfrac{x}{a}\,dx = 2a\sinh\tfrac{x}{a}\,\Big|_0^{L/2} = 2a\sinh\tfrac{L}{2a}$$

Elliptic Integrals

Suppose you tried to find the circumference s of the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, with $a > b > 0$, which has eccentricity $e = \frac{c}{a}$, where $c = \sqrt{a^2 - b^2}$. By symmetry, s is quadruple the arc length of the upper hemisphere $y = b\sqrt{1-\frac{x^2}{a^2}}$ over $[0,a]$:



$$s = 4\int_0^a \sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx = 4\int_0^a \sqrt{1+\left(-\frac{bx}{a\sqrt{a^2-x^2}}\right)^2}\,dx$$

$$= 4\int_0^a \sqrt{\frac{a^2(a^2-x^2)+b^2x^2}{a^2(a^2-x^2)}}\,dx$$

$$= \frac{4}{a}\int_0^a \sqrt{\frac{a^4-(a^2-b^2)x^2}{a^2-x^2}}\,dx$$

Now try the trigonometric substitution $x = a\sin\theta$, $dx = a\cos\theta\,d\theta$:

[3] See pp.162-163 in SMITH, C.E., *Applied Mechanics: Statics*, New York: John Wiley & Sons, Inc., 1976.

$$\begin{aligned}
s &= \frac{4}{a}\int_0^{\pi/2}\sqrt{\frac{a^4-(a^2-b^2)a^2\sin^2\theta}{a^2-a^2\sin^2\theta}}\; a\cos\theta\; d\theta \\
&= 4\int_0^{\pi/2}\sqrt{\frac{a^2-(a^2-b^2)\sin^2\theta}{1-\sin^2\theta}}\;\cos\theta\; d\theta \\
&= 4a\int_0^{\pi/2}\sqrt{1-\frac{a^2-b^2}{a^2}\sin^2\theta}\; d\theta \\
s &= 4a\int_0^{\pi/2}\sqrt{1-e^2\sin^2\theta}\; d\theta \quad (\text{since } e^2=\frac{c^2}{a^2}=\frac{a^2-b^2}{a^2})
\end{aligned}$$

The last integral is a special case of the **elliptic integral of the second kind**

$$E(k,\phi) = \int_0^{\phi}\sqrt{1-k^2\sin^2\theta}\; d\theta\;,$$

with $k=e$ and $\phi=\frac{\pi}{2}$. This special case is denoted by $E(k)=E(k,\frac{\pi}{2})$. Thus, the circumference s of the ellipse is:

$$s = 4a\,E(e) = 4a\,E(e,\tfrac{\pi}{2})$$

The integral $E(k)$ for $0<k<1$ cannot be evaluated in a closed form. There are tables[4] for certain values of k between 0 and 1, but a number of scientific computing applications have built-in functions to evaluate elliptic integrals.

For example, suppose you want to find the circumference s of the ellipse $\frac{x^2}{25}+\frac{y^2}{9}=1$. Then $a=5$, $b=3$, $c=\sqrt{a^2-b^2}=4$, and $e=\frac{c}{a}=0.8$, so $s=4a\,E(e)=20\,E(0.8)$. In the Python-based open-source mathematical software system Sage[5], the elliptic integral $E(k,\phi)$ is provided by the function `elliptic_e(`ϕ`,`k^2`)`. Use $k=e=0.8$ and $\phi=\frac{\pi}{2}$:

```
In [1]: 20*elliptic_e(pi/2,0.8^2)
Out[1]: 25.5269988633981
```

The circumference is thus approximately 25.5269988633981. In Octave/MATLAB the function `ellipke(`e^2`)` evaluates the elliptic integral $E(e)$, with one extra step:

```
MATLAB>> [K,E] = ellipke(0.8^2);
MATLAB>> 20*E
ans = 25.526998863398131
```

[4]See p.609 in Abramowitz, M. and I.A. Stegun, *Handbook of Mathematical Functions*, New York: Dover Publications, Inc., 1965.

[5]Available at `https://www.sagemath.org`

The parametric formula for arc length can be derived by dividing all sides of the infinitesimal right triangle in Figure 8.3.1(b) by dt, then applying the Pythagorean Theorem to the resulting noninfinitesimal right triangle:

$$\frac{ds}{dt} = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} \quad \Rightarrow \quad ds = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt$$

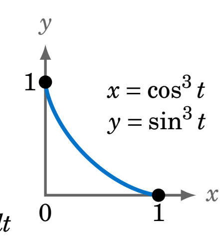
Sum up those infinitesimal lengths ds to obtain the arc length s:

The arc length s of a parametric curve $x = x(t)$, $y = y(t)$, $a \le t \le b$ is:

$$s = \int_a^b ds = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt \tag{8.5}$$

Example 8.15

Find the arc length of the parametric curve $x = \cos^3 t$, $y = \sin^3 t$, $0 \le t \le \pi/2$.



Solution: Since $\frac{dx}{dt} = -3\cos^2 t \sin t$ and $\frac{dy}{dt} = 3\sin^2 t \cos t$, then the arc length s is:

$$\begin{aligned}
s &= \int_0^{\pi/2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt = \int_0^{\pi/2} \sqrt{(-3\cos^2 t \sin t)^2 + (3\sin^2 t \cos t)^2}\, dt \\
&= \int_0^{\pi/2} \sqrt{9\cos^4 t \sin^2 t + 9\sin^4 t \cos^2 t}\, dt = \int_0^{\pi/2} 3\sqrt{\sin^2 t \cos^2 t\,(\cos^2 t + \sin^2 t)}\, dt \\
&= \int_0^{\pi/2} 3\sin t \cos t\, dt \quad (\text{since } \cos t \ge 0 \text{ and } \sin t \ge 0 \text{ for } 0 \le t \le \pi/2) \\
&= \frac{3}{2}\sin^2 t \,\Big|_0^{\pi/2} = \frac{3}{2}
\end{aligned}$$

The polar formula for arc length can be considered a special case of the parametric formula. A polar curve $r = r(\theta)$ for $\alpha \le \theta \le \beta$ has Cartesian coordinates $x = r(\theta)\cos\theta$ and $y = r(\theta)\sin\theta$, so that

$$\frac{dx}{d\theta} = \frac{dr}{d\theta}\cos\theta - r\sin\theta \quad \text{and} \quad \frac{dy}{d\theta} = \frac{dr}{d\theta}\sin\theta + r\cos\theta\,.$$

It is left as an exercise to show that putting these derivatives into formula (8.5)—using the parameter θ instead of t—yields the polar arc length formula:

The arc length s of a polar curve $r = r(\theta)$ for $\alpha \le \theta \le \beta$ is:

$$s = \int_\alpha^\beta ds = \int_\alpha^\beta \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2}\, d\theta \tag{8.6}$$

Example 8.16

Prove that the circumference of a circle of radius R is $2\pi R$.

Solution: Use the polar curve $r = R$ for $0 \le \theta \le 2\pi$. Then $\frac{dr}{d\theta} = 0$, so:

$$s = \int_0^{2\pi} \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2}\, d\theta = \int_0^{2\pi} \sqrt{R^2 + 0^2}\, d\theta = \int_0^{2\pi} R\, d\theta = R\theta \Big|_0^{2\pi} = 2\pi R \quad \checkmark$$

Curvature

In Chapter 4 you saw one simple measure of curvature: the second derivative. From Figure 8.3.2 it is clear that the parabola $y = x^2$ is less curved at the point $(1,1)$ than at the origin, yet $\frac{d^2y}{dx^2} = 2$ at each point. So the second derivative—the rate of change of the slopes $\frac{dy}{dx}$ of the tangent lines—does not fully capture the curvature of a curve; more information is needed. It turns out that the rate of change of the *angles* of the tangent lines is the key to curvature.

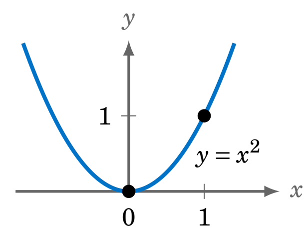


Figure 8.3.2

First consider a curve with arc length s between two points A and B on the curve. Let α be the angle between the tangent lines to the curve at A and B, as in Figure 8.3.3(a).

(a) Small curvature **(b)** Larger curvature and α

Figure 8.3.3 Curvature between A and B: same s, different α

For the same arc length s but larger angle α as in Figure 8.3.3(b), the curvature appears greater. This suggests that curvature should measure α relative to s:

The **average curvature** $\bar{\kappa}_{AB}$ of a curve between two points A and B on the curve is

$$\bar{\kappa}_{AB} = \frac{\alpha}{s} \tag{8.7}$$

where s is the arc length of the curve between A and B, and α is the angle between the tangent lines to the curve at A and B, called the **angle of contingence**.

Similar to how the instantaneous rate of change of a function at a point is the average rate of change over an infinitesimal interval, the curvature of a curve at a point can be defined as the average curvature over an infinitesimal length of the curve:

> The **curvature** κ of a curve at a point A on the curve is
>
> $$\kappa = \frac{d\phi}{ds} \tag{8.8}$$
>
> where s is the arc length function of the curve and ϕ is the angle that the tangent line to the curve at A makes with the positive x-axis.

The idea is that moving an infinitesimal length ds along the curve induces an infinitesimal difference $d\phi$ in the angles of the tangent lines. Roughly, as B moves toward A:

$$\lim_{B \to A} \bar{\kappa}_{AB} = \lim_{B \to A} \frac{\alpha}{s} = \frac{d\phi}{ds} = \kappa$$

Curvature can be viewed as the instantaneous rate of change of *direction* of a curve, but with respect to arc length (i.e. distance traveled) instead of time.

For a curve $y = f(x)$, recall from formula (3.2) in Section 3.1 that $\phi = \phi(x) = \tan^{-1} f'(x)$ at each point $(x, f(x))$ on the curve. Thus, by the Chain Rule:

$$\kappa = \frac{d\phi}{ds} = \frac{d\,(\tan^{-1} f'(x))}{ds} = \frac{\dfrac{f''(x)\,dx}{1+(f'(x))^2}}{ds}$$

So since $ds = \sqrt{1+(f'(x))^2}\,dx$ by formula (8.4), then:

> The curvature κ of a curve $y = f(x)$ is:
>
> $$\kappa = \frac{f''(x)}{(1 + (f'(x))^2)^{3/2}} \tag{8.9}$$

The above formula makes κ a function of x. Note also that $\kappa = 0$ for a straight line, and that the curve $y = f(x)$ is concave up if $\kappa > 0$ and concave down if $\kappa < 0$.

Example 8.17

Find the curvature of the curve $y = x^2$ for $x = 0$ and $x = 1$.

Solution: For $f(x) = x^2$, $f'(x) = 2x$ and $f''(x) = 2$, so for any x the curvature $\kappa = \kappa(x)$ is:

$$\kappa = \frac{f''(x)}{(1+(f'(x))^2)^{3/2}} = \frac{2}{(1+4x^2)^{3/2}}$$

In particular, $\kappa(0) = 2$ and $\kappa(1) = \frac{2}{5^{3/2}} \approx 0.1789$. So $y = x^2$ has more curvature at the origin than at $(1,1)$.

For a parametric curve $x = x(t)$, $y = y(t)$, the curvature κ will be a function of the parameter t. Since $\frac{dy}{dx} = \frac{y'(t)}{x'(t)}$ by formula (7.14) in Section 7.6, then by formula (7.15):

$$\frac{d^2y}{dx^2} = \frac{\frac{d}{dt}\left(\frac{dy}{dx}\right)}{\frac{dx}{dt}} = \frac{\frac{d}{dt}\left(\frac{y'(t)}{x'(t)}\right)}{x'(t)} = \frac{x'(t)\,y''(t) - y'(t)\,x''(t)}{(x'(t))^3}$$

So by formula (8.9):

$$\kappa = \frac{\frac{d^2y}{dx^2}}{\left(1+\left(\frac{dy}{dx}\right)^2\right)^{3/2}} = \frac{\frac{x'(t)\,y''(t) - y'(t)\,x''(t)}{(x'(t))^3}}{\left(1+\left(\frac{y'(t)}{x'(t)}\right)^2\right)^{3/2}} = \frac{x'(t)\,y''(t) - y'(t)\,x''(t)}{\left((x'(t))^2\right)^{3/2}\left(1+\left(\frac{y'(t)}{x'(t)}\right)^2\right)^{3/2}}$$

Simplify the denominator to obtain the parametric curvature formula:

The curvature κ of a parametric curve $x = x(t)$, $y = y(t)$ is:

$$\kappa = \frac{x'(t)\,y''(t) - y'(t)\,x''(t)}{\left((x'(t))^2 + (y'(t))^2\right)^{3/2}} \tag{8.10}$$

The derivation of the curvature formula in polar coordinates is left as an exercise:

The curvature κ of a polar curve $r = r(\theta)$ is:

$$\kappa = \frac{(r(\theta))^2 + 2(r'(\theta))^2 - r(\theta)\,r''(\theta)}{\left((r(\theta))^2 + (r'(\theta))^2\right)^{3/2}} \tag{8.11}$$

Example 8.18

Find the curvature of a circle of radius R.

Solution: Use the polar curve $r = r(\theta) = R$, so that $r'(\theta) = 0 = r''(\theta)$:

$$\kappa = \frac{(r(\theta))^2 + 2(r'(\theta))^2 - r(\theta)\,r''(\theta)}{\left((r(\theta))^2 + (r'(\theta))^2\right)^{3/2}} = \frac{R^2 + 2\cdot 0^2 - R\cdot 0}{(R^2 + 0^2)^{3/2}} = \frac{R^2}{R^3} = \frac{1}{R}$$

A circle thus has constant curvature, as you would expect by symmetry. It turns out that any planar curve with constant curvature is either a line or part of a circle.[6]

[6] See pp.62-63 in O'NEILL, B., *Elementary Differential Geometry*, New York: Academic Press, Inc., 1966.

Exercises

A

For Exercises 1-10, find the arc length of the given curve over the given interval.

1. $y = x^{3/2}$; $1 \le x \le 4$

2. $y = x^2$; $0 \le x \le 1$

3. $y = x^{2/3}$; $1 \le x \le 8$

4. $y = \dfrac{x^2}{4} - \dfrac{\ln x}{2}$; $1 \le x \le 2$

5. $y = \dfrac{x^4}{4} + \dfrac{1}{8x^2}$; $1 \le x \le 2$

6. $y = \ln \dfrac{e^x + 1}{e^x - 1}$; $1 \le x \le 2$

7. $x = e^t \cos t,\ y = e^t \sin t$; $0 \le \theta \le \pi$

8. $x = \cos t + t \sin t,\ y = \sin t - t \cos t$; $0 \le t \le \pi$

9. polar curve $r = 1 + \cos\theta$; $0 \le \theta \le 2\pi$

10. polar curve $r = e^\theta$; $0 \le \theta \le 2$

11. Find the arc length of the curve in Example 8.15 for $0 \le t \le \pi$.

12. Use formula (8.5) to find the circumference of the unit circle using two different parametrizations:

(a) $x = \cos t,\ y = \sin t,\ 0 \le t \le 2\pi$

(b) $x = \dfrac{1-t^2}{1+t^2},\ y = \dfrac{2t}{1+t^2},\ -\infty < t < \infty$

For Exercises 13-18 find the curvature of the given curve at the indicated points.

13. $y = \sin x$ at $x = 0$ and $x = \frac{\pi}{2}$

14. $y = \ln x$ at $x = 1$

15. $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ at $(a,0)$ and $(0,b)$

16. $y = e^x$ at $x = 0$

17. $x^2 - y^2 = 1$ at $(1,0)$

18. $r = 1 + \cos\theta$ at $\theta = 0$

B

19. Prove formula (8.6).

20. Prove formula (8.11).

21. Let α and β be nonzero constants. Show that the arc length s of $y = \beta \sin \frac{x}{\alpha}$ over the interval $[0, x_0]$ can be put in terms of the elliptic integral $E(k,\phi)$:

$$s = \sqrt{\alpha^2 + \beta^2} \cdot E\left(\sqrt{\frac{\beta^2}{\alpha^2+\beta^2}},\ \frac{x_0}{\alpha}\right)$$

22. For $-1 < k < 1$ and $-1 \le x \le 1$, define $u(x)$ and K by

$$u(x) = \int_0^x \frac{dt}{\sqrt{1-t^2}\,\sqrt{1-k^2t^2}} \quad \text{and} \quad K = u(1) = \int_0^1 \frac{dt}{\sqrt{1-t^2}\,\sqrt{1-k^2t^2}}$$

so that all the square roots are defined and positive.

(a) Show that u is an increasing function of x and thus has an inverse function, call it $x = \operatorname{sn} u$, with domain $[-K,K]$ and range $[-1,1]$.

(b) Define $\operatorname{cn} u = \sqrt{1 - \operatorname{sn}^2 u}$ and $\operatorname{dn} u = \sqrt{1 - k^2 \operatorname{sn}^2 u}$. Show that:

$$\frac{d}{du}(\operatorname{sn} u) = \operatorname{cn} u \operatorname{dn} u \quad , \quad \frac{d}{du}(\operatorname{cn} u) = -\operatorname{sn} u \operatorname{dn} u \quad , \text{and} \quad \frac{d}{du}(\operatorname{dn} u) = -k^2 \operatorname{sn} u \operatorname{cn} u$$

The functions $\operatorname{sn} u$, $\operatorname{cn} u$, and $\operatorname{dn} u$ are called the *Jacobian elliptic functions*.

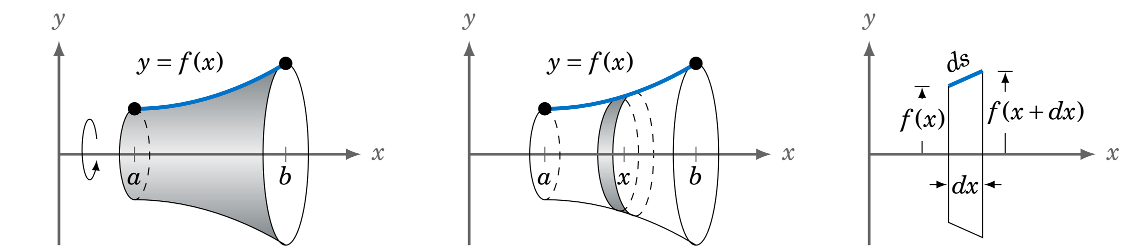
(c) Suppose that $\sin\phi = \operatorname{sn} u$. Show that $E(k,\phi) = \displaystyle\int_0^u \operatorname{dn}^2 v\, dv$.

23. The ends of a 50 ft long catenary are fastened 40 ft apart. Use a numerical method to find how much the apex dips below the ends. *(Hint: Solve for a in Example 8.14, then use symmetry.)*

8.4 Surfaces and Solids of Revolution

Long before calculus was invented the ancient Greeks (e.g. Archimedes) discovered the formulas for the volume and surface area of familiar three-dimensional objects such as the sphere.[7] Volumes and surface areas of arbitrary solids and surfaces can be found using multivariable calculus. However, single-variable calculus can be used in the special case of the objects possessing symmetry about an axis, via methods that involve revolving a curve or region in the xy-plane around an axis.

For example, revolve a curve $y = f(x) \geq 0$ around the x-axis, for $a \leq x \leq b$. This produces a **surface of revolution** in three dimensions, as in Figure 8.4.1(a).



(a) Revolve $y = f(x)$ around x-axis **(b)** dS over $[x, x+dx]$ **(c)** Side view of frustrum

Figure 8.4.1 Surface of revolution over $[a,b]$ with lateral surface area S

To find the total lateral surface area S, pick x in $[a,b)$ then find the infinitesimal surface area dS swept out over the infinitesimal interval $[x, x+dx]$, as in Figure 8.4.1(b). By the Microstraightness Property, the curve $y = f(x)$ is a straight line segment of length ds over that interval, so that the infinitesimal surface is a **frustrum**—a right circular cone with the vertex chopped off by a plane parallel to the base circle. From geometry[8] you might recall the formula for the lateral surface area of the frustrum in Figure 8.4.2: $\pi(r_1 + r_2)l$. Use that formula with $r_1 = f(x)$, $r_2 = f(x+dx) = f(x) + dy$, and $l = ds = \sqrt{1 + (f'(x))^2}\,dx$ (by formula (8.4) in Section 8.3) as in Figure 8.4.1(c), so that dS is

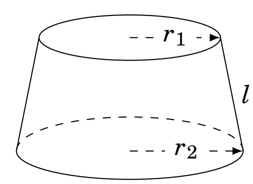


Figure 8.4.2

$$\begin{aligned} dS &= \pi(f(x) + (f(x) + dy))\sqrt{1 + (f'(x))^2}\,dx \\ &= 2\pi f(x)\sqrt{1 + (f'(x))^2}\,dx \;+\; \pi\sqrt{1 + (f'(x))^2}\,dy\,dx \\ &= 2\pi f(x)\sqrt{1 + (f'(x))^2}\,dx \;+\; 0 \end{aligned}$$

since $dy\,dx = f'(x)(dx)^2 = 0$. The surface area S is then the sum of all the areas dS:

[7]See Propositions 33 and 34 in *On the Sphere and Cylinder, Book I*, appearing in HEATH, T.L., *The Works of Archimedes*, Mineola, NY: Dover Publications, Inc., 2002. This work is also available at `https://archive.org`

[8]See pp.136-137 in WELCHONS A.M. AND W.R. KRICKENBERGER, *Solid Geometry*, Boston: Ginn & Co., 1936.

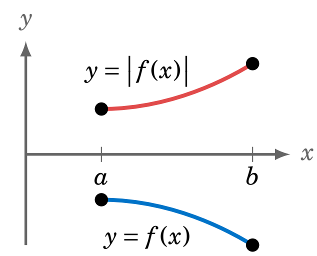
The surface area S of the surface of revolution obtained by revolving the curve $y = f(x) \geq 0$ around the x-axis for $a \leq x \leq b$ is:

$$S = \int_a^b dS = \int_a^b 2\pi f(x)\sqrt{1+(f'(x))^2}\,dx \tag{8.12}$$

For a general curve $y = f(x)$, possibly negative in $[a,b]$, the surface area S is:

$$S = \int_a^b dS = \int_a^b 2\pi |y|\,ds = \int_a^b 2\pi \left|f(x)\right|\sqrt{1+(f'(x))^2}\,dx \tag{8.13}$$

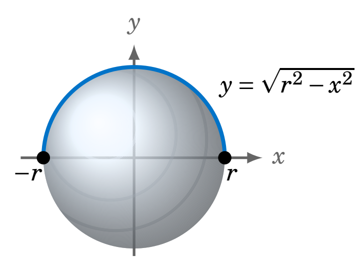
Note that formula (8.13) holds by symmetry and formula (8.12). A curve $y = f(x) < 0$ and the curve $y = \left|f(x)\right| = -f(x)$ are symmetric with respect to the x-axis, as in the figure on the right. Thus, both curves sweep out the same surface of revolution when revolved around the x-axis. This means formula (8.13) also holds if $y = f(x)$ changes sign in $[a,b]$: similar to the area between two curves, you would split the integral over different subintervals depending on the sign.



Example 8.19

Show that the surface area of a sphere of radius r is $4\pi r^2$.

Solution: Use the circle $x^2+y^2 = r^2$. The upper half of that circle is the curve $y = f(x) = \sqrt{r^2-x^2}$ over the interval $[-r,r]$, as in the figure on the right. Revolving that curve around the x-axis produces a sphere of radius r, whose surface area S is:

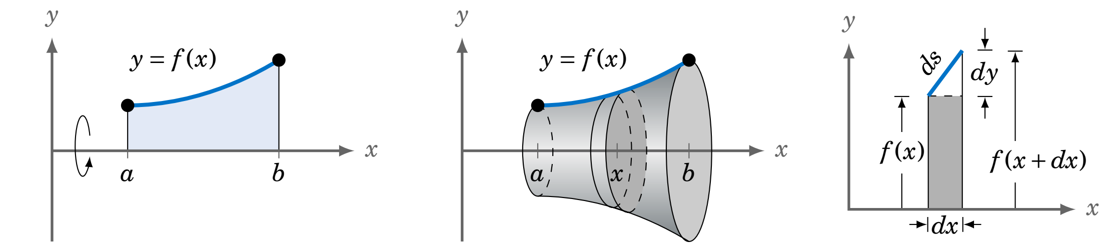


$$\begin{aligned} S &= \int_{-r}^{r} 2\pi f(x)\sqrt{1+(f'(x))^2}\,dx \\ &= \int_{-r}^{r} 2\pi\sqrt{r^2-x^2}\sqrt{1+\left(\frac{-x}{\sqrt{r^2-x^2}}\right)^2}\,dx \\ &= \int_{-r}^{r} 2\pi\sqrt{r^2-x^2}\sqrt{\frac{r^2}{r^2-x^2}}\,dx \\ &= 2\pi r x\,\Bigg|_{-r}^{r} = 4\pi r^2 \quad \checkmark \end{aligned}$$

A similar derivation using a frustrum yields the surface area S of the surface of revolution obtained by revolving a curve $y = f(x)$ around the y-axis, for $0 \leq a \leq x \leq b$:

$$S = \int_a^b dS = \int_a^b 2\pi |x|\,ds = \int_a^b 2\pi x\sqrt{1+(f'(x))^2}\,dx \tag{8.14}$$

Now suppose you revolve the region between a curve $y = f(x) \geq 0$ and the x-axis around the x-axis, for $a \leq x \leq b$ (see Figure 8.4.3(a)). This produces a **solid of revolution** in three dimensions, as in Figure 8.4.3(b). Notice that this solid consists of the surface of revolution as before along with its interior.



(a) Revolve region around x-axis **(b)** Solid **(c)** Infinitesimal strip

Figure 8.4.3 Solid of revolution over $[a,b]$ with volume area V

The goal is to find the volume V of this solid. The idea is to divide the solid into slices, like a loaf of bread. First, the infinitesimal volume dV of the frustrum swept out by a strip of infinitesimal width dx at x in $[a,b)$—shown in Figure 8.4.3(c)—is needed. By the Microstraightness Property the curve $y = f(x)$ is a straight line of length ds over the interval $[x, x+dx]$. There is thus a right triangle at the top of the strip—unshaded in Figure 8.4.3(c)—whose area A is zero: $A = \frac{1}{2}(dy)(dx) = \frac{1}{2}f'(x)(dx)^2 = 0$.

That triangle thus contributes no volume when revolved around the x-axis: the volume dV swept out by that strip all comes from the shaded rectangle of height $f(x)$ and width dx. That rectangle sweeps out a right circular cylinder of radius $f(x)$ and height dx (see Figure 8.4.4). The volume of a right circular cylinder of radius r and height h is defined as the area of the base circle times the height: $\pi r^2 h$. Hence,

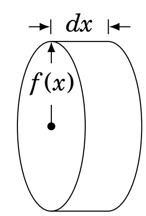


Figure 8.4.4

$$dV = \pi (f(x))^2 \, dx \; .$$

The total volume V of the solid is then the sum of all those infinitesimal volumes dV:

The volume V of the solid of revolution obtained by revolving the region between the curve $y = f(x)$ and the x-axis around the x-axis for $a \leq x \leq b$ is:

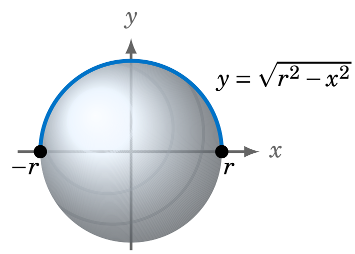
$$V = \int_a^b dV = \int_a^b \pi (f(x))^2 \, dx \qquad (8.15)$$

This method for finding the volume is called the **disc method**, since the cylinder of volume dV resembles a disc. Think of the discs as being similar to infinitesimally thin slices of a loaf of bread. Notice that absolute values are not needed in formula (8.15) since $f(x)$ is squared, so the formula holds even when $f(x)$ is negative.

Example 8.20

Show that the volume of a sphere of radius r is $\frac{4}{3}\pi r^3$.

Solution: Use the circle $x^2 + y^2 = r^2$. Revolve the region between the upper half of the circle $y = f(x) = \sqrt{r^2 - x^2}$ and the x-axis around the x-axis over the interval $[-r, r]$, as in the figure on the right. The solid of revolution swept out is a sphere of radius r, whose volume V is:



$$V = \int_{-r}^{r} \pi (f(x))^2 \, dx = \int_{-r}^{r} \pi (r^2 - x^2) \, dx = \pi r^2 x - \frac{1}{3}\pi x^3 \Big|_{-r}^{r}$$

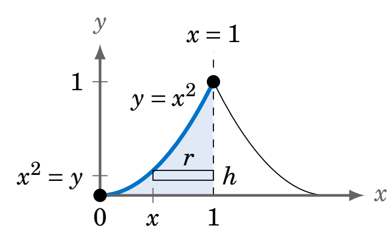
$$= \left(\pi r^3 - \frac{1}{3}\pi r^3\right) - \left(-\pi r^3 + \frac{1}{3}\pi r^3\right) = \frac{4}{3}\pi r^3 \quad \checkmark$$

Instead of memorizing formula (8.15), try to remember the more generic approach of revolving an infinitesimal rectangular strip around an axis, which might not be the x-axis. The idea is to find the radius r and height h—typically dx or dy—of the disc swept out by that strip, so that the disc's volume is $dV = \pi r^2 h$. Then integrate dV over the appropriate interval to find the volume V of the entire solid.

Example 8.21

Suppose the region bounded by the curve $y = x^2$ and the x-axis for $0 \le x \le 1$ is revolved around the line $x = 1$. Find the volume of the resulting solid of revolution.

Solution: The region is shaded in the figure on the right. Since the region is revolved around a vertical axis, the disc method will use discs with height dy, not dx. At a point x in $[0, 1]$ go up to the curve $y = x^2$ and draw a horizontal rectangular strip to the line $x = 1$, as shown in the figure. Let $h = dy$ and revolve that strip around the line $x = 1$, producing a disc of radius $r - 1 - x$ and height $h = dy$. Since $y = x^2$ implies $x = \sqrt{y}$, the volume dV of that disc is

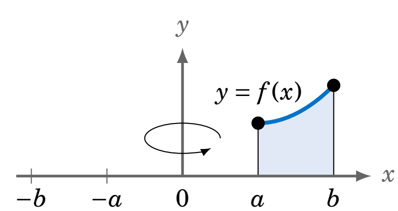


$$dV = \pi r^2 h = \pi (1 - x)^2 \, dy = \pi (1 - \sqrt{y})^2 \, dy = \pi (1 - 2\sqrt{y} + y) \, dy \, .$$

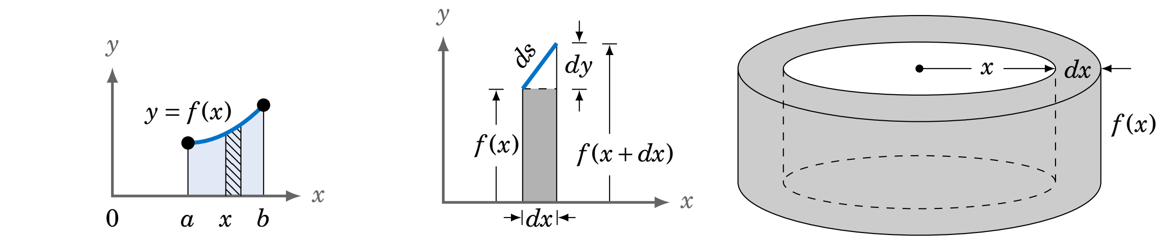
The volume V of the entire solid is then the sum of those volumes dV along the y-axis for $0 \le y \le 1$:

$$V = \int_0^1 dV = \int_0^1 \pi (1 - 2\sqrt{y} + y) \, dy = \pi \left(y - \frac{4}{3} y^{3/2} + \frac{1}{2} y^2\right) \Big|_0^1 = \pi \left(1 - \frac{4}{3} + \frac{1}{2}\right) = \frac{\pi}{6}$$

The **shell method** can be used for finding the volume of a solid with a "hole" in the middle, as in the solid of revolution produced by revolving the shaded region in the figure on the right around the y-axis. The hole in the solid between $x = -a$ and $x = a$ is a result of the gap between the y-axis and the region.



To find the volume V of that solid, at a point x in $[a,b)$ form an infinitesimal strip of width dx from the x-axis up to the curve $y = f(x)$, as in Figure 8.4.5(a).



(a) Revolve region around y-axis **(b)** Infinitesimal strip **(c)** Cylindrical shell

Figure 8.4.5 Shell method

Just like the strip in the disc method, the right triangle at the top of this strip—as in Figure 8.4.5(b)—has zero area and thus does not contribute to the volume dV of the right circular cylindrical shell swept out by the strip, shown in Figure 8.4.5(c). The volume of that shell is just the volume of the "outer" cylinder of radius $x+dx$ minus the volume of the "inner" cylinder of radius x, both with height $f(x)$:

$$\begin{aligned} dV &= \pi(x+dx)^2 f(x) - \pi x^2 f(x) \\ &= \cancel{\pi x^2 f(x)} + 2\pi x f(x)\,dx + \pi\cancelto{0}{(dx)^2} f(x) - \cancel{\pi x^2 f(x)} \\ &= 2\pi x f(x)\,dx \end{aligned}$$

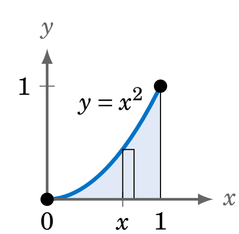
The volume V of the entire solid is then the sum of those volumes dV, using an absolute value to handle any sign for $f(x)$:

> The volume V of the solid of revolution obtained by revolving the region between the curve $y = f(x)$ and the x-axis around the y-axis for $0 \le a \le x \le b$ is:
>
> $$V = \int_a^b dV = \int_a^b 2\pi x \left|f(x)\right| dx \tag{8.16}$$

Example 8.22

Suppose the region bounded by the curve $y = x^2$ and the x-axis for $0 \le x \le 1$ is revolved around the y-axis. Find the volume of the resulting solid of revolution.

Solution: The region is shaded in the figure on the right. The vertical strip at x in $[0,1)$ with infinitesimal width dx and height $\left|f(x)\right| = f(x)$ is shown in the figure. That strip produces the shell with volume dV in formula (8.16), so by the shell method the volume V of the solid of revolution is:



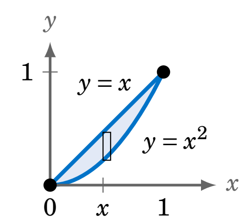
$$V = \int_0^1 dV = \int_0^1 2\pi x \left|f(x)\right| dx = \int_0^1 2\pi x \cdot x^2\, dx = \frac{\pi}{2}x^4 \Big|_0^1 = \frac{\pi}{2}$$

The volume dV in formula (8.16) can be generalized to $dV = 2\pi rhw$, where r is the distance from the axis of revolution to a generic vertical strip of infinitesimal width w in the region, and h is the height of the strip.

Example 8.23

Suppose the region bounded by the curves $y = x^2$ and $y = x$ is revolved around the y-axis. Find the volume of the resulting solid of revolution.

Solution: The region is shaded in the figure on the right, along with a vertical strip with infinitesimal width $w = dx$ at the distance $r = x$ from the y-axis in the region and height $h = x - x^2$. That strip produces the shell with volume $dV = 2\pi rhw = 2\pi x(x - x^2)\,dx$, so that the volume V of the solid of revolution is:



$$V = \int_0^1 dV = \int_0^1 2\pi x(x - x^2)\,dx = \left.\frac{2\pi}{3}x^3 - \frac{\pi}{2}x^4\right|_0^1 = \frac{\pi}{6}$$

Exercises

A

For Exercises 1-3, find the surface area of the surface of revolution produced by revolving the given curve around the x-axis for the given interval.

1. $y = \sqrt{4 - x^2}$ for $1 \le x \le 2$ **2.** $y = \cosh x$ for $0 \le x \le 1$ **3.** $y = \frac{x^3}{6} + \frac{1}{2x}$ for $1 \le x \le 3$

For Exercises 4-6, find the volume of the solid of revolution produced by revolving the region between the given curve and the x-axis around the x-axis for the given interval.

4. $y = x^3$ for $0 \le x \le 1$ **5.** $y = \sin x$ for $0 \le x \le \pi$ **6.** $y = \sqrt{x}$ for $0 \le x \le 1$

For Exercises 7-9, find the volume of the solid of revolution produced by revolving the region between the given curve and the x-axis around the y-axis for the given interval.

7. $y = \sin(x^2)$ for $0 \le x \le \sqrt{\pi}$ **8.** $y = \sin x$ for $0 \le x \le \pi$ **9.** $y = x^2 - x^3$ for $0 \le x \le 1$

10. Revolve the region in Example 8.23 around the line $x = 1$ and find the volume of the resulting solid.

11. Revolving the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ around the x-axis produces an *ellipsoid*, for $a > b > 0$. Show that the surface area of the ellipsoid is $2\pi b^2\left(1 + \frac{a}{eb}\sin^{-1} e\right)$, where e is the eccentricity of the ellipse.

12. Show that the volume inside the ellipsoid from Exercise 11 is $\frac{4}{3}\pi a b^2$.

13. Find the surface area and volume of a right circular cone of radius r and height h.

14. Formulas (8.13), (8.15) and (8.16) can be extended to include regions over infinite intervals—the integrals in those formulas simply become improper integrals. Consider the region between the curve $y = \frac{1}{x}$ and the x-axis over the interval $[1,\infty)$. Revolve that region around the x-axis.

(a) Show that the surface area of the resulting surface of revolution is infinite.

(b) Show that the volume of the resulting solid of revolution is π.

15. For $0 < a < b$, revolving the region inside the circle $(x - b)^2 + y^2 = a^2$ around the y-axis produces a donut-shaped solid of revolution called a *torus*. Show that the volume of the torus is $2\pi^2 a^2 b$.

16. Use formula (8.14) and symmetry to show that the torus from Exercise 15 has surface area $4\pi^2 ab$.

8.5 Applications in Physics and Statistics

This chapter concludes with a few applications showing how some familiar discrete sums can be replaced by integrals, which are essentially continuous sums.

Center of Gravity

Suppose a thin uniform rod has $n > 1$ masses $m_1, \ldots, m_n$ attached, with m_1 and m_n at the ends. The **center of gravity** of the masses is the point where—due to the Earth's gravity—the rod would be balanced if a fulcrum were placed there (see Figure 8.5.1(a)). Imagine the rod as part of the x-axis and the weights as point masses—with each mass m_k at x_k—and let the center of gravity be at $\bar{x}$, as in Figure 8.5.1(b).

(a) Rod balanced on fulcrum **(b)** Masses and $\bar{x}$ on x-axis

Figure 8.5.1 Center of gravity $\bar{x}$ for masses $m_1, \ldots, m_n$

The rod is balanced if the masses do not rotate the rod, i.e. the total **torque** is zero. Torque is defined here as force times position relative to $\bar{x}$. Each mass m_k applies a force $m_k g$ to the rod—where g is the (downward) acceleration due to the Earth's gravity—at position $(x_x - \bar{x})$ relative to $\bar{x}$. The total torque is thus zero if

$$(m_1 g)(x_1 - \bar{x}) + (m_2 g)(x_2 - \bar{x}) + \cdots + (m_n g)(x_n - \bar{x}) = 0$$

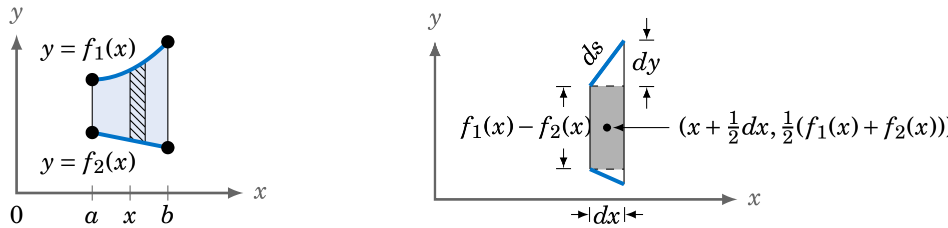
so that solving for $\bar{x}$ yields:

$$\bar{x} = \frac{m_1 g x_1 + \cdots + m_n g x_n}{m_1 g + \cdots + m_n g} = \frac{m_1 x_1 + \cdots + m_n x_n}{m_1 + \cdots + m_n} = \frac{\sum_{k=1}^{n} m_k x_k}{\sum_{k=1}^{n} m_k} \tag{8.17}$$

Each quantity $m_k x_k$ is called the **moment** of the mass m_k. Thus, $\bar{x}$ is the sum of the moments divided by the total mass. This idea can be extended to regions in the xy-plane, using an integral of a continuum of moments instead of a finite sum. The **center of gravity** of a planar region is defined as the point such that any force along a line through that point produces no rotation of the region about that line.[9] There should thus be zero torque in both the x and y directions, so the idea is to apply formula (8.17) in both directions to obtain the region's center of gravity $(\bar{x}, \bar{y})$.

[9]For a proof that such a point exists, see p.206 in BROWN, F.L., *Engineering Mechanics*, 2nd ed., New York: John Wiley & Sons, Inc., 1942. Some texts use the terms "center of mass" or "centroid" instead of "center of gravity," and there are differences in the meanings. However, for the situation presented here, where the gravitational field is assumed to have constant magnitude and direction throughout the region, they all mean the same thing.

A region can be thought of as a **lamina**—a thin plate with uniform density. Take the area of the region as its mass, which makes sense given the uniform density. For the region between two curves $y = f_1(x)$ and $y = f_2(x)$ over $[a,b]$, with $f_1(x) \ge f_2(x)$, take a vertical slice of width dx at some x, as in Figure 8.5.2(a). By the same arguments used in Section 8.4, all the area from that strip comes from the rectangle of height $f_1(x) - f_2(x)$ and width dx (see the shaded rectangle in Figure 8.5.2(b)).



(a) Region between $y = f_1(x)$ and $y = f_2(x)$ **(b)** Infinitesimal strip over $[x, x+dx]$

Figure 8.5.2 Center of gravity of a region

By the assumption of uniform density, the center of gravity of that rectangle is clearly its geometric center, whose coordinates are $\left(x + \frac{1}{2}dx, \frac{1}{2}(f_1(x) + f_2(x))\right)$. The entire mass of the strip can be treated as if it is concentrated at that point. The **moment** m_x of the strip about the x-axis is its mass times the position of its center of gravity relative to the x-axis (i.e. its y coordinate):

$$m_x = (f_1(x) - f_2(x))\,dx \cdot \left(\tfrac{1}{2}(f_1(x) + f_2(x))\right) = \tfrac{1}{2}\left((f_1(x))^2 - (f_2(x))^2\right)dx$$

Similarly the moment m_y of the strip about the y-axis is its mass times the x coordinate of its center of gravity:

$$\begin{aligned} m_y &= (f_1(x) - f_2(x))\,dx \cdot (x + \tfrac{1}{2}dx) = x(f_1(x) - f_2(x))\,dx + \tfrac{1}{2}(f_1(x) - f_2(x))(dx)^2 \\ &= x(f_1(x) - f_2(x))\,dx \end{aligned}$$

The moments M_x and M_y of the entire region about the x-axis and y-axis, respectively, are defined as the sum of the respective moments m_x and m_y of all strips over $[a,b]$:

$$M_x = \int_a^b m_x = \int_a^b \tfrac{1}{2}\left((f_1(x))^2 - (f_2(x))^2\right)dx \quad\text{and}\quad M_y = \int_a^b m_y = \int_a^b x(f_1(x) - f_2(x))\,dx$$

Note in formula (8.17) that the denominator is the sum of all the masses in the system. For the region that total mass would simply be its area M:

$$M = \int_a^b (f_1(x) - f_2(x))\,dx$$

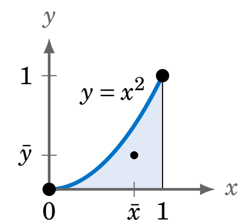
Dividing the moments M_x adn M_y by M yields the formula for the center of gravity:

The center of gravity $(\bar{x},\bar{y})$ of the region between the curves $y = f_1(x)$ and $y = f_2(x)$ over $[a,b]$, with $f_1(x) \ge f_2(x)$, is given by:

$$\bar{x} = \frac{M_y}{M} = \frac{\int_a^b x\,(f_1(x) - f_2(x))\,dx}{\int_a^b (f_1(x) - f_2(x))\,dx} \quad\text{and}\quad \bar{y} = \frac{M_x}{M} = \frac{\int_a^b \frac{1}{2}\left((f_1(x))^2 - (f_2(x))^2\right)dx}{\int_a^b (f_1(x) - f_2(x))\,dx} \tag{8.18}$$

Example 8.24

Find the center of gravity of the region bounded by the curve $y = x^2$ and the x-axis for $0 \le x \le 1$.



Solution: The region is shaded in the figure on the right. Using $y = f_1(x) = x^2$ and $y = f_2(x) = 0$ in formula (8.18) yields

$$M_x = \int_0^1 \tfrac{1}{2}(f_1(x))^2\,dx = \int_0^1 \tfrac{1}{2}x^4\,dx = \tfrac{1}{10}x^5\,\Big|_0^1 = \tfrac{1}{10}$$

$$M_y = \int_0^1 x\,f_1(x)\,dx = \int_0^1 x^3\,dx = \tfrac{1}{4}x^4\,\Big|_0^1 = \tfrac{1}{4}$$

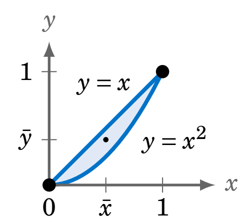
$$M = \int_0^1 f_1(x)\,dx = \int_0^1 x^2\,dx = \tfrac{1}{3}x^3\,\Big|_0^1 = \tfrac{1}{3}$$

so that the center of gravity $(\bar{x},\bar{y})$ is:

$$\bar{x} = \frac{M_y}{M} = \frac{1/4}{1/3} = \frac{3}{4} \quad\text{and}\quad \bar{y} = \frac{M_x}{M} = \frac{1/10}{1/3} = \frac{3}{10}$$

Example 8.25

Find the center of gravity of the region bounded by the curves $y = x$ and $y = x^2$.



Solution: The region is shaded in the figure on the right. Using $y = f_1(x) = x$ and $y = f_2(x) = x^2$ in formula (8.18) yields

$$M_x = \int_0^1 \tfrac{1}{2}\left((f_1(x))^2 - (f_2(x))^2\right)dx = \int_0^1 \tfrac{1}{2}(x^2 - x^4)\,dx = \tfrac{1}{6}x^3 - \tfrac{1}{10}x^5\,\Big|_0^1 = \tfrac{1}{15}$$

$$M_y = \int_0^1 x\,(f_1(x) - f_2(x))\,dx = \int_0^1 (x^2 - x^3)\,dx = \tfrac{1}{3}x^3 - \tfrac{1}{4}x^4\,\Big|_0^1 = \tfrac{1}{12}$$

$$M = \int_0^1 (f_1(x) - f_2(x))\,dx = \int_0^1 (x - x^2)\,dx = \tfrac{1}{2}x^2 - \tfrac{1}{3}x^3\,\Big|_0^1 = \tfrac{1}{6}$$

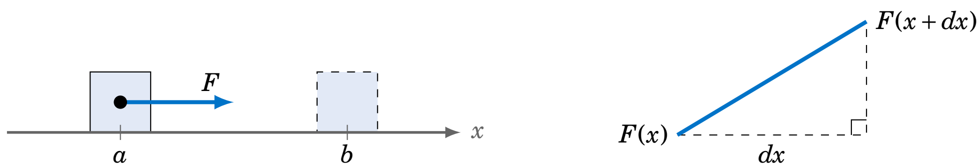
so that the center of gravity $(\bar{x},\bar{y})$ is:

$$\bar{x} = \frac{M_y}{M} = \frac{1/12}{1/6} = \frac{1}{2} \quad\text{and}\quad \bar{y} = \frac{M_x}{M} = \frac{1/15}{1/6} = \frac{2}{5}$$

Work

Suppose that a constant force displaces an object along a line in the same direction in which the force is applied. The **work** done by the force is defined as the force times the displacement. For example, if the constant force F moves an object from position $x = a$ to $x = b$ on the x-axis, as in Figure 8.5.3(a), then the work W done by the force is:

$$W \;=\; \text{force} \times \text{displacement} \;=\; \text{force} \times (\text{final position} - \text{initial position}) \;=\; F \cdot (b-a)$$



(a) Constant force F **(b)** Variable force F over $[x, x+dx]$

Figure 8.5.3 Work W as the effect of a force F displacing an object from $x = a$ to $x = b$

Suppose now that the force F is a function of position x over $[a,b]$: $F = F(x)$. By the Microstraightness Property, over an infinitesimal interval $[x, x+dx]$ the curve $y = F(x)$ is a straight line, as in Figure 8.5.3(b). How should the work dW performed by F over this infinitesimal interval be defined? After all, F is not constant over $[x, x+dx]$—it takes every value between $F(x)$ and $F(x+dx)$. It is left as an exercise to show that *any* value in that range can be used—they all result in the same amount $F(x)\,dx$ for the work performed.[10]

For example, suppose you use the value halfway between $F(x)$ and $F(x+dx)$ as the value of F: $\frac{1}{2}(F(x)+F(x+dx))$. Then the work dW as force times displacement is:

$$\begin{aligned} dW &= \frac{1}{2}(F(x)+F(x+dx))\;dx \;=\; \frac{1}{2}(F(x)+F(x)+F'(x)\,dx)\;dx \\ &= F(x)\,dx \;+\; \frac{1}{2}F'(x)\cancelto{0}{(dx)^2} \\ &= F(x)\,dx \end{aligned}$$

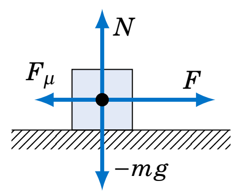
Define the total work W over $[a,b]$ as the sum of all the dW:

> The **work** performed by a force $F(x)$ in displacing an object along the x-axis from $x = a$ to $x = b$ is:
>
> $$W \;=\; \int_a^b dW \;=\; \int_a^b F(x)\;dx \qquad (8.19)$$

[10]Note how this is different than claiming that F is "essentially constant over small intervals," as most textbooks do. Instead, the additional infinitesimal force beyond $F(x)$ contributes zero work over $[x, x+dx]$.

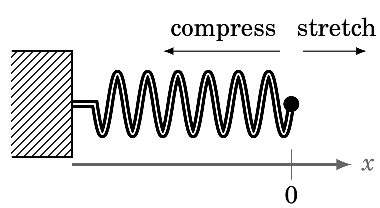
Before continuing, some possible confusion needs to be cleared up. First, force is always a **vector**—it has both a magnitude and a direction. For the forces considered here, which act in a single dimension (e.g. along the x-axis), by convention the direction of the force is indicated by its sign: positive in the direction toward $+\infty$, negative in the direction toward $-\infty$. So a force of 3 N acts in the opposite direction as a force of -3 N, but they have the same magnitude $|3| = 3$.

Second, work is not a vector—it is a **scalar**, meaning it has a magnitude but no direction. That magnitude can have any sign, though. Work is positive if the object is displaced in the same direction as the force, but is negative if the displacement is in the opposite direction of the force. For example, if you lift an object straight up from the ground, then you did positive work—the object moved in the same direction as the force you used. However, the force of gravity did negative work on the object as you lifted, since gravity works downward yet the object moved upward.



Last, zero work is done by a force if no displacement in its direction occurs. In particular, forces acting perpendicular to the line of displacement perform no work. For example, consider an object of mass m on a flat horizontal table top as in the figure on the right. If you push that object to the right with a force F (performing positive work), then both the downward force of gravity $-mg$ and the upward normal force N exerted by the table perform zero work on the object. The force of friction F_μ from the table surface does negative work, as it opposes the force F. As an another example, no work is performed by holding a 100 lb object still and above the ground.

Example 8.26



Hooke's law states that a coiled spring has an elastic *restoring force* $F = -kx$, where x is the displacement of the end of the spring from its equilibrium position as the spring is stretched or compressed, and $k > 0$ is the *spring constant*—or *stiffness coefficient*—specific to the spring. This force always tries to restore the spring to its equilibrium position, and the law holds only for a limited range of x. For a spring laid horizontally imagine it lies on the x-axis with the equilibrium position at $x = 0$, as in the figure on the right.

(a) Find the spring constant k if a force of 2 N stretches the spring by 4 cm.

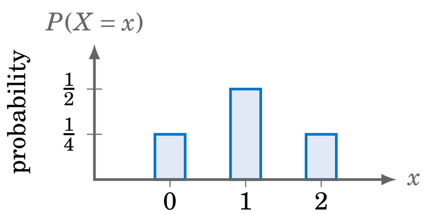
(b) Use part (a) to find the work performed by compressing the spring 3 cm.

Solution: (a) The force required to stretch the spring by an amount x is $F = kx$, since that force must counter the restoring force. Thus, $k = \frac{F}{x} = \frac{2\text{N}}{4\text{cm}} = \frac{2\text{N}}{0.04\text{m}} = 50$ N/m.
(b) By part (a) the force required to compress the string to position x is $F(x) = kx = 50x$, since again it must counter the restoring force. Thus, since 3 cm is 0.03 m, the work W performed is:

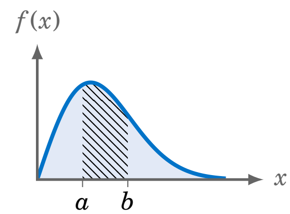
$$W = \int_0^{-0.03} F(x)\,dx = \int_0^{-0.03} 50x\,dx = 25x^2\,\Big|_0^{-0.03} = 25(-0.03)^2 - 0 = 0.0225 \text{ Nm}$$

Probability

Suppose you flip two evenly balanced pennies and let X be the number of heads in the result. Then X is a **discrete random variable**—discrete because it can take only a discrete set of values (0, 1 and 2); random because its value is left to chance. The **probability** of a penny being flipped heads is $50\% = \frac{1}{2}$, i.e. that is its theoretical likelihood since heads and tails are equally likely. The **sample space** S of all possible outcomes is the set $S = \{TT, TH, HT, HH\}$, where H is heads and T is tails (e.g. HT means the first penny came up heads and the second came up tails). Figure 8.5.4(a) shows a bar chart of the probabilities—as numbers between 0 and 1—with $P(X = x)$ denoting the probability of the **event** that X equals the number x. Notice that the sum of the probabilities is 1, and $P(X = x) = 0$ if x is not 0, 1, or 2.



(a) Discrete: $P(X = x)$



(b) Continuous: $P(a < X < b)$

Figure 8.5.4 Probability: discrete vs continuous random variables

The idea behind a **continuous random variable** X is to fill in those gaps between the bars in Figure 8.5.4(a), so that X would represent a continuous quantity, e.g. time, distance, temperature. Rather than finding $P(X = x)$ you would find the probability that X is in a continuum such as an interval, e.g. $P(a < X < b)$ (see Figure 8.5.4(b)).

For a continuous random variable X define $P(X = x) = 0$ for all real x, and define the **probability density function** for X as a function $f(x) \ge 0$ such that:

(a) $\displaystyle\int_{-\infty}^{\infty} f(x)\,dx \;=\; 1$

(b) $P(a < X < b) \;=\; \displaystyle\int_a^b f(x)\,dx$ for all $a < b$ (including $a = -\infty$ and $b = \infty$)

Notice that since $P(X = a) = 0$ then $P(a \le X < b) = P(a < X < b)$. In general, $<$ and $\le$ are interchangeable for events involving continuous random variables (as well as $>$ and $\ge$). In the remainder of this section it will be assumed that all random variables are continuous, for which the sample space is typically all of $\mathbb{R}$ or some interval, finite or infinite (e.g. $(0, \infty)$).

Example 8.27

Let X be the lifetime—i.e. the time to failure—of an electronic component. If the average lifetime of the component is 700 days, then the probability density function $f(x)$ for the random variable X is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0, \\ 0 & \text{if } x < 0 \end{cases} \tag{8.20}$$

where $\lambda = \frac{1}{700}$ and x is the number of days. In this case X is said to have the *exponential distribution with parameter* λ. Find the probability that the lifetime of the component is:

(a) between 600 and 800 days

(b) greater than 700 days

Solution: (a) The probability is:

$$P(600 < X < 800) = \int_{600}^{800} f(x)\,dx = \int_{600}^{800} \frac{1}{700} e^{-\frac{x}{700}}\,dx = -e^{-\frac{x}{700}} \Bigg|_{600}^{800} = -e^{-\frac{800}{700}} + e^{-\frac{600}{700}} \approx 0.1055$$

Thus, there is about a 10.55% chance that the component's lifetime will be between 600 and 800 days. (b) The probability is:

$$P(X > 700) = \int_{700}^{\infty} f(x)\,dx = \int_{700}^{\infty} \frac{1}{700} e^{-\frac{x}{700}}\,dx = -e^{-\frac{x}{700}} \Bigg|_{700}^{\infty} = 0 + e^{-1} \approx 0.3679$$

Exercises

A

For Exercises 1-3, find the center of gravity of the region bounded by the given curves over the given interval.

1. $y = x^3$ and $y = 0$; $0 \le x \le 1$ **2.** $y = -x + 1$ and $y = 0$; $0 \le x \le 1$ **3.** $y = x^2$ and $y = x^3$; $0 \le x \le 1$

4. Find the center of gravity of the region inside the circle $x^2 + y^2 = r^2$ and above the x-axis.

5. Find the center of gravity of the region inside the circle $x^2 + y^2 = r^2$ in the first quadrant.

6. Find the center of gravity of the region inside the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ and above the x-axis.

7. Find the center of gravity of the region between the circle $x^2 + y^2 = 4$ and the ellipse $\frac{x^2}{4} + y^2 = 1$ above the x-axis.

8. Would formula (8.18) for the center of gravity change if the mass of a region were proportional—but not equal—to its area, say, by a constant positive proportion $\delta \ne 1$? Explain.

9. If a spring requires 3 N of force to be compressed 5 cm, how much work would be performed in stretching the spring 8 cm?

10. The gravitational force $F(x)$ exerted by the Earth on an object of mass m at a distance x from the center of the Earth is

$$F(x) = -\frac{mgr_e^2}{x^2}$$

where r_e is the radius of the Earth. If the object is released from rest at a distance r_o from the center of the Earth, find the work performed by gravity in bringing the object to the Earth's surface.

11. Recall that the ideal gas law states that $PV = RT$, where R is a constant, P is the pressure, V is the volume, and T is the temperature. It can be shown that the work W done by an ideal gas in expanding the volume from V_a to V_b is

$$W = \int_{V_a}^{V_b} P\,dV\,.$$

Calculate W.

12. Verify that $\int_{-\infty}^{\infty} f(x)\,dx = 1$ for the function $f(x)$ in formula (8.20) in Example 8.27 for all $\lambda > 0$.

13. Find $P(X < 300)$ in Example 8.27.

14. The *distribution function* $F(x)$ for a random variable X is defined as $F(x) = P(X \le x)$ for all x. Show that $F'(x) = f(x)$, where $f(x)$ is the probability density function for X.

B

15. Formula (8.18) can be extended to regions over an infinite interval, provided the area is finite. Use that fact to find the center of gravity of the region between $y = e^{-x}$ and the x-axis for $0 \le x < \infty$.

16. The *expected value* (or *mean*) $E[X]$ of a random variable X with probability density function $f(x)$ is

$$E[X] = \int_{-\infty}^{\infty} x\,f(x)\,dx\,.$$

Show that $E[X] = \frac{1}{\lambda}$ if X has the exponential distribution with parameter $\lambda > 0$.
Note: The expected value can be thought of as the weighted average of all possible values of X, with weights determined by probability. It is analogous to the idea of a center of gravity.

17. A random variable X is said to have a *normal distribution* if its probability density function $f(x)$ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for all } x$$

where $\sigma > 0$ and μ are constants. This is the famous "bell curve" in statistics.

(a) Verify that $\int_{-\infty}^{\infty} f(x)\,dx = 1$. *(Hint: Use Example 6.25 and a substitution.)*

(b) Show that $E[X] = \mu$.

(c) Use numerical integration to show that $P(-1 < X < 1) \approx 0.6827$ when $\mu = 0$ and $\sigma = 1$.

18. A random variable X has the *beta distribution* if its probability density function $f(x)$ is

$$f(x) = \begin{cases} \frac{1}{B(a,b)}\,x^{a-1}(1-x)^{b-1} & \text{if } 0 \le x \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

for positive constants a and b, where $B(a,b)$ is the Beta function. Show that $E[X] = \frac{a}{a+b}$.

19. Show that any value between $F(x)$ and $F(x+dx)$ for the force over $[x, x+dx]$ gives the same formula $dW = F(x)\,dx$ for the work performed over that interval. *(Hint: Consider $F(x+\alpha\,dx)$ for $0 \le \alpha \le 1$.)*

20. A drop of water of mass M is released from rest at a height sufficient for the drop to evaporate completely, losing mass m each second (i.e. at a constant rate). Ignoring air resistance, show that the work performed by gravity on the drop up to complete evaporation is $\frac{g^2M^2}{6m^2}$.

21. This exercise is related to Einstein's famous law $E = mc^2$. The *relativistic momentum* p of a particle of mass m moving at a speed v along a straight line (say, the x-axis) is

$$p = \frac{mv}{\sqrt{1 - \frac{v^2}{c^2}}} ,$$

where c is the speed of light. The *relativistic force* on the particle along that line is

$$F = \frac{dp}{dt} ,$$

which is the same formula as Newton's Second Law of motion in classical mechanics. Assume that the particle starts at rest at position x_1 and ends at position x_2 along the x-axis. The work done by the force F on the particle is:

$$W = \int_{x_1}^{x_2} F\,dx = \int_{x_1}^{x_2} \frac{dp}{dt}\,dx$$

(a) Show that

$$\frac{dp}{dv} = \frac{m}{\left(1 - \frac{v^2}{c^2}\right)^{3/2}} .$$

(b) Use the Chain Rule formula

$$\frac{dp}{dt} = \frac{dp}{dv}\,\frac{dv}{dx}\,\frac{dx}{dt}$$

to show that

$$F\,dx = v\,\frac{dp}{dv}\,dv .$$

(c) Use parts (a) and (b) to show that

$$W = \int_0^v \frac{dp}{dv}\,v\,dv = \int_0^v \frac{mv}{\left(1 - \frac{v^2}{c^2}\right)^{3/2}}\,dv .$$

(d) Use part (c) to show that

$$W = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}} - mc^2 .$$

(e) Define the *relativistic kinetic energy* K of the particle to be $K = W$, and define the *total energy* E to be

$$E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}} .$$

So by part (d), $K = E - mc^2$. Show that

$$E^2 = p^2c^2 + (mc^2)^2 .$$

(Hint: Expand the right side of that equation.)

(f) What is E when the particle is at rest?

22. A *median* of a triangle is a line segment from a vertex to the midpoint of the opposite side, and the three medians intersect at a common point. Show that this point is a triangle's center of gravity.

CHAPTER 9

Infinite Sequences and Series

9.1 Sequences and Series

In the 5$^{\text{th}}$ century B.C. the ancient Greek philosopher Zeno of Elea devised several paradoxes, the most famous of which—*The Dichotomy*—asserts that if space is infinitely divisible then motion is impossible. The argument goes like this: imagine a line segment of finite length, say 1 m, with a person at one end as in Figure 9.1.1.

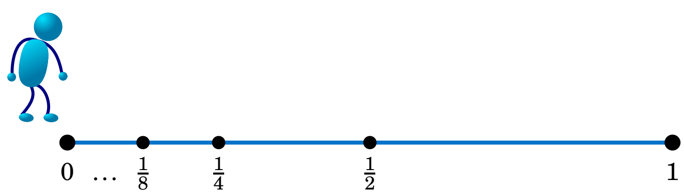


Figure 9.1.1 Zeno's motion paradox

Before traversing the entire distance the person would first need to travel one half the distance. Before doing that, though, he would need to travel one fourth the distance, and before that one eighth the distance, and so on. There is thus no "first" distance for him to traverse, meaning his motion cannot even begin!

Obviously motion is possible, or you would not be reading this. Does that mean Zeno's reasoning is flawed? More on that later. In the meantime, notice a few things in Figure 9.1.1. First, the distance markers $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$ form an **infinite sequence** of numbers approaching 0. Second, the sum of the distances between successive markers is an **infinite series** which should equal the total segment length 1:

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 1$$

It will be shown shortly that the sum is indeed 1, which turns out to have no bearing on Zeno's paradox. First some definitions are needed.

A **sequence** is an ordered list of objects, which in this book will always be real numbers. Sequences can be finite or infinite: finite if there is a last number in the list, infinite if every number in the list is followed by another number (i.e. a "successor"). Sequences should not be confused with *sets*—order matters in a sequence but not in a set, and numbers may repeat in a sequence but not in a set.

For example, the sequences $\langle 1,2,3\rangle$ and $\langle 1,3,2\rangle$ are different, since order matters. However, the numbers in those sequences comprise the same set $\{1,2,3\}$.

The simplest example of an infinite sequence is $\mathbb{N}$, the set of natural numbers: $0,1,2,3,\ldots$. In fact, the numbers $a_0, a_1, a_2, \ldots$ in any infinite sequence of real numbers can be written as the range of a function f mapping $\mathbb{N}$ into $\mathbb{R}$:

$$f(n) = a_n \tag{9.1}$$

Typically notation such as $\{a_n\}_{n=0}^{\infty}$ is used for representing an infinite sequence, or simply $\{a_n\}$ when the initial value of the index n is understood (and n is always an integer). The intuitive notion of the **limit** of an infinite sequence can be stated formally:

A real number L is the **limit** of an infinite sequence $\{a_n\}$, written as

$$\lim_{n\to\infty} a_n = L \quad \text{or simply} \quad a_n \to L\,,$$

if for any given number $\epsilon > 0$ there exists an integer N such that

$$\left|a_n - L\right| < \epsilon \quad \text{for all } n > N\,. \tag{9.2}$$

In this case the sequence $\{a_n\}$ is said to **converge** to L and is called a **convergent sequence**. A sequence that is not convergent is **divergent**.

In other words, a sequence $\{a_n\}$ converges to L if the terms a_n can be made arbitrarily close to L for n sufficiently large. In most cases the formal definition will not be needed, since by formula (9.1) the same rules and formulas from Chapters 1 and 3 for the limit of a function $f(x)$ as x approaches ∞ apply to sequences (e.g. sums and products of limits, L'Hôpital's Rule). All you have to do is replace x by n.

Example 9.1

For integers $n \ge 1$ define $a_n = \frac{1}{2^n}$. Find $\lim_{n\to\infty} a_n$ if it exists.

Solution: Since $\lim_{x\to\infty} \dfrac{1}{2^x} = 0$ then replacing x by n shows that

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{1}{2^n} = 0\,.$$

Example 9.2

For integers $n \ge 0$ define $a_n = \frac{2n+1}{3n+2}$. Is $\{a_n\}$ a convergent sequence? If so then find its limit.
Solution: By L'Hôpital's Rule, treating an integer $n \ge 0$ as a real-valued variable x,

$$\lim_{n\to\infty} a_n \;=\; \lim_{n\to\infty} \frac{2n+1}{3n+2} \;=\; \lim_{n\to\infty} \frac{2}{3} \;=\; \frac{2}{3}\,.$$

Thus the sequence is convergent and its limit is $\frac{2}{3}$.

Example 9.3

For integers $n \ge 0$ define $a_n = \frac{e^n}{3n+2}$. Is $\{a_n\}$ a convergent sequence? If so then find its limit.
Solution: By L'Hôpital's Rule, treating an integer $n \ge 0$ as a real-valued variable x,

$$\lim_{n\to\infty} a_n \;=\; \lim_{n\to\infty} \frac{e^n}{3n+2} \;=\; \lim_{n\to\infty} \frac{e^n}{3} \;=\; \infty$$

Thus the sequence is divergent.

Example 9.4

The famous *Fibonacci sequence*[1] $\{F_n\}$ starts with the numbers 0 and 1, then each successive term is the sum of the previous two terms: $F_0 = 0$, $F_1 = 1$,

$$F_n \;=\; F_{n-1} \,+\, F_{n-2} \quad \text{for integers } n \ge 2 \tag{9.3}$$

Equation (9.3) is a *recurrence relation*. The first ten Fibonacci numbers are $0, 1, 1, 2, 3, 5, 8, 13, 21, 34$. Clearly $\{F_n\}$ is a divergent sequence, since $F_n \to \infty$. For $n \ge 2$ define $a_n = F_n/F_{n-1}$. The first few values are:

$$a_2 \;=\; \frac{F_2}{F_1} \;=\; \frac{1}{1} \;=\; 1 \quad,\quad a_3 \;=\; \frac{F_3}{F_2} \;=\; \frac{2}{1} \;=\; 2 \quad,\quad a_4 \;=\; \frac{F_4}{F_3} \;=\; \frac{3}{2} \;=\; 1.5 \quad,\quad a_5 \;=\; \frac{F_5}{F_4} \;=\; \frac{5}{3} \;\approx\; 1.667$$

Show that $\{a_n\}$ is convergent. In other words, in the Fibonacci sequence the ratios of each term to the previous term converge to some number.

Solution: There are many ways to prove this, perhaps the simplest being to assume the sequence is convergent and then find the limit (which would be impossible if the sequence were divergent). So assume that $a_n \to a$ for some real number a. Then divide both sides of formula (9.3) by F_{n-1}, so that

$$\frac{F_n}{F_{n-1}} \;=\; 1 \,+\, \frac{F_{n-2}}{F_{n-1}} \quad\Rightarrow\quad a_n \;=\; 1 \,+\, \frac{1}{a_{n-1}} \quad \text{for integers } n \ge 2$$

Now take the limit of both sides of the last equation as $n \to \infty$:

$$\lim_{n\to\infty} a_n \;=\; \lim_{n\to\infty} \left(1 \,+\, \frac{1}{a_{n-1}}\right) \quad\Rightarrow\quad a \;=\; 1 \,+\, \frac{1}{a} \quad\Rightarrow\quad a^2 - a - 1 \;=\; 0 \quad\Rightarrow\quad a \;=\; \frac{1 \pm \sqrt{5}}{2}$$

Since a must be positive then $a = \frac{1+\sqrt{5}}{2}$. Thus, the sequence is convergent and converges to $\frac{1+\sqrt{5}}{2} \approx 1.618$. Note: This number is the famous *golden ratio*, the subject of many claims regarding its appearance in nature and supposed aesthetic appeal as a ratio of sides of a rectangle.[2]

---

[1] Due to the Italian mathematician Leonardo Fibonacci (ca. 1170-1250).
[2] For example, see HUNTLEY, H.E., *The Divine Proportion*, New York: Dover Publications, Inc., 1970.

An **infinite series** is the sum of an infinite sequence. If the infinite sequence is $\{a_n\}_{n=0}^{\infty}$ then the series can be written as

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots + a_n + \cdots$$

or simply as $\sum a_n$ when the initial value of the index n is understood. There is a natural way to define the sum of such a series:

The **sum** of an infinite series $\sum_{n=0}^{\infty} a_n$ is

$$\sum_{n=0}^{\infty} a_n = \lim_{n\to\infty} s_n \tag{9.4}$$

where $\{s_n\}_{n=0}^{\infty}$ is the sequence of **partial sums** of the series:

$$s_n = \sum_{k=0}^{n} a_k = a_0 + a_1 + a_2 + \cdots + a_n \quad \text{for } n \ge 0$$

If the partial sums $\{s_n\}$ converge to a real number s then the series is **convergent**, and **converges** to s; if $\{s_n\}$ diverges than the series is **divergent**.

One important convergent series is a **geometric progression**:

$$a + ar + ar^2 + ar^3 + \cdots + ar^n + \cdots \tag{9.5}$$

with $a \ne 0$ and $|r| < 1$. Multiply the n-th partial sum s_n by r:

$$rs_n = r(a + ar + ar^2 + ar^3 + \cdots + ar^{n-1} + ar^n) = ar + ar^2 + ar^3 + ar^4 + \cdots + ar^n + ar^{n+1}$$

Now subtract rs_n from s_n:

$$s_n - rs_n = a + ar + ar^2 + ar^3 + \cdots + ar^n - (ar + ar^2 + ar^3 + ar^4 + \cdots + ar^n + ar^{n+1})$$

$$s_n - rs_n = a - ar^{n+1} \quad \text{so that}$$

$$\lim_{n\to\infty} s_n = \lim_{n\to\infty} \frac{a(1-r^{n+1})}{1-r} = \frac{a}{1-r}$$

since $r^{n+1} \to 0$ as $n \to \infty$ when $|r| < 1$. Thus:

For $a \ne 0$ and $|r| < 1$ the geometric progression $\sum_{n=0}^{\infty} ar^n$ is convergent:

$$\sum_{n=0}^{\infty} ar^n = a + ar + ar^2 + ar^3 + \cdots + ar^n + \cdots = \frac{a}{1-r} \tag{9.6}$$

Example 9.5

Show that $\sum_{n=1}^{\infty} \frac{1}{2^n} = 1$.

Solution: This is a geometric progression with $a = \frac{1}{2}$ and $r = \frac{1}{2}$. So by formula (9.6) the sum is:

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right)^n = \frac{a}{1-r} = \frac{\frac{1}{2}}{1-\frac{1}{2}} = 1 \quad \checkmark$$

Example 9.6

Write the repeating decimal $0.\overline{17} = 0.17171717\ldots$ as a rational number.

Solution: This is a geometric progression with $a = 0.17 = \frac{17}{100}$ and $r = 0.01 = \frac{1}{100}$:

$$\begin{aligned} 0.171717\ldots &= 0.17 + 0.0017 + 0.000017 + \cdots = 0.17(0.01)^0 + 0.17(0.01)^1 + 0.17(0.01)^2 + \cdots \\ &= \sum_{n=0}^{\infty} 0.17(0.01)^n = \frac{a}{1-r} = \frac{0.17}{1-0.01} = \frac{0.17}{0.99} = \frac{17}{99} \end{aligned}$$

Back to Zeno's motion paradox, by Example 9.5 the sum of the infinite number of distances between successive markers in Figure 9.1.1 is 1, as expected. This fact is often mistaken as proof that Zeno was wrong, yet it does not actually address Zeno's argument, since the person is attempting to begin motion at the tail end—the "infinite end"—of the geometric progression, not at the beginning (i.e. at $n = 0$). Zeno's point remains that there is no "first step" at that tail end.

In fact, even if you were to reverse the person's position to start at the other end, so that he would first move a distance $\frac{1}{2}$, then a distance $\frac{1}{4}$, and so on, as in Figure 9.1.2, then a new problem is introduced: the person keeps moving no matter how close he is to the end point. There is now no "last step" and motion is thus still impossible.

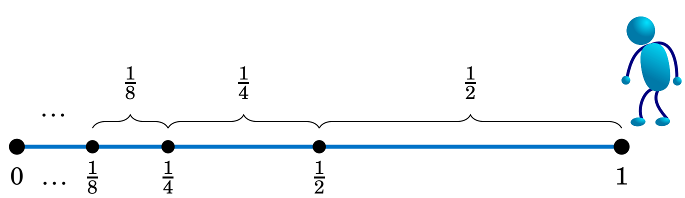


Figure 9.1.2 Zeno's motion paradox in reverse

The convergence of the geometric progression has misled many people to argue that Zeno is wrong, by claiming it shows an infinite number of movements can be completed in a finite amount of time. But this line of argument fails on (at least) two counts. First, Zeno never argued about time—it is irrelevant to his paradox. The more fundamental flaw is that the introduction of time brings along the concept of speed, typically taken to be constant (though it need not be). Speed is distance over time, but it is precisely the *distance* part of that ratio that Zeno rejects—that distance can never be traveled. In other words, it is circular—and hence faulty—reasoning to "prove" that motion is possible by *assuming* it is possible.

The geometric progression also doesn't help when considering only the distances and ignoring time. Even assuming each of those distances could be traveled, the partial sums *approach* 1 but never actually *reach* it. The limit of a sequence is defined in terms of an inequality—you can get arbitrarily close to the limit, and that is all. The equality in formula (9.6) is merely a shorthand way of saying that. It is an abstraction based on properties of the real number system, not necessarily based on physical reality.

Zeno's paradox is not purely mathematical—it is about space and hence is physical, with a tinge of philosophy. Physicists have recognized this—and noted the flaw in the purely mathematical line of attack—and devised new arguments against Zeno, some more sophisticated than others. However, they all invariably end up in some sort of circular reasoning. All this calls into question the original assumption: the infinite divisibility of space. If space had some smallest unit that could not be divided any further then there is no paradox—motion over a finite distance could always be decomposed into a large but finite number of irreducible steps.[3]

Exercises

A

For Exercises 1-8, determine if the given sequence is convergent. If so then find its limit.

1. $\{ne^{-n}\}_{n=0}^{\infty}$ **2.** $\left\{\dfrac{n^2}{3n^2+7n-2}\right\}_{n=1}^{\infty}$ **3.** $\left\{\dfrac{n^2}{3n^3+7n-2}\right\}_{n=1}^{\infty}$ **4.** $\left\{\dfrac{n^3}{3n^2+7n-2}\right\}_{n=1}^{\infty}$

5. $\left\{\dfrac{n}{\ln n}\right\}_{n=2}^{\infty}$ **6.** $\left\{\dfrac{n!}{(n+2)!}\right\}_{n=0}^{\infty}$ **7.** $\left\{\sin\left(\dfrac{n\pi}{2}\right)\right\}_{n=0}^{\infty}$ **8.** $\{(-1)^n\cos n\pi\}_{n=0}^{\infty}$

For Exercises 9-12 determine whether the given series is convergent. If so then find its sum.

9. $\displaystyle\sum_{n=0}^{\infty} 2\left(\frac{2}{3}\right)^n$ **10.** $\displaystyle\sum_{n=1}^{\infty} 7^{-n}$ **11.** $\displaystyle\sum_{n=0}^{\infty} \frac{3^n+5^{n+1}}{6^n}$ **12.** $\displaystyle\sum_{n=0}^{\infty} n$

For Exercises 13-16 use a geometric progression to write the given repeating decimal as a rational number.

13. $0.\overline{113}$ **14.** $0.\overline{9}$ **15.** $0.24\overline{9}$ **16.** $0.01\overline{7}$

B

17. In Example 9.4 define $b_n = F_n/F_{n+1}$ for $n \ge 0$ and show that $\{b_n\}_{n=0}^{\infty}$ converges to $\frac{\sqrt{5}-1}{2}$.

18. Show that formula (4.1) for Newton's method with $f(x) = x^2 - 2$ and $x_0 = 1$ yields the sequence $\{x_n\}_{n=0}^{\infty}$ with

$$x_n = \frac{1}{2}x_{n-1} + \frac{1}{x_{n-1}} \quad \text{for } n \ge 1,$$

then assuming $\{x_n\}$ is convergent show that it must converge to $\sqrt{2}$. *(Hint: See Example 9.4.)*

[3]As of this writing there is not yet a definitive answer as to whether space is continuous or discrete (quantized). A "smallest unit" would have to be below the *Planck level*—around 10^{-33} cm, well below current measurement capabilities. Some recent advances in the field of *loop quantum gravity* suggest the possibility of quantized space. See CHAMSEDDINE, A.H., CONNES, A. & MUKHANOV, V., "Geometry and the quantum: basics." *J. High Energy Phys.* **2014**, 98 (2014). `https://doi.org/10.1007/JHEP12(2014)098`

19. In this exercise you will prove a formula for the general number F_n in the Fibonacci sequence. Denote the positive and negative solutions to the equation $x^2 - x - 1 = 0$ by $\phi_+ = \frac{1+\sqrt{5}}{2}$ and $\phi_- = \frac{1-\sqrt{5}}{2}$, respectively. Note that ϕ_+ is the golden ratio mentioned in Example 9.4.

(a) Use the equation $x^2 - x - 1 = 0$ to show that

$$\phi_+^{n+1} = \phi_+^n + \phi_+^{n-1} \quad \text{and} \quad \phi_-^{n+1} = \phi_-^n + \phi_-^{n-1}.$$

(b) Use part (a) and induction to show that

$$F_n = \frac{\phi_+^n - \phi_-^n}{\sqrt{5}} \quad \text{for } n \ge 0.$$

20. A ball is dropped from a height of 4 ft above the ground, and upon each bounce off the ground the ball bounces straight up to a height equal to 65% of its previous height. Find the theoretical total distance the ball could travel if it could bounce indefinitely. Why is this physically unrealistic?

21. A *continued fraction* is a type of infinite sum involving a fraction with a denominator that continues indefinitely:

$$a = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}$$

(a) Show that the golden ratio $\phi_+ = \frac{1+\sqrt{5}}{2}$ (a solution of $x^2 - x - 1 = 0$) can be written as

$$\phi_+ = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cdots}}}.$$

(Hint: Look for a recurrence relation in the fraction.)

(b) Show that

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cdots}}}.$$

22. Show that the golden ratio $\phi_+ = \frac{1+\sqrt{5}}{2}$ can be written as $\phi_+ = \sqrt{1 + \sqrt{1 + \sqrt{1 + \cdots}}}$.

23. Write a computer program to approximate the result in Exercise 22 using 100 terms. After how many iterations do the approximate values start repeating?

24. Would the existence of infinitesimals as a measure of space resolve Zeno's paradox? Explain.

9.2 Tests for Convergence

There are many ways to determine if a sequence converges—two are listed below. In all cases changing or removing a finite number of terms in a sequence does not affect its convergence or divergence:

1. **Monotone Bounded Test**: A sequence that is bounded and *monotone*—i.e. either always increasing or always decreasing—is convergent.
2. **Comparison Test**: If $\{b_n\}$ diverges to ∞ and if $\{a_n\}$ is a sequence such that $a_n \ge b_n$ for all $n > N$ for some N, then $\{a_n\}$ also diverges to ∞.
   Likewise, if $\{c_n\}$ diverges to $-\infty$ and $a_n \le c_n$ for all $n > N$ for some N, then $\{a_n\}$ also diverges to $-\infty$.

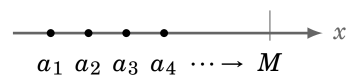


Figure 9.2.1

The Comparison Test makes sense intuitively, since something larger than a quantity going to infinity must also go to infinity. The Monotone Bounded Test can be understood by thinking of a bound on a sequence as a wall that the sequence can never pass, as in Figure 9.2.1. The increasing sequence $\{a_n\}$ in the figure moves toward M but can never pass it. The sequence thus cannot diverge to ∞, and it cannot fluctuate back and forth since it always increases. Thus it must converge somewhere before or at M.[4] Notice that the Monotone Bounded Test tells you only that the sequence converges, not what it converges to.

Example 9.7

Show that the sequence $\{a_n\}_{n=1}^{\infty}$ defined for $n \ge 1$ by

$$a_n = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2 \cdot 4 \cdot 6 \cdots (2n)}$$

is convergent.

Solution: Notice that $\{a_n\}$ is always decreasing, since

$$a_{n+1} = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1) \cdot (2n+1)}{2 \cdot 4 \cdot 6 \cdots (2n) \cdot (2n+2)} = a_n \cdot \frac{2n+1}{2n+2} < a_n \cdot (1) = a_n$$

for $n \ge 1$. The sequence is also bounded, since $0 < a_n$ and

$$a_n = \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{5}{6} \cdots \frac{2n-1}{2n} < 1 \quad \text{for } n \ge 1$$

since each fraction in the above product is less than 1. Thus, by the Monotone Bounded Test the sequence is convergent.

Note that for a decreasing sequence only the lower bound is needed for the Monotone Bounded Test, not the upper bound. Similarly, for an increasing sequence only the upper bound matters.

[4] For a proof see pp.48-49 in BUCK, R.C., *Advanced Calculus*, 2nd ed., New York: McGraw-Hill Book Co., 1965.

Some tests for convergence of a series are listed below:

1. **n-th Term Test**: If $\sum a_n$ converges then $\lim_{n\to\infty} a_n = 0$.

2. **Ratio Test**: For a series $\sum a_n$ of positive terms let

$$R = \lim_{n\to\infty} \frac{a_{n+1}}{a_n}\,.$$

   Then

   (a) if $R < 1$ then the series converges,

   (b) if $R > 1$ (including $R = \infty$) then the series diverges,

   (c) if $R = 1$ then the test fails.

3. **Integral Test**: For a series $\sum_{n=1}^{\infty} a_n$ of positive terms let $f(x)$ be a decreasing function on $[1,\infty)$ such that $f(n) = a_n$ for all integers $n \ge 1$. Then

$$\sum_{n=1}^{\infty} a_n \quad \text{and} \quad \int_1^{\infty} f(x)\,dx$$

   either both converge or both diverge.

4. **p-series Test**: The series $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges for $p > 1$, and diverges for $p \le 1$.

5. **Comparison Test**: If $0 \le a_n \le b_n$ for $n > N$ for some N, and if $\sum b_n$ is convergent then $\sum a_n$ is convergent. Similarly, if $0 \le b_n \le a_n$ for $n > N$ for some N, and if $\sum b_n$ is divergent then $\sum a_n$ is divergent.

6. **Limit Comparison Test**: For two series $\sum a_n$ and $\sum b_n$ of positive terms let

$$L = \lim_{n\to\infty} \frac{a_n}{b_n}\,.$$

   Then

   (a) if $0 < L < \infty$ then $\sum a_n$ and $\sum b_n$ either both converge or both diverge,

   (b) if $L = 0$ and $\sum b_n$ converges then $\sum a_n$ converges,

   (c) if $L = \infty$ and $\sum b_n$ diverges then $\sum a_n$ diverges.

7. **Telescoping Series Test**: Suppose $\sum a_n = \sum (b_n - b_{n+1})$ for some sequence $\{b_n\}$. Then $\sum a_n$ converges if and only if $b_n \to L$, in which case $\sum a_n = b_1 - L$.

Most of the above tests have fairly short proofs or at least intuitive explanations. For example, the n-th Term Test follows from the definition of convergence of a series: if $\sum a_n$ converges to a number L then since each term $a_n = s_n - s_{n-1}$ is the difference of successive partial sums, taking the limit yields

$$\lim_{n\to\infty} a_n \;=\; \lim_{n\to\infty} (s_n - s_{n-1}) \;=\; L - L \;=\; 0$$

by definition of the convergence of a series. ✓
Since the n-th Term Test can never be used to prove convergence of a series, it is often stated in the following logically equivalent manner:

> **n-th Term Test**: If $\lim\limits_{n\to\infty} a_n \neq 0$ then $\sum a_n$ diverges.

Example 9.8

Show that $\displaystyle\sum_{n=1}^{\infty} \frac{n}{2n+1} = \frac{1}{3} + \frac{2}{5} + \frac{3}{7} + \cdots$ is divergent.

Solution: Since

$$\lim_{n\to\infty} \frac{n}{2n+1} \;=\; \frac{1}{2} \neq 0$$

then by the n-th Term Test the series diverges.

The Ratio Test takes a bit more effort to prove.[5] When the ratio R in the Ratio Test is larger than 1 then that means the terms in the series do not approach 0, and thus the series diverges by the n-th Term Test. When $R = 1$ the test fails, meaning it is inconclusive—another test would need to be used. When the test shows convergence it does not tell you what the series converges to, merely that it converges.

Example 9.9

Determine if $\displaystyle\sum_{n=1}^{\infty} \frac{n}{2^n}$ is convergent.

Solution: For the series general term $a_n = \frac{n}{2^n}$,

$$R \;=\; \lim_{n\to\infty} \frac{a_{n+1}}{a_n} \;=\; \lim_{n\to\infty} \frac{\frac{n+1}{2^{n+1}}}{\frac{n}{2^n}} \;=\; \lim_{n\to\infty} \frac{n+1}{2n} \;=\; \frac{1}{2} \;<\; 1\,,$$

so by the Ratio Test the series converges.

---

[5]See pp.612-613 in TAYLOR, A.E. AND W.R. MANN, *Advanced Calculus*, 2nd ed., New York: John Wiley & Sons, Inc., 1972.

Figure 9.2.2 shows why the Integral Test works.

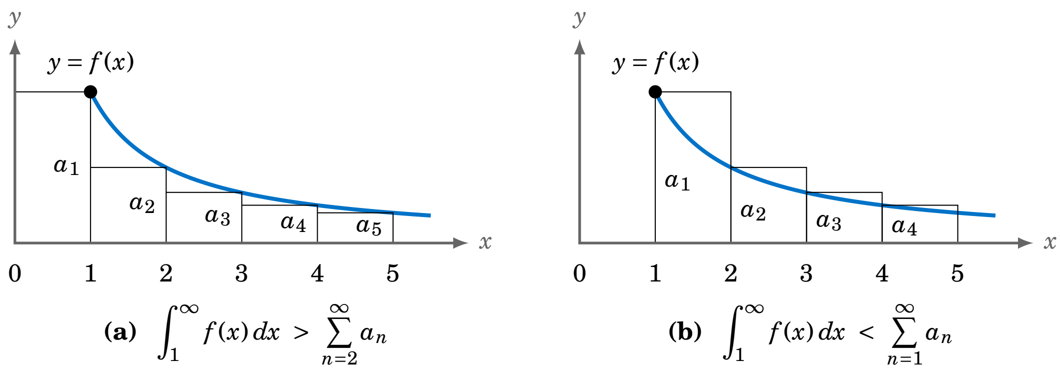


Figure 9.2.2 Integral Test

In Figure 9.2.2(a) the area $\int_1^{\infty} f(x)\,dx$ is greater than the total area S of all the rectangles under the curve. Since each rectangle has height a_n and width 1, then $S = \sum_2^{\infty} a_n$. Thus, since removing the single term a_1 from the series does not affect the convergence or divergence of the series, the series converges if the improper integral converges, and conversely the integral diverges if the series diverges. Similarly, in Figure 9.2.2(b) the area $\int_1^{\infty} f(x)\,dx$ is less than the total area $S = \sum_1^{\infty} a_n$ of all the rectangles, so the integral converges if the series converges, and the series diverges if the integral diverges. Notice how in both graphs the rectangles are either all below the curve or all protrude above the curve due to $f(x)$ being a decreasing function.

Example 9.10

Show that the *p-series* $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges for $p > 1$ and diverges for $p = 1$.

Solution: For $p \ge 1$ let $f(x) = \frac{1}{x^p}$ on $[1,\infty)$. Then $f(x)$ is decreasing, and $f(n) = a_n = \frac{1}{n^p}$ for all integers $n \ge 1$. For $p > 1$,

$$\int_1^{\infty} f(x)\,dx = \int_1^{\infty} \frac{dx}{x^p} = \left.\frac{-1}{(p-1)x^{p-1}}\right|_1^{\infty} = 0 - \left(\frac{-1}{(p-1)(1)^{p-1}}\right) = \frac{1}{p-1} < \infty$$

so the integral converges. Thus, by the Integral Test, the series converges for $p > 1$. For $p = 1$,

$$\int_1^{\infty} f(x)\,dx = \int_1^{\infty} \frac{dx}{x} = \ln x\,\Big|_1^{\infty} = \infty$$

so the integral diverges. So by the Integral Test, the series diverges for $p = 1$. The **harmonic series**

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

thus diverges even though $a_n = \frac{1}{n} \to 0$ (which is hence not a sufficient condition for a series to converge). Note that this example partly proves the p-series Test. The remaining case ($p < 1$) is left as an exercise.

The divergence part of the Comparison Test is clear enough to understand, but for the convergence part with $0 \le a_n \le b_n$ for all n larger than some N, ignore the (finite) number of terms before a_N and b_N. Since $\sum b_n$ converges then its partial sums must be bounded. The partial sums for $\sum a_n$ then must also be bounded, since $0 \le a_n \le b_n$ for $n > N$. So since $a_n \ge 0$ means that the partial sums for $\sum a_n$ are increasing, by the Monotone Bounded Test the partial sums for $\sum a_n$ must converge, i.e. $\sum a_n$ is convergent.

Example 9.11

Determine if $\sum\limits_{n=1}^{\infty} \frac{1}{n^n}$ is convergent.

Solution: Since $n^n \ge n^2 > 0$ for $n > 2$, then

$$0 < \frac{1}{n^n} \le \frac{1}{n^2}$$

for $n > 2$. Thus, since $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges (by the p-series Test with $p = 2 > 1$, as in Example 9.10), the series $\sum_{n=1}^{\infty} \frac{1}{n^n}$ converges by the Comparison Test.

For the Limit Comparison Test with $\frac{a_n}{b_n} \to L < \infty$ and $L > 0$, by definition of the limit of a sequence, $\frac{a_n}{b_n}$ can be made arbitrarily close to L. In particular there is an integer N such that

$$\frac{L}{2} < \frac{a_n}{b_n} < \frac{3L}{2}$$

for all $n > N$. Then

$$0 < a_n < \frac{3L}{2} b_n \quad \text{and} \quad \sum b_n \text{ converges} \quad \Rightarrow \quad \sum a_n \text{ converges}$$

by the Comparison test. Likewise,

$$0 < \frac{L}{2} b_n < a_n \quad \text{and} \quad \sum b_n \text{ diverges} \quad \Rightarrow \quad \sum a_n \text{ diverges}$$

by the Comparison Test again. The cases $L = 0$ and $L = \infty$ are handled similarly.

Example 9.12

Determine if $\sum\limits_{n=1}^{\infty} \frac{n+3}{n \cdot 2^n}$ is convergent.

Solution: Since $\sum_{n=1}^{\infty} \frac{1}{2^n}$ is convergent (as part of a geometric progression) and

$$\lim_{n\to\infty} \frac{(n+3)/(n \cdot 2^n)}{1/2^n} = \lim_{n\to\infty} \frac{n+3}{n} = 1$$

then by the Limit Comparison Test $\sum_{n=1}^{\infty} \frac{n+3}{n \cdot 2^n}$ is convergent..

A series $\sum a_n$ is **telescoping** if $a_n = b_n - b_{n+1}$ for some sequence $\{b_n\}$. Assume the series $\sum a_n$ and sequence $\{b_n\}$ both start at $n = 1$. Then the partial sum s_n for $\sum a_n$ is

$$s_n = a_1 + a_2 + \cdots + a_n = (b_1 - b_2) + (b_2 - b_3) + \cdots + (b_n - b_{n+1}) = b_1 - b_{n+1}$$

for $n \geq 1$. Thus, since b_1 is a fixed number, $\lim_{n\to\infty} s_n$ exists if and only if $\lim_{n\to\infty} b_{n+1}$ exists, i.e. $\sum a_n$ converges if and only if $\{b_n\}$ converges. So if $b_n \to L$ then $s_n \to b - L$, i.e. $\sum a_n$ converges to L, which proves the Telescoping Series Test. Note that the number b_1, as the first number in the sequence $\{b_n\}$, could be replaced by whatever the first number is, in case the index n starts at a number different from 1.

Example 9.13

Determine if $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$ is convergent. If it converges then can you find its sum?

Solution: For the sequence $\{b_n\}$ with $b_n = \frac{1}{n}$ for $n \geq 1$, each term in the series can be written as

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1} = b_n - b_{n+1}$$

Thus, since $\{b_n\}$ converges to 0, by the Telescoping Series Test the series also converges, to $b_1 - 0 = 1$.

Convergent series have the following properties (based on similar properties of limits):

> Let $\sum a_n$ and $\sum b_n$ be convergent series, and let c be a number. Then:
> **(a)** $\sum(a_n \pm b_n)$ is convergent, with $\sum(a_n \pm b_n) = \sum a_n \pm \sum b_n$
>
> **(b)** $\sum ca_n$ is convergent, with $\sum ca_n = c \cdot \sum a_n$

Exercises

A

For Exercises 1-5 show that the given sequence $\{a_n\}_{n=1}^{\infty}$ is convergent.

1. $a_n = \dfrac{2 \cdot 4 \cdot 6 \cdots (2n)}{3 \cdot 5 \cdot 7 \cdots (2n+1)}$ **2.** $a_n = 1 - \dfrac{2^n}{n!}$ **3.** $a_n = \dfrac{2 \cdot 4 \cdot 6 \cdots (2n)}{1 \cdot 3 \cdot 5 \cdots (2n-1)} \cdot \dfrac{1}{2n+2}$

4. $a_n = \dfrac{1}{n}\left(\dfrac{2 \cdot 4 \cdot 6 \cdots (2n)}{1 \cdot 3 \cdot 5 \cdots (2n-1)}\right)^2$ **5.** $a_n = \dfrac{1}{2} \cdot \dfrac{3}{2} \cdot \dfrac{3}{4} \cdot \dfrac{5}{4} \cdot \dfrac{5}{6} \cdot \dfrac{7}{6} \cdots \dfrac{2n-1}{2n} \cdot \dfrac{2n+1}{2n}$

For Exercises 6-17 determine whether the given series is convergent.

6. $\displaystyle\sum_{n=0}^{\infty} \sin\left(\frac{n\pi}{2}\right)$ **7.** $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n(n+2)}$ **8.** $\displaystyle\sum_{n=1}^{\infty} \frac{1}{2n}$ **9.** $\displaystyle\sum_{n=1}^{\infty} \frac{n}{(n+1)2^n}$

10. $\sum_{n=2}^{\infty} \frac{1}{n\sqrt{\ln n}}$ **11.** $\sum_{n=1}^{\infty} \frac{n!}{(2n)!}$ **12.** $\sum_{n=1}^{\infty} \frac{n}{e^n}$ **13.** $\sum_{n=1}^{\infty} \frac{1}{\cosh^2 n}$

14. $\sum_{n=1}^{\infty} \frac{n!}{2^n}$ **15.** $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}$ **16.** $\sum_{n=1}^{\infty} \frac{1}{n(2n-1)}$ **17.** $\sum_{n=1}^{\infty} \frac{\ln(n+1)}{n^2}$

For Exercises 18-21 determine whether the given series is convergent. If convergent then find its sum.

18. $\sum_{n=1}^{\infty} \frac{1}{(2n+1)(2n+3)}$ **19.** $\sum_{n=1}^{\infty} \frac{1}{(2n+3)(2n+5)}$ **20.** $\sum_{n=1}^{\infty} \frac{2}{(3n+1)(3n+4)}$ **21.** $\sum_{n=1}^{\infty} \frac{1}{4n^2-1}$

B

22. Continue Example 9.10 with a proof of the p-series Test for $p < 1$.

23. Show that $\{a_n\}_{n=1}^{\infty}$ is convergent, where

$$a_n = \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots + \frac{1}{n!}$$

for $n \geq 1$. *(Hint: Use the Monotone Bounded test by using a bound on $\frac{1}{n!}$ for $n > 2$.)*

24. Consider the series $\sum_{n=1}^{\infty} \frac{1}{2n-1} = 1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots$.

(a) Show that the series is divergent.

(b) The textbook *Applied Mathematics for Physical Chemistry* (3rd ed.), J. Barrante, provides the following argument that the above series converges: Since

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots \quad < \quad 1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots \quad < \quad 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

where the series on the left converges (by the p-series Test with $p = 2$) and the series on the right diverges (by the p-series Test with $p = 1$), and since each term in the middle series is between its corresponding terms in the left series and right series, then there must be a p-series for some value $1 < p < 2$ such that each term in the middle series is less than the corresponding term in that p-series. That is,

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots \quad < \quad 1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots \quad < \quad 1 + \frac{1}{2^p} + \frac{1}{3^p} + \frac{1}{4^p} + \cdots$$

for that value of p between 1 and 2. But $p > 1$ for that p-series on the right, so it converges, which means that the middle series converges! Find and explain the flaw in this argument.

25. *Wallis' formula*[6] for π is given by the infinite product

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdots \frac{2n}{2n-1} \cdot \frac{2n}{2n+1} \cdots .$$

Notice that this is the limit of the reciprocal of the sequence in Exercise 5. Write a computer program to approximate the limit using 1 million iterations. How close is your approximation to $\frac{\pi}{2}$?

[6] Due to the English mathematician and theologian John Wallis (1616-1703). For a proof of the formula see pp.738-739 in TAYLOR, A.E. AND W.R. MANN, *Advanced Calculus*, 2nd ed., New York: John Wiley & Sons, Inc., 1972.

9.3 Alternating Series

In the last section the harmonic series

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots$$

was shown to diverge. If you were to alternate the signs of successive terms, as in

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots \tag{9.7}$$

then it turns out that this new series—called an **alternating series**—converges, due to the following test:

> **Alternating Series Test**: If $\sum a_n$ is an **alternating series**—i.e. the signs of the terms a_n alternate between positive and negative—such that the absolute values of the terms are decreasing to 0, then the series converges.

The condition for the test means that $|a_{n+1}| \le |a_n|$ for all n and $|a_n| \to 0$ as $n \to \infty$. To see why the test works, consider the alternating series given above by formula (9.7), with $a_n = \frac{-1^{n-1}}{n}$. The odd-numbered partial sums $s_1, s_3, s_5, \ldots$, can be written as

$$s_1 = 1 \quad , \quad s_3 = 1 - \underbrace{\left(\frac{1}{2} - \frac{1}{3}\right)}_{>0} \quad , \quad s_5 = 1 - \underbrace{\left(\frac{1}{2} - \frac{1}{3}\right)}_{>0} - \underbrace{\left(\frac{1}{4} - \frac{1}{5}\right)}_{>0} \quad , \quad \ldots$$

while the even-numbered partial sums $s_2, s_4, s_6, \ldots$, can be written as

$$s_2 = 1 - \frac{1}{2} \quad , \quad s_4 = 1 - \frac{1}{2} + \underbrace{\left(\frac{1}{3} - \frac{1}{4}\right)}_{>0} \quad , \quad s_6 = 1 - \frac{1}{2} + \underbrace{\left(\frac{1}{3} - \frac{1}{4}\right)}_{>0} + \underbrace{\left(\frac{1}{5} - \frac{1}{6}\right)}_{>0} \quad , \quad \ldots$$

Thus the odd-numbered partial sums $\{s_{2n-1}\}$ are decreasing from $s_1 = 1$, and the even-numbered ones $\{s_{2n}\}$ are increasing from $s_2 = 1 - \frac{1}{2} = \frac{1}{2}$, with $\frac{1}{2} < s_n < 1$ for all n, i.e. the partial sums are bounded. So by the Monotone Bounded Test, both sequences $\{s_{2n-1}\}$ and $\{s_{2n}\}$ must converge. Since $s_{2n} - s_{2n-1} = a_{2n} = \frac{-1}{2n}$ for all $n \ge 1$, then

$$\lim_{n\to\infty} (s_{2n} - s_{2n-1}) = \lim_{n\to\infty} \frac{-1}{2n} = 0 \quad \Rightarrow \quad \lim_{n\to\infty} s_{2n} = \lim_{n\to\infty} s_{2n+1}$$

Thus, the partial sums s_n have a common limit, so the series converges. Notice that the key to the convergence was having the terms decreasing in absolute value to zero.

Example 9.14

Determine if $\sum_{n=2}^{\infty} \frac{(-1)^n}{\ln n}$ is convergent.

Solution: For the general term $a_n = \frac{(-1)^n}{\ln n}$, since $\ln(n+1) > \ln n$ for $n \ge 2$ and $\ln n \to \infty$ as $n \to \infty$, then $|a_n|$ decreases to 0 as $n \to \infty$. Thus, by the Alternating Series Test the series converges.

The series $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}$ converges by the Alternating Series Test, though the series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges. This makes $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}$ an example of a **conditionally convergent** series:

A series $\sum a_n$ is **conditionally convergent** if $\sum a_n$ converges but $\sum |a_n|$ diverges. If $\sum |a_n|$ converges then $\sum a_n$ is **absolutely convergent**.

For example, $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}$ is not absolutely convergent, since $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges.

Example 9.15

Is $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^2}$ conditionally convergent or absolutely convergent?

Solution: Since $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges (by the p-series Test) then $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^2}$ is absolutely convergent.

It turns out that absolute convergence implies ordinary convergence:

Absolute Convergence Test: If $\sum |a_n|$ converges then $\sum a_n$ converges.

The test is obvious if the terms a_n are all positive, so assume the series has both positive terms (denoted by the sequence $\{a_{\text{pos}}\}$) and negative terms (denoted by $\{a_{\text{neg}}\}$). Then the series can be decomposed into the difference of two series:

$$\sum a_n = \sum a_{\text{pos}} - \sum |a_{\text{neg}}|$$

Since each of the sums on the right side of the equation is part of the convergent series $\sum |a_n|$, then each sum itself converges (being part of a finite sum). Thus their difference, namely $\sum a_n$, is finite, i.e. $\sum a_n$ converges.

The test can be stated in the following logically equivalent manner:

Absolute Convergence Test: If $\sum a_n$ diverges then $\sum |a_n|$ diverges.

One unusual feature of a conditionally convergent series is that its terms can be rearranged to converge to any number, a result known as **Riemann's Rearrangement Theorem**. For example, the alternating harmonic series

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots$$

consists of one divergent series of positive terms subtracted from another series of positive terms, namely:

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots = \left(1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots\right) - \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{6} + \cdots\right)$$

The idea is that since the first series on the right diverges then that series has some partial sum that could be made just larger than any positive number A. Likewise, since the second series on the right that is being subtracted also diverges, it has some partial sum that when subtracted from the first partial sum results in a number just less than A. Continue like this indefinitely, first adding a partial sum to get a number just bigger than A then subtracting another partial sum to get just less than A. Since the terms in each series approach zero, the overall series can be made to converge to A! It also turns out that an absolutely convergent series does *not* have this feature—any rearrangement of terms results in the same sum.[7]

Exercises

A

For Exercises 1-5 determine whether the given alternating series is convergent. If convergent, then determine if it is conditionally convergent or absolutely convergent.

1. $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{\sqrt{n}}$ **2.** $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n-1}\,n!}{2^n}$ **3.** $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{2n-1}$ **4.** $\displaystyle\sum_{n=2}^{\infty} \frac{(-1)^{n}}{n \ln n}$ **5.** $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n!}$

6. Rearrangements of a divergent alternating series can make it appear to converge to different numbers. For example, find different rearrangements of the terms in the divergent series

$$\sum_{n=0}^{\infty} (-1)^n = 1 - 1 + 1 - 1 + \cdots$$

so that the series appears to converge to 0, 1, -1, and 2.

7. A guest arrives at the Aleph Null Hotel,[8] which has an infinite number of rooms, numbered Room 0, Room 1, Room 2, and so on. The hotel manager says all the rooms are taken, but he can still give the guest his own room. How is that possible? Would it still be possible if an infinite number of guests showed up, each wanting his own room? Explain your answers.

[7] For formal proofs of these statements, see pp.442-444 in KLAMBAUER, G., *Aspects of Calculus*, New York: Springer-Verlag, 1986.

[8] The symbol $\aleph_0$ is called *aleph null* and represents the *cardinality* of $\mathbb{N}$, i.e. its size. $\aleph_0$ is the smallest infinity.

9.4 Power Series

A **power series** is an infinite series whose terms involve constants a_n and powers of $x-c$, where x is a variable and c is a constant: $\sum a_n (x-c)^n$. In many cases c will be 0. For example, the geometric progression

$$\sum_{n=0}^{\infty} r^n = 1 + r + r^2 + r^3 + \cdots = \frac{1}{1-r}$$

converges when $|r| < 1$, i.e. for $-1 < r < 1$, as shown in Section 9.1. Replacing the constant r by a variable x yields the power series

$$\sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x} \tag{9.8}$$

that converges to $\frac{1}{1-x}$ when $-1 < x < 1$. Note that the series diverges for $|x| \geq 1$, by the n-th Term Test.

In general a power series of the form $\sum f_n(x)$, where $f_n(x) = a_n(x-c)^n$ is a sequence of functions, has an **interval of convergence** defined as the set of all x such that the series converges. The interval can be any combination of open or closed, as well as the extreme cases of a single point or all real numbers. On its interval of convergence the power series is thus a function of x. The **radius of convergence** R of a power series is defined as half the length of the interval of convergence. In the case where the interval of convergence is all of $\mathbb{R}$ you would say $R = \infty$.

For example, for the above power series $\sum_{n=0}^{\infty} f_n(x)$, where $f_n(x) = x^n$ for $n \geq 0$, the interval of convergence is $-1 < x < 1$, so the radius of convergence is $R = 1$. Notice that

$$f(x) = \sum_{n=0}^{\infty} x^n \quad \text{for } -1 < x < 1$$

is thus a well-defined function on the interval $(-1, 1)$, where it happens to equal $\frac{1}{1-x}$. This power series can be thought of as a polynomial of infinite degree.

To find the interval of convergence of a power series $\sum f_n(x)$, you typically would use the Ratio Test on the absolute values of the terms (since the Ratio Test requires positive terms):

$$r(x) = \lim_{n\to\infty} \left| \frac{f_{n+1}(x)}{f_n(x)} \right| \tag{9.9}$$

Note that the limit $r(x)$ in this case is a function of x. When taking the limit, though, treat x as fixed. By the Ratio Test the power series will then converge for all x such that $r(x) < 1$, and diverge when $r(x) > 1$. When $r(x) = 1$ the test is inconclusive, so you would have to check those cases individually to see if those values of x should be added to the interval of convergence (along with the points where $r(x) < 1$).

Example 9.16

Find the interval of convergence of the power series $\sum_{n=0}^{\infty} \frac{x^n}{n!}$.

Solution: For $f_n(x) = \frac{x^n}{n!}$,

$$r(x) = \lim_{n\to\infty} \left| \frac{f_{n+1}(x)}{f_n(x)} \right| = \lim_{n\to\infty} \left| \frac{x^{n+1}/(n+1)!}{x^n/n!} \right| = |x| \cdot \lim_{n\to\infty} \left| \frac{1}{n+1} \right| = |x| \cdot 0 = 0$$

for any fixed x. Thus, $r(x) = 0 < 1$ for all x, so the interval of convergence is all of $\mathbb{R}$, i.e. $-\infty < x < \infty$.

Example 9.17

Find the interval of convergence of the power series $\sum_{n=1}^{\infty} \frac{x^n}{n}$.

Solution: For $f_n(x) = \frac{x^n}{n}$,

$$r(x) = \lim_{n\to\infty} \left| \frac{f_{n+1}(x)}{f_n(x)} \right| = \lim_{n\to\infty} \left| \frac{x^{n+1}/(n+1)}{x^n/n} \right| = \lim_{n\to\infty} \left| \frac{nx}{n+1} \right| = |x| \cdot \lim_{n\to\infty} \left| \frac{n}{n+1} \right| = |x| \cdot 1 = |x|$$

for any fixed x. Thus, the series converges when $r(x) = |x| < 1$ and diverges when $r(x) = |x| > 1$. The cases $r(x) = |x| = 1$ need to be checked individually. For $x = 1$ the series is $\sum_{n=1}^{\infty} \frac{1}{n}$, which diverges. For $x = -1$ the series is $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}$, which converges. Thus, the interval of convergence is $-1 \le x < 1$.

Example 9.18

Find the interval of convergence of the power series $\sum_{n=0}^{\infty} n!\, x^n$.

Solution: For $f_n(x) = n!\, x^n$,

$$r(x) = \lim_{n\to\infty} \left| \frac{f_{n+1}(x)}{f_n(x)} \right| = \lim_{n\to\infty} \left| \frac{(n+1)!\, x^{n+1}}{n!\, x^n} \right| = |x| \cdot \lim_{n\to\infty} |n+1| = \begin{cases} 0 & \text{if } x = 0, \\ \infty & \text{if } x \neq 0. \end{cases}$$

Thus, $r(x) = \infty > 1$ for all $x \neq 0$, so the series diverges for $x \neq 0$. So since $r(x) = 0 < 1$ only for $x = 0$, the interval of convergence is the single point $x = 0$.

It turns out that power series can be both differentiated and integrated term by term:[9]

For a power series $f(x) = \sum_{n=0}^{\infty} a_n (x-c)^n$ that converges for $|x-c| < R$, both

$$f'(x) = \sum_{n=1}^{\infty} n\, a_n (x-c)^{n-1} \quad \text{and} \quad \int f(x)\, dx = C + \sum_{n=0}^{\infty} \frac{a_n}{n+1} (x-c)^{n+1} \tag{9.10}$$

converge for $|x-c| < R$.

[9]The proofs require *uniform convergence*, a stronger condition than ordinary convergence. See pp.129-134 in BROMWICH, T.J., *An Introduction to the Theory of Infinite Series*, 2nd ed., London: Macmillan & Co. Ltd., 1955.

Notice that the above statement says nothing about the convergence of $f'(x)$ or $\int f(x)\,dx$ at the endpoints of the interval $|x-c| < R$. In each case convergence at the endpoints can be checked individually.

Example 9.19

Write the power series form of the derivative of $f(x) = \sum_{n=0}^{\infty} x^n$ and find its interval of convergence. Can $f'(x)$ be written in a non-series form?.

Solution: Differentiate $f(x)$ term by term:

$$f'(x) = \frac{d}{dx}\left(\sum_{n=0}^{\infty} x^n\right) = \frac{d}{dx}(1 + x + x^2 + x^3 + \cdots) = 0 + 1 + 2x + 3x^2 + \cdots = \sum_{n=1}^{\infty} nx^{n-1}$$

Since $f(x)$ converges for $-1 < x < 1$ then so does $f'(x)$. Checking the endpoints, at $x = 1$ and $x = -1$ the series for $f'(x)$ are $\sum_{n=1}^{\infty} n$ and $\sum_{n=1}^{\infty}(-1)^{n-1} n$, respectively, both of which diverge by the n-th Term Test. Thus, the interval of convergence for $f'(x)$ is $-1 < x < 1$.

Since $f(x) = \frac{1}{1-x}$ for $-1 < x < 1$ then $f'(x) = \frac{1}{(1-x)^2}$ for $-1 < x < 1$. Thus,

$$\sum_{n=1}^{\infty} nx^{n-1} = 1 + 2x + 3x^2 + \cdots = \frac{1}{(1-x)^2} \quad \text{for } -1 < x < 1.$$

Bessel Functions

Many applications in engineering and physics—especially those involving oscillations or mechanical vibrations—require solving differential equations of the form

$$\frac{d^2y}{dx^2} + \frac{1}{x}\frac{dy}{dx} + y = 0 \ , \tag{9.11}$$

known as **Bessel's equation**. This equation has a solution $J_0(x)$, known as **Bessel's function of order zero**,[10] defined in terms of a power series:

$$J_0(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(n!)^2 \cdot 2^{2n}} = 1 - \frac{x^2}{2^2} + \frac{x^4}{2^2 \cdot 4^2} - \frac{x^6}{2^2 \cdot 4^2 \cdot 6^2} + \cdots \tag{9.12}$$

The Ratio Test shows that $J_0(x)$ converges for all x, since for any fixed x,

$$r(x) = \lim_{n\to\infty} \left| \frac{(-1)^{n+1} \dfrac{x^{2n+2}}{((n+1)!)^2 \cdot 2^{2n+2}}}{(-1)^n \dfrac{x^{2n}}{(n!)^2 \cdot 2^{2n}}} \right| = x^2 \cdot \lim_{n\to\infty} \left| \frac{1}{4(n+1)^2} \right| = 0 < 1 \ .$$

[10]Due to the German astronomer Friedrich Wilhelm Bessel (1784-1846), from a study of elliptic planetary motion. For some historical background and examples of how Bessel's equation arises in applications, see § I.2 and § I.8 in McLachlan, N.W., *Bessel Functions for Engineers*, London: Oxford University Press, 1934.

The **general Bessel equation of order *m***,

$$\frac{d^2y}{dx^2} + \frac{1}{x}\frac{dy}{dx} + \left(1 - \frac{m^2}{x^2}\right) y = 0 \tag{9.13}$$

for $m = 0, 1, 2, \ldots$, has a solution $J_m(x)$, called **Bessel's function of order *m***:

$$J_m(x) = \sum_{n=0}^{\infty} (-1)^n \frac{1}{n!\cdot(n+m)!} \left(\frac{x}{2}\right)^{2n+m} \tag{9.14}$$

For example, the Bessel function $J_1(x)$ of order 1 is

$$J_1(x) = \sum_{n=0}^{\infty} (-1)^n \frac{1}{n!\cdot(n+1)!} \left(\frac{x}{2}\right)^{2n+1} = \frac{x}{2} - \frac{x^3}{2^2\cdot 4} + \frac{x^5}{2^2\cdot 4^2\cdot 6} - \frac{x^7}{2^2\cdot 4^2\cdot 6^2\cdot 8} + \cdots$$

Term by term differentiation shows that $J_0'(x) = -J_1(x)$:

$$\begin{aligned}\frac{d}{dx}(J_0(x)) &= \frac{d}{dx}\left(1 - \frac{x^2}{2^2} + \frac{x^4}{2^2\cdot 4^2} - \frac{x^6}{2^2\cdot 4^2\cdot 6^2} + \frac{x^8}{2^2\cdot 4^2\cdot 6^2\cdot 8^2} - \cdots\right)\\ &= -\frac{x}{2} + \frac{x^3}{2^2\cdot 4} - \frac{x^5}{2^2\cdot 4^2\cdot 6} + \frac{x^7}{2^2\cdot 4^2\cdot 6^2\cdot 8} - \cdots = -J_1(x)\end{aligned}$$

Graphs of $J_0(x)$ and $J_1(x)$ are shown in Figure 9.4.1 below. As you can see, $J_0(x)$ and $J_1(x)$ behave as sort of "poor man's" cosine and sine functions, respectively.

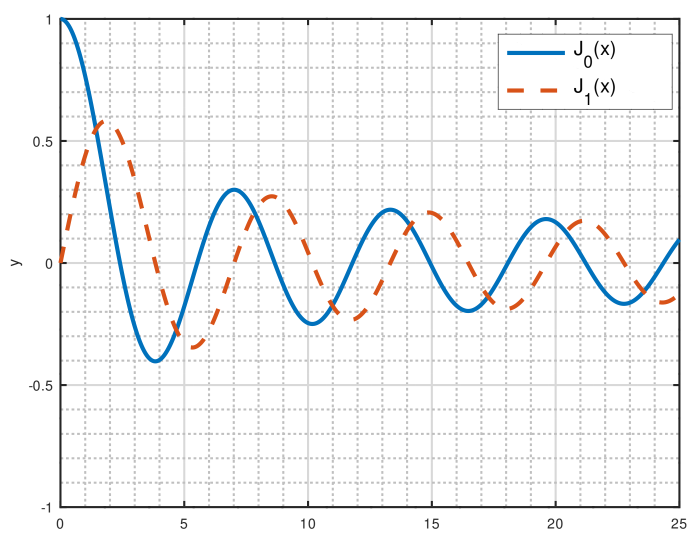


Figure 9.4.1 Bessel functions $J_0(x)$ and $J_1(x)$

Exercises

A

For Exercises 1-8 find the interval of convergence of the given power series.

1. $\displaystyle\sum_{n=1}^{\infty} \frac{nx^n}{(n+1)^2}$ **2.** $\displaystyle\sum_{n=1}^{\infty} \frac{nx^n}{2^n}$ **3.** $\displaystyle\sum_{n=1}^{\infty} n^2(x-2)^n$ **4.** $\displaystyle\sum_{n=0}^{\infty} \frac{(x+4)^n}{2^n}$

5. $\displaystyle\sum_{n=1}^{\infty} \frac{(x+1)^n}{n^n}$ **6.** $\displaystyle\sum_{n=1}^{\infty} n^n x^n$ **7.** $\displaystyle\sum_{n=0}^{\infty} (-1)^n x^n$ **8.** $\displaystyle\sum_{n=1}^{\infty} \frac{nx^n}{n+1}$

9. Note that power series of the form $\sum_{n=0}^{\infty} a_n x^n$ have an issue at $x=0$ when $n=0$: 0^0 is an indeterminate form—it can equal anything (or nothing). What value has it implicitly been assigned so far? What would be the technically correct way to write the series $\sum_{n=0}^{\infty} a_n x^n$ so that this issue goes away?

10. Show that

$$\sum_{n=1}^{\infty} n x^n = \frac{x}{(1-x)^2} \quad \text{for } -1 < x < 1.$$

11. Write the following infinite series as a rational number:

$$\frac{1}{10} + \frac{2}{100} + \frac{3}{1000} + \frac{4}{10000} + \cdots$$

B

12. Differentiating term by term, verify that the Bessel function $J_0(x)$ satisfies Bessel's equation (see equation (9.11)).

13. Show that for all $m \ge 1$ the Bessel functions $J_m(x)$ converge for all x.

14. For all $m \ge 1$ verify that the Bessel functions $J_m(x)$ satisfy the general Bessel equation of order m (see equation(9.13)).

15. For the Bessel functions $J_0(x)$ and $J_1(x)$ show that:

(a) $\dfrac{d}{dx}(x\,J_1(x)) = x\,J_0(x)$

(b) $\int J_0(x)J_1(x)\,dx = -\frac{1}{2}J_0^2(x)$

(c) $\int x\,J_0(x)J_1(x)\,dx = -\frac{1}{2}x\,J_0^2(x) + \frac{1}{2}\int J_0^2(x)\,dx$

(d) For all integers $n \ge 2$,

$$\int x^n J_0(x)\,dx = x^n J_1(x) + (n-1)x^{n-1}J_0(x) - (n-1)^2 \int x^{n-2} J_0(x)\,dx\,.$$

(Hint: Use part (a) and integration by parts twice.)

16. For all integers $m \ge 2$ show that the Bessel functions $J_m(x)$ satisfy the relations:

(a) $m\,J_m(x) + x\,J_m'(x) = x\,J_{m-1}(x)$

(b) $J_{m-1}(x) - J_{m+1}(x) = 2J_m'(x)$

17. Use long division to obtain the first three terms of $\frac{1}{xJ_0^2(x)}$, then integrate term by term to show that

$$J_0(x)\int \frac{dx}{xJ_0^2(x)} = J_0(x)\cdot \ln x + \frac{x^2}{4} - \frac{3x^4}{128} + \cdots\,.$$

This function is a *Bessel function of the second kind* and is another solution of Bessel's equation.

9.5 Taylor's Series

In the previous section a few functions, e.g. $f(x) = \frac{1}{1-x}$, turned out to be the sum of a power series. This section will discuss a general method for representing a function as a power series, called a **Taylor's series**.[11] Suppose that a function $f(x)$ can be written as

$$f(x) \;=\; \sum_{n=0}^{\infty} a_n\,(x-c)^n$$

either for all x or for $|x-c| < R$, for some $R > 0$. Then $f(c) = a_0$, and differentiating term by term yields

$$\begin{aligned}
f'(x) &= \sum_{n=1}^{\infty} n\,a_n\,(x-c)^{n-1} & \Rightarrow\quad f'(c) &= 1\cdot a_1\\
f''(x) &= \sum_{n=2}^{\infty} n\,(n-1)a_n\,(x-c)^{n-2} & \Rightarrow\quad f''(c) &= 2\cdot 1\cdot a_2\\
f'''(x) &= \sum_{n=3}^{\infty} n\,(n-1)(n-2)a_n\,(x-c)^{n-3} & \Rightarrow\quad f'''(c) &= 3\cdot 2\cdot 1\cdot a_3\\
&\cdots & &\cdots\\
f^{(k)}(x) &= \sum_{n=k}^{\infty} n\,(n-1)(n-2)\cdots(n-k+1)a_n\,(x-c)^{n-k} & \Rightarrow\quad f^{(k)}(c) &= k!\,a_k
\end{aligned}$$

so that in general (since $f^{(0)}(x) = f(x)$ and $0! = 1$):

$$a_n \;=\; \frac{f^{(n)}(c)}{n!} \quad \text{for } \; n \ge 0 \tag{9.15}$$

These $\{a_n\}$ are the **Taylor's series coefficients** of $f(x)$ at $x = c$. The full power series representation of $f(x)$ can now be stated:

Taylor's formula: If $f(x)$ has a power series representation in powers of $x-c$, where $x = c$ is inside the interval of convergence, then that representation is unique in that interval and is given by

$$f(x) \;=\; \sum_{n=0}^{\infty} \frac{f^{(n)}(c)}{n!}(x-c)^n \tag{9.16}$$

for all x in the interval. This is the **Taylor's series** for $f(x)$ about $x = c$.

[11] Named after English mathematician Brook Taylor (1685-1731), though such series were known to others (e.g. James Gregory, Johann Bernoulli) before Taylor.

Example 9.20

Find the Taylor's series for $f(x) = e^x$ about $x = 0$.

Solution: Since $\frac{d}{dx}(e^x) = e^x$, then for all $n \ge 0$,

$$f^{(n)}(x) = e^x \quad \Rightarrow \quad f^{(n)}(0) = e^0 = 1 .$$

Thus, by Taylor's formula with $c = 0$:[12]

$$\begin{aligned} e^x &= \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = \sum_{n=0}^{\infty} \frac{x^n}{n!} \\ &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots \end{aligned}$$

For what x is this Taylor's series valid? Recall that Example 9.16 showed the interval of convergence is all of $\mathbb{R}$. Thus the above Taylor's series holds for all x.

Before continuing, you might be wondering why you should even bother with finding the Taylor's series—after all, in the above example why replace a simple function like e^x by a far more complicated expression? One reason is that it often helps simplify some computations, especially in integrals. The idea is to use only a few terms in the series, i.e. a polynomial, as an approximation, since polynomials are generally easier to work with. Perhaps surprisingly, in many practical applications no more than two terms are needed, and often only one.

For example, using only the first two terms of the Taylor's series for e^x in Example 9.20, $e^x \approx 1 + x$ is a good approximation when x is close to 0 (i.e. $|x| \ll 1$). Using more terms does not necessarily help—for $|x| \ll 1$ and $n > 1$, x^n will be effectively 0. So the added complexity would not make the approximation significantly better.

Example 9.21

The energy density E of electromagnetic radiation at wavelength λ from a black-body at temperature T degrees Kelvin is given by *Planck's Law* of black-body radiation,

$$E(\lambda) = \frac{8\pi hc}{\lambda^5 (e^{hc/\lambda kT} - 1)}$$

where h is Planck's constant, c is the speed of light, and k is Boltzmann's constant. Show that for $\lambda \gg 1$:

$$E(\lambda) \approx \frac{8\pi kT}{\lambda^4}$$

Solution: Since $e^x \approx 1 + x$ for $|x| \ll 1$ by the Taylor series for e^x, let $x = hc/\lambda kT$. Then $x \ll 1$ and so

$$E(\lambda) = \frac{8\pi hc}{\lambda^5 (e^{hc/\lambda kT} - 1)} \approx \frac{8\pi hc}{\lambda^5 \left(\left(1 + \frac{hc}{\lambda kT}\right) - 1 \right)} \approx \frac{8\pi hc}{\lambda^5 \frac{hc}{\lambda kT}} \approx \frac{8\pi kT}{\lambda^4}$$

---

[12]The special case of $c = 0$ in Taylor's formula yields what is sometimes called the *Maclaurin's series* for $f(x)$, though that terminology is typically not used in fields of study outside mathematics.

Example 9.22

Find the Taylor's series for $f(x) = \sin x$ about $x = 0$.

Solution: The derivatives of $f(x) = \sin x$ repeat every four derivatives:

$$f(x) = \sin x \quad , \quad f'(x) = \cos x \quad , \quad f''(x) = -\sin x \quad , \quad f'''(x) = -\cos x \quad , \quad f^{(4)}(x) = \sin x$$

So at $x = 0$:

$$f(0) = 0 \quad , \quad f'(x) = 1 \quad , \quad f''(x) = 0 \quad , \quad f'''(x) = -1 \quad , \quad f^{(4)}(x) = 0$$

So for $n \ge 0$,

$$f^{(n)}(0) = \begin{cases} 0 & \text{if } n \text{ is even,} \\ 1 & \text{if } n = 1,5,9,\ldots, \\ -1 & \text{if } n = 3,7,11,\ldots. \end{cases}$$

Thus, by Taylor's formula with $c = 0$

$$\begin{aligned} \sin x &= \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n \\ &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \end{aligned}$$

By the Ratio Test this series converges for all x, since for any fixed x,

$$r(x) = \lim_{n\to\infty} \left| \frac{(-1)^{n+1} \dfrac{x^{2n+3}}{(2n+3)!}}{(-1)^n \dfrac{x^{2n+1}}{(2n+1)!}} \right| = x^2 \cdot \lim_{n\to\infty} \left| \frac{1}{(2n+3)(2n+2)} \right| = 0 < 1 \,.$$

Example 9.23

Find the Taylor's series for $f(x) = \cos x$ about $x = 0$.

Solution: The Taylor's series can be found using the same procedure as in Example 9.22, but it is simpler to just differentiate the Taylor's series for $\sin x$ term by term for all x:

$$\begin{aligned} \cos x &= \frac{d}{dx}(\sin x) = \frac{d}{dx}\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots \right) \\ &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \cdots \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} \end{aligned}$$

Since the Taylor's series for $\sin x$ converges for all x then so does its derivative. Thus the Taylor's series for $\cos x$ converges for all x.

Notice that the Taylor's series for $\cos x$ has only even powers of x, while the series for $\sin x$ has only odd powers of x. This makes sense since $\cos x$ and $\sin x$ are even and odd functions, respectively.

Example 9.24

The function $\ln x$ is not defined at $x = 0$ and hence has no Taylor's series about $x = 0$. Instead, find the Taylor's series for $f(x) = \ln(1+x)$ about $x = 0$.

Solution: Take successive derivatives:

$$f(x) = \ln(1+x) \quad , \quad f'(x) = \frac{1}{1+x} \quad , \quad f''(x) = -\frac{1}{(1+x)^2} \quad , \quad f'''(x) = \frac{1\cdot 2}{(1+x)^3} \quad , \quad f^{(4)}(x) = -\frac{1\cdot 2\cdot 3}{(1+x)^4}$$

So $f(0) = 0$ and for $n \geq 1$:

$$f^{(n)}(x) = (-1)^{n-1}\frac{(n-1)!}{(1+x)^n} \quad \Rightarrow \quad f^{(n)}(0) = (-1)^{n-1}(n-1)!$$

Thus, by Taylor's formula,

$$\begin{aligned}\ln(1+x) &= \sum_{n=0}^{\infty}\frac{f^{(n)}(0)}{n!}x^n = \sum_{n=1}^{\infty}(-1)^{n-1}\frac{(n-1)!\,x^n}{n!}\\ &= \sum_{n=1}^{\infty}(-1)^{n-1}\frac{x^n}{n}\\ &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \cdots\end{aligned}$$

Use the Ratio Test to find the interval of convergence:

$$r(x) = \lim_{n\to\infty}\left|\frac{(-1)^n\,\dfrac{x^{n+1}}{n+1}}{(-1)^{n-1}\,\dfrac{x^n}{n}}\right| = |x|\cdot\lim_{n\to\infty}\left|\frac{n}{n+1}\right| = |x|\cdot 1 = |x|$$

So the series converges when $|x| < 1$. Check the cases $r(x) = |x| = 1$ individually. For $x = 1$ the series is the alternating harmonic series $\sum_{n=1}\frac{(-1)^{n-1}}{n}$, which converges. For $x = -1$ the series is $-\sum_{n=1}\frac{1}{n}$, the negative of the harmonic series, which diverges. Thus, the series converges for $-1 < x \leq 1$.

Example 9.25

Find the Taylor's series for $f(x) = e^{x^2}$ about $x = 0$.

Solution: The Taylor's series can be found using the same procedure as in Example 9.20, but it is simpler to just replace each occurrence of x in the Taylor's series for e^x by x^2 (since the series for e^x converges for all x). In other words, make the substitution $u = x^2$ in the Taylor's series for e^u about $u = 0$:

$$\begin{aligned}e^u &= \sum_{n=0}^{\infty}\frac{u^n}{n!}\\ e^{x^2} &= \sum_{n=0}^{\infty}\frac{(x^2)^n}{n!} = \sum_{n=0}^{\infty}\frac{x^{2n}}{n!}\\ &= 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!} + \frac{x^{10}}{5!} + \cdots\end{aligned}$$

Define the ***n*-th degree Taylor polynomial** $P_n(x)$ for a function $f(x)$ about $x = c$ by

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k$$
$$= f(c) + \frac{f'(c)}{1!}(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \cdots + \frac{f^{(n)}(c)}{n!}(x-c)^n$$

for x in the interval of convergence for the full Taylor's series. In other words, $P_n(x)$ is the n-th partial sum of the Taylor's series. Since some of the coefficients could be zero, $P_n(x)$ is a polynomial of degree at most n. Thus, $P_n(x) = O(x^n)$. For that reason $P_n(x)$ is sometimes called the $\boldsymbol{O(x^n)}$ **approximation** to $f(x)$.

Figure 9.5.1 shows a comparison of $\sin x$ with a few of its approximations:

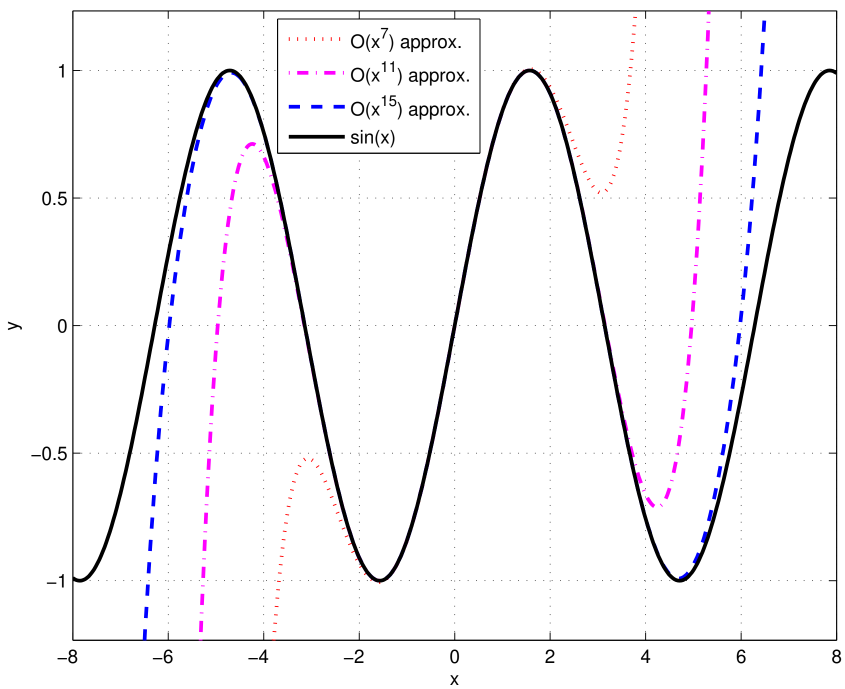


Figure 9.5.1 $\sin x$ and Taylor's series approximations

As you can see, the Taylor polynomials of degree 7, 11 and 15 are all good approximations over the interval $[-2, 2]$, with the $O(x^{15})$ approximation still being fairly good over $[-6, 6]$. Clearly those approximations all become poor quite quickly for $|x| > 6$; they approach $\pm\infty$, unlike $\sin x$.

The following theorem shows how to measure the accuracy of the approximations:

Remainder Theorem:[13] If $P_n(x)$ is the n-th degree Taylor polynomial about $x = c$ for a function $f(x)$ in some interval containing $x = c$, then for all x in that interval,

$$f(x) \;=\; P_n(x) \;+\; R_n(x)\;, \tag{9.17}$$

where

$$R_n(x) \;=\; \frac{f^{(n+1)}(c+\theta(x-c))}{(n+1)!}\,(x-c)^{n+1} \tag{9.18}$$

for some number θ between 0 and 1. Alternatively,

$$R_n(x) \;=\; \frac{1}{n!}\int_c^x (x-t)^n\, f^{(n+1)}(t)\; dt\;. \tag{9.19}$$

Since the number θ is unknown in equation (9.18), usually only an upper bound on the remainder $R_n(x)$ can be found by that formula. For practical purposes formula (9.19) for $R_n(x)$ might be easier to use (via numerical integration).

A common misconception is that hand-held calculators use Taylor's series to compute values of functions like $\sin x$, $\cos x$, e^x, etc. However, that is typically well beyond their capability, especially for large values of x—far too many terms would be required. Instead, many calculators use an algorithm called CORDIC[14] (Coordinate Rotation Digital Computer), and—perhaps surprisingly—lookup tables. CORDIC uses the computationally inexpensive operation of bit-shifting to translate large input values into a smaller range, then uses tables stored in memory for values in that range, along with interpolation for numbers between those in the tables.[15]

Exercises

A

For Exercises 1-9 write out the first three nonzero terms in the Taylor's series for the given function $f(x)$ about the given value c. You may use any method you like.

1. $f(x) = \sin x\;;\; c = \pi/2$

2. $f(x) = \sinh x\;;\; c = 0$

3. $f(x) = \cosh x\;;\; c = 0$

4. $f(x) = \tan x\;;\; c = 0$

5. $f(x) = \tanh x\;;\; c = 0$

6. $f(x) = \sec x\;;\; c = 0$

7. $f(x) = \dfrac{1}{1+x^2}\;;\; c = 0$

8. $f(x) = \dfrac{1}{1+x^2}\;;\; c = 1$

9. $f(x) = \sqrt{1+x^2}\;;\; c = 0$

10. Use Example 9.19 from Section 9.4 to write out the first three nonzero terms in the Taylor's series for $f(x) = \frac{1}{(1-x)^3}$ about $x = 0$.

11. Use the derivative of the Taylor's series for $\sqrt{1+x^2}$ from Exercise 9 to write out the first three nonzero terms in the Taylor's series for $f(x) = \frac{x^2}{\sqrt{1+x^2}}$ about $x = 0$.

[13] For a proof see pp.171-172 in KLAMBAUER, G., *Aspects of Calculus*, New York: Springer-Verlag, 1986.

[14] See Ch.7 in SCHMID, H., *Decimal Computation*, New York: John Wiley & Sons, Inc., 1974.

[15] To see how inaccurate calculators can be, compute tan(355/226) in radian mode on a calculator. The true value to 3 decimal places is −7497258.185, but few calculators produce an answer close to that.

For Exercises 12-15 replace the function $f(x)$ by its Taylor's series about $x = 0$ to evaluate the given indefinite integral $\int f(x)\,dx$ (up to the first three nonzero terms in the series).

12. $\displaystyle\int \frac{\sin x}{x}\,dx$ **13.** $\displaystyle\int \cos(x^2)\,dx$ **14.** $\displaystyle\int e^{-x^2}\,dx$ **15.** $\displaystyle\int \sqrt{1+x^6}\,dx$

16. Use $\frac{d}{dx}(\tan^{-1}x) = \frac{1}{1+x^2}$ along with Exercise 7 to find the Taylor's series for $f(x) = \tan^{-1}x$ about $x = 0$, along with its interval of convergence.

17. Use Exercise 16 to show that

$$\pi = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} - \cdots\right).$$

18. Use the first three nonzero terms in the Taylor's series about $x = 0$ for $e^{-x^2/2}$ to evaluate the definite integral

$$\int_{-1}^{1} \frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}\,dx \quad .$$

Note: The actual value (rounded to 4 decimal places) is 0.6826.

19. Recall that the surface area S of the solid obtained by revolving the curve $y = x^2$ around the x-axis between $x = 0$ and $x = 2$ is given by the integral

$$S = \int_0^2 2\pi x^2\sqrt{1 + 4x^2}\,dx \quad .$$

The exact value of the integral rounded to 3 decimal places is $S = 53.226$. Use the first two nonzero terms in the Taylor's series for $\sqrt{1+4x^2}$ about $x = 0$ to approximate the integral. How close is this approximation to the actual value? Does the approximation become better if you use the first three nonzero terms in the Taylor's series? Justify your answer.

20. The tangential component of a space shuttle's velocity during reentry is approximately

$$v(t) = v_c \tanh\left(\frac{g}{v_c}t + \tanh^{-1}\left(\frac{v_0}{v_c}\right)\right)$$

where v_0 is the velocity at time $t = 0$ and v_c is the terminal velocity. If $\tanh^{-1}\left(\frac{v_0}{v_c}\right) = \frac{1}{2}$ then show that $v(t) \approx gt + \frac{1}{2}v_c$.

21. The velocity of a water wave of length L in water of depth h satisfies the equation $v^2 = \frac{gL}{2\pi}\tanh\left(\frac{2\pi h}{L}\right)$. Show that $v \approx \sqrt{gh}$.

22. A disk of radius a has a charge of constant density σ. A point P lies at a distance r directly above the disk. The *electrical potential* V at point P is given by $V = 2\pi\sigma(\sqrt{r^2 + a^2} - r)$. Show that $V \approx \frac{\pi a^2\sigma}{r}$ for large r (i.e. $r \gg 1$).

23. The fifth-degree *Padé approximation* uses rational functions to approximate $\tanh x$:

$$\tanh x \approx \frac{x^5 + 105x^3 + 945x}{15x^4 + 420x^2 + 945}$$

Compare the values of the Padé approximation and the fifth-degree Taylor's series approximation from Exercise 5, evaluated at $x = 1$. Which is better? The actual value of tanh(1) is 0.7615941559558. How do the two approximations compare at $x = 2$?

APPENDIX A

Answers and Hints to Selected Exercises

Chapter 1

Section 1.1 (p. 6)

1. $2t$ **2.** $19.6t$ **3.** $-32t+2$ **4.** $3t^2$

Section 1.2 (p. 14)

1. 0 **3.** $2x+2$ **5.** $-\frac{1}{(x+1)^2}$ **7.** $-\frac{2}{x^3}$
9. $\frac{1}{2\sqrt{x+1}}$ **11.** $\frac{2x+3}{2\sqrt{x^2+3x+4}}$

Section 1.3 (p. 20)

5. Hint: Use the sine double-angle formula. **7.** Hint: Use Exercise 5 and the sine addition formula.

Section 1.4 (p. 26)

1. $2x-1$ **3.** $4x^5+\frac{9}{x^7}$ **5.** $x\cos x+\sin x$
7. $\frac{x\cos x-\sin x}{x^2}$ **9.** $\frac{2-2t^2}{(1+t^2)^2}$ **11.** $\frac{ad-bc}{(cx+d)^2}$
13. $2\pi r$

Section 1.5 (p. 30)

1. $-20(1-5x)^3$ **3.** $-\frac{1}{\sqrt{1-2x}}$ **5.** $\frac{1-x}{2\sqrt{x}(x+1)^2}$
7. $-\frac{8(1-t)^3}{(1+t)^5}$ **9.** $2\sin x\cos x$ **11.** $15\sec^2(5x)$
13. $2x\sec(x^2)\tan(x^2)$ **15.** $\beta(1-\beta^2)^{-3/2}$
17. $\sin(\cos x)\sin x$ **21.** Hint for part(b): Use part(a) and the Chain Rule to find S_p, then recall how to convert from radians per second to revolutions per minute.

Section 1.6 (p. 35)

1. $6x+2$ **3.** $-9\cos 3x$ **5.** $\frac{2}{x^3}$

Chapter 2

Section 2.1 (p. 40)

1. $f^{-1}(x)=x$, $\left(f^{-1}\right)'(x)=1$
3. $f^{-1}(x)=\sqrt{x}$, $\left(f^{-1}\right)'(x)=\frac{1}{2\sqrt{x}}$
5. $f^{-1}(x)=\frac{1}{x}$, $\left(f^{-1}\right)'(x)=-\frac{1}{x^2}$
7. $f^{-1}(x)=\frac{1}{\sqrt{x}}$, $\left(f^{-1}\right)'(x)=-\frac{1}{2}x^{-3/2}$

Section 2.2 (p. 44)

1. $6\sec^2 3x\tan 3x$ **3.** $-3\csc^2 3x$ **5.** $\frac{3}{9+x^2}$
7. $-\frac{3}{1+9x^2}$ **9.** $\frac{1}{1+x^2}$ **11.** $\frac{6\sin^{-1}3x}{\sqrt{1-9x^2}}$
13. $\frac{1}{1+x^2}$ **15.** $\cot^{-1}x-\frac{x}{1+x^2}$

Section 2.3 (p. 52)

1. $2e^{2x}$ **3.** $-e^{-x}-e^{x}$ **5.** $\frac{2e^x}{(1-e^x)^2}$ **7.** $e^{e^x}e^x$
9. $\frac{1}{x}$ **11.** $\frac{6x(\ln(\tan x^2))^2\sec^2 x^2}{\tan x^2}$

15. $x^{x^2}(x+2x\ln x)$
17. $x^{\sin x}\left(\cos x \ln x + \frac{\sin x}{x}\right)$ **19.** 15.5 hours
21. 12 hours

Section 2.4 (p. 55)

1. $\frac{\ln 3(3^x - 3^{-x})}{2}$ **3.** $(\ln 2)^2\, 2^{2^x}\, 2^x$
5. $\frac{2x}{(\ln 2)(x^2+1)}$ **7.** $\frac{\cos(\log_2 \pi x)}{x \ln 2}$ **9.** $3x^2$

Chapter 3

Section 3.1 (p. 61)

1. $y = 4x - 3$ **3.** $y = -6x + 10$ **5.** $y = 4x$
7. $y = x + 3$ **9.** $y = 240x + 176$ **11.** $y = 2x$
13. $y = 3x + \frac{31}{27}$, $y = 3x + 1$ **15.** 75.96°
17. 0° **19.** 116.6° **21.** 5.71°
23. $y = -\frac{1}{4}x + \frac{11}{2}$
25. $y = -\frac{1}{4}x - \frac{81}{4}$, $y = -\frac{1}{4}x + \frac{1159}{108}$

Section 3.2 (p. 72)

1. $\frac{7}{3}$ **3.** 0 **5.** −1 **7.** 0 **9.** 2 **11.** 0
14. $\frac{1}{2}$ **15.** 0 **17.** 0

Section 3.3 (p. 78)

1. continuous **3.** discontinuous
5. discontinuous **7.** discontinuous
9. continuous **11.** continuous
13. continuous **15.** discontinuous
17. continuous **19.** 1 **21.** e^{-1}
25. Hint: Use the Intermediate Value Theorem.

Section 3.4 (p. 81)

1. $\frac{-3x^2y + 4y^2 + 2x}{x^3 - 8xy - 1}$ **3.** $\frac{2(x-y+1) - 3(x+y)^2}{2(x-y+1) + 3(x+y)^2}$
5. $\frac{2x(1-(x^2-y^2))}{y(2(y^2-x^2)-1)}$ **7.** $-\frac{y}{x}$
9. $-\frac{-2x - y + 3x^2y^2e^{\sin(xy)} + x^3y^3e^{\sin(xy)}\cos(xy)}{x^4y^2e^{\sin(xy)}\cos(xy) + 2x^3ye^{\sin(xy)} - 3y^2 - x}$
13. $-\frac{x^2+y^2}{y^3}$

Section 3.5 (p. 83)

1. 80π ft/s **3.** 2.4 ft/s **5.** 10 ft/s
7. -76π cm^3/min **9.** 45.14 mph
11. 155.8 ft/min

Section 3.6 (p. 88)

1. $(2x-2)\,dx$ **2.** $4x \sin(x^2)\cos(x^2)\,dx$
11. Hint: Mimic Example 3.35.

Chapter 4

Section 4.1 (p. 98)

1. $(1,1)$ **3.** 125,000 sq yd **5.** $U = \frac{V}{2}$
7. $R = r$ **9.** $2ab$ **13.** $Q = \sqrt{\frac{2DP}{I+W}}$
15. $r = \sqrt{r_0^2 + x_0^2}$ **17.** 380.62 minutes
19. $12\pi\sqrt{3}$ **21.** 12.8 ft **22.** Hint: Place the right angle of the triangle at the origin in the xy-plane. **25.** $x = \sqrt{\frac{RN}{r}}$
27. $x = \frac{a}{\sqrt{2}}$ **33.** $(a^{2/3} + b^{2/3})^{3/2}$
34. Hint: You can leave your answer in terms of R and an angle satisfying a certain equation.

Section 4.2 (p. 108)

2. local maximum at $x = 0$, local minimum at $x = 2$, inflection pt at $x = 1$, increasing for $x < 0$ and $x > 2$, decreasing for $0 < x < 2$, concave up for $x > 1$, concave down for $x < 1$
3. local maximum at $x = 1$, inflection pt at $x = 2$, increasing for $x < 1$, decreasing for $x > 1$, concave up for $x > 2$, concave down for $x < 2$, horizontal asymptote: $y = 0$
5. local maximum at $x = 0$, inflection pts at $x = \pm\frac{1}{\sqrt{3}}$, increasing for $x < 0$, decreasing for $x > 0$, concave up for $x < -\frac{1}{\sqrt{3}}$ and $x > \frac{1}{\sqrt{3}}$, concave down for $-\frac{1}{\sqrt{3}} < x < \frac{1}{\sqrt{3}}$, horizontal asymptote: $y = 0$

7. local maximum at $x = \ln 2$, inflection pt at $x = \ln 4$, increasing for $x < \ln 2$, decreasing for $x > \ln 2$, concave up for $x > \ln 4$, concave down for $x < \ln 4$, horizontal asymptote: $y = 0$

Section 4.3 (p. 117)

1. $x = 0.450184$ **3.** $x = 0.567143$ **5.** $x = 1.414213$ **11.** global maximum at $x = 2.8214$ **14.** 50

Section 4.4 (p. 122)

1. No **3.** Yes **6.** No **18.** Hint: Calculate $f'(x)$ and use the cosine addition formula.

Chapter 5

Section 5.1 (p. 131)

1. $\frac{x^3}{3} + \frac{5x^2}{2} - 3x + C$ **3.** $4e^x + C$
5. $-5\cos x + C$ **7.** $6\ln|x| + C$
9. $-\frac{4}{3}x^{3/2} + C$ **11.** $\frac{x^2}{2} + \frac{3}{7}x^{7/3} + C$
13. $3\sec x + C$ **15.** $-7\cot x + C$

Section 5.2 (p. 139)

3. $\frac{1}{2}$ **4.** $\frac{1}{3}$ **5.** 1 **6.** $\frac{1}{4}$

Section 5.3 (p. 145)

1. $\frac{1}{3}$ **3.** $\frac{1}{4}$ **5.** $\frac{1}{2}$ **7.** 1 **9.** $2e - 2e^{-1}$
11. $\frac{16}{3}$

Section 5.4 (p. 150)

1. $\frac{3\sin 5x - 4\cos 5x}{5} + C$
3. $-\frac{1}{2}e^{-x^2} + \frac{1}{3}\sin x^3 + C$
5. $\ln(1 + e^x) + C$
7. $\frac{2}{5}(x+4)^{5/2} - \frac{8}{3}(x+4)^{3/2} + C$
9. $\tan x - x + C$ **11.** $\frac{3}{10}\tan^{-1}\left(\frac{5x}{2}\right) + C$
13. 10 **15.** $\frac{1192}{15}$ **17.** 1 **19.** $-\frac{1}{48}$
21. $\frac{\pi}{6}$ **23.** $\frac{1}{2}$

Section 5.5 (p. 158)

1. $\frac{1}{2}$ **3.** 1 **5.** divergent **7.** $\frac{1}{\ln 2}$
9. divergent **11.** 6 **13.** divergent
15. $\frac{\pi}{2}$ **19.** Yes **20.** No

Chapter 6

Section 6.1 (p. 165)

1. $\frac{x^2\ln x}{2} - \frac{x^2}{4} + C$ **2.** $(x^2 - 2x + 2)e^x + C$
3. $x\sin x + \cos x + C$
5. $\frac{x^2a^x}{\ln a} - \frac{2xa^x}{\ln^2 a} + \frac{2a^x}{\ln^3 a} + C$
7. $x\ln x^2 - 2x + C$
9. Hint: Use a double-angle identity.
11. $x\sin^{-1}x + \sqrt{1-x^2} + C$
13. $x\tan^{-1}3x - \frac{1}{6}\ln(1+9x^2) + C$
15. $-\frac{3}{8}\sin x\cos 3x + \frac{1}{8}\cos x\sin 3x + C$
17. $\frac{1}{4}x^4\ln^2 x - \frac{1}{8}x^4\ln x + \frac{1}{32}x^4 + C$ **19.** $\frac{16}{3}$
20. $\frac{2\sqrt{2}+2}{15}$ **21.** $\frac{x\sin(\ln x)}{2} - \frac{x\cos(\ln x)}{2} + C$
23. $\frac{x^2\tan^{-1}x}{2} - \frac{x}{2} + \frac{\tan^{-1}x}{2} + C$
24. $x\cot^{-1}\sqrt{x} + \sqrt{x} + \cot^{-1}\sqrt{x} + C$
25. Hint: Try the substitution $t = \sqrt{x}$.

Section 6.2 (p. 171)

1. $-\frac{1}{14}\cos 7x + \frac{1}{6}\cos 3x + C$
3. $-\frac{1}{14}\sin 7x + \frac{1}{6}\sin 3x + C$
5. $-\frac{2}{5}\cos^{5/2}x + \frac{2}{9}\cos^{9/2}x + C$
7. $-\frac{1}{4}\sin 2x + \frac{1}{48}\sin^2 2x + \frac{5}{16}x + \frac{3}{64}\sin 4x + C$
9. $\frac{1}{3}\tan^3 x + \tan x + C$ **11.** $\frac{1}{4}\sin^4 x + C$

Section 6.3 (p. 177)

1. $\frac{1}{2}x\sqrt{9+4x^2} + \frac{9}{4}\ln|2x + \sqrt{9+4x^2}| + C$
3. $\frac{1}{2}x\sqrt{4x^2-9} - \frac{9}{4}\ln|2x + \sqrt{4x^2-9}| + C$
5. $-\sin^{-1}x - \frac{\sqrt{1-x^2}}{x} + C$
7. $\ln|x| - \ln|1 + \sqrt{1+x^2}| + C$
9. $\frac{1}{3}(x^2+4)^{3/2} - 4\sqrt{x^2+4} + C$
11. $\frac{1}{108}\tan^{-1}\left(\frac{2x}{3}\right) + \frac{x}{18(9+4x^2)} + C$

13. $-9\sqrt{9-x^2}+\frac{1}{3}(9-x^2)^{3/2}+C$
15. $-\frac{1}{9}\sqrt{-9x^2+36x-32}-\frac{2}{3}\sin^{-1}\left(\frac{3x-6}{2}\right)+C$

Section 6.4 (p. 184)

1. $-\ln|x|+\ln|x-1|+C$
3. $\frac{1}{5}\ln|2x-1|-\frac{1}{5}\ln|x+2|+C$
5. $\frac{1}{x}+\frac{1}{2}\ln|x-1|-\frac{1}{2}\ln|x+1|+C$
7. $2\ln|x|+\frac{1}{x}-2\ln|x+1|+C$
9. $-3\ln|x|+\frac{2}{x}+3\ln|x-1|+\frac{1}{x-1}+C$
11. $\frac{1}{3}\tan^{-1}x-\frac{1}{6}\tan^{-1}\left(\frac{x}{2}\right)+C$

Section 6.5 (p. 193)

1. $2\ln\left|\sec\frac{1}{2}\theta\right|-\ln|\sin\theta|+C$
3. $\frac{2}{\sqrt{3}}\tan^{-1}\left(\frac{2\tan\frac{1}{2}\theta-1}{\sqrt{3}}\right)+C$
5. $\frac{4}{\sqrt{3}}\tan^{-1}\left(\frac{2\tan\frac{1}{2}\theta-1}{\sqrt{3}}\right)-\theta+C$
7. $\ln\left|\tan\frac{1}{2}\theta\right|-\ln\left|\tan\frac{1}{2}\theta+1\right|+C$
9. $\ln\left|\tan\frac{1}{2}\theta\right|-2\ln\left|\tan\frac{1}{2}\theta+1\right|+C$
11. $\sqrt{2\pi}$ **23.** $\frac{3}{2\Gamma(\frac{2}{3})}x^{2/3}$

Section 6.6 (p. 201)

1. The true value is $P\approx 7.4163\sqrt{l/g}$ (i.e. the integral is ≈ 1.8541)
2. $7.416331870724302\sqrt{l/g}$
3. 0.8948311310564181
7. 119.9785845899309
9. 0.5967390281992041 (The true value is 0.5963473623231939)

Chapter 7

Section 7.1 (p. 209)

2. Foci: $(\pm 3,0)$, vertexes: $(\pm 5,0)$, $e=\frac{3}{5}$
3. Foci: $(0,\pm\sqrt{5})$, vertexes: $(0,\pm 3)$, $e=\frac{\sqrt{5}}{3}$
5. Foci: $\left(\pm\frac{\sqrt{3}}{2},0\right)$, vertexes: $(\pm 1,0)$, $e=\frac{\sqrt{3}}{2}$
9. $\left(\pm\frac{ab}{\sqrt{a^2+b^2}},\pm\frac{ab}{\sqrt{a^2+b^2}}\right)$ **13.** Hint: Use the two points you know for certain are on the ellipse to find the location of the directrix.
16. Hint: Use Exercise 15 and formula (7.3).

Section 7.2 (p. 215)

2. Focus: $(0,2)$, vertex: $(0,0)$, directrix: $y=-2$
3. Focus: $\left(0,\frac{1}{32}\right)$, vertex: $(0,0)$, directrix: $y=-\frac{1}{32}$
4. Focus: $\left(\frac{1}{4},0\right)$, vertex: $(0,0)$, directrix: $x=-\frac{1}{4}$
5. Focus: $\left(\frac{-1}{12},0\right)$, vertex: $(0,0)$, directrix: $x=\frac{1}{12}$
7. $(0,0)$ and $(4p,4p)$; $y=x$ **9.** $|4p|$
11. $(3p,\pm 2\sqrt{3}p)$
16. Focus: $\left(\frac{-b}{2a},\frac{4ac-b^2+1}{4a}\right)$, vertex: $\left(\frac{-b}{2a},\frac{4ac-b^2}{4a}\right)$, directrix: $y=\frac{4ac-b^2-1}{4a}$

Section 7.3 (p. 223)

2. Foci: $(\pm 5,0)$, vertexes: $(\pm 4,0)$, directrices: $x=\pm\frac{16}{5}$, asymptotes: $y=\pm\frac{3}{4}x$, $e=\frac{5}{4}$
3. Foci: $(\pm\sqrt{23},0)$, vertexes: $(\pm 2\sqrt{2},0)$, directrices: $x=\pm\frac{8}{\sqrt{23}}$,
asymptotes: $y=\pm\frac{\sqrt{15}}{2\sqrt{2}}x$, $e=\frac{\sqrt{23}}{2\sqrt{2}}$
4. Foci: $(\pm\frac{\sqrt{41}}{2},0)$, vertexes: $(\pm\frac{5}{2},0)$, directrices: $x=\pm\frac{25}{2\sqrt{41}}$, asymptotes: $y=\pm\frac{4}{5}x$,
$e=\frac{\sqrt{41}}{5}$
5. Foci: $(\pm\frac{\sqrt{5}}{2},0)$, vertexes: $(\pm 1,0)$, directrices: $x=\pm\frac{2}{\sqrt{5}}$, asymptotes: $y=\pm\frac{1}{2}x$, $e=\frac{\sqrt{5}}{2}$
6. Foci: $(0,\pm\sqrt{34})$, vertexes: $(0,\pm 3)$, directrices: $y=\pm\frac{9}{\sqrt{34}}$, asymptotes: $y=\pm\frac{3}{5}x$, $e=\frac{\sqrt{34}}{3}$
7. $\frac{x^2}{9}-\frac{y^2}{16}=1$ **17.** $\frac{x^2}{302500}-\frac{y^2}{697500}=1$

Section 7.4 (p. 229)

1. Foci: $(0,2)$ and $(6,2)$, vertexes: $(-2,2)$ and $(8,2)$
3. Foci: $(-3,1\pm 2\sqrt{3})$, vertexes: $(-3,-3)$ and $(-3,5)$

5. Focus: $(-3,-\frac{239}{16})$, vertex: $(-3,-15)$, directrix: $y=-\frac{241}{16}$
7. Focus: $\left(\frac{1}{2},\frac{5}{4}\right)$, vertex: $\left(\frac{1}{2},\frac{3}{2}\right)$, directrix: $y=\frac{7}{4}$
9. Foci: $(-1\pm\sqrt{13},-3)$, vertexes: $(-4,-3)$ and $(2,-3)$, directrices: $x=-1\pm\frac{9}{\sqrt{13}}$, asymptotes: $y=\pm\frac{2}{3}(x+1)-3$
11. Foci: $(\sqrt{2},\sqrt{2})$ and $(-\sqrt{2},-\sqrt{2})$, vertexes: $(1,1)$ and $-1,-1)$, directices: $y=-x\pm\sqrt{2}$, asymptotes: $x=0$ and $y=0$
15. hyperbola

Section 7.5 (p. 236)

12. Hint: Use Exercise 11.
24. local maximum at $x=\ln\sqrt{3}$, inflection pt at $x=\ln 3$, horizontal asymptote: $y=0$
26. (b) Hint: See Exercise 10.
27. (b) $k_1=c_1+c_2$, $k_2=c_1-c_2$
30. $s_0\approx 1.006237835313385$,
$e^{\pi/s_0}=22.69438187638412$
33. (a) $x=\frac{cx+cy}{2}-\frac{y-x}{2c}$, $y=\frac{cx+cy}{2}+\frac{y-x}{2c}$
(b) $c=\cosh\alpha+\sinh\alpha$
(c) Hint: Use Example 7.12.

Section 7.6 (p. 243)

5. Yes **6.** (a) Hint: Solve for t in terms of x then substitute into y.
(b) Hint: Use the distance formula.
8. Hint: Does $BP=\widehat{AB}$? **9.** (a) $-\frac{7}{(2t+1)^3}$
(c) $\left(\frac{36}{25},\frac{14}{5}\right)$ **12.** $x=t^2-1$, $y=t(t^2-1)$

Section 7.7 (p. 250)

1. $r=6\cos\theta$ **3.** $r^2=\sec 2\theta$
5. $\theta=\frac{3\pi}{4}$ **7.** $r=\sec\theta\tan\theta$
9. $x^4+2x^2y^2+y^4=4x^2-4y^2$
11. $y-1=-\frac{3\sqrt{3}}{5}(x+\sqrt{3})$
13. local maxima at $\left(2,\frac{\pi}{2}\right)$ and $\left(0,\frac{3\pi}{2}\right)$, local minima at $\left(\frac{1}{2},\frac{7\pi}{6}\right)$ and $\left(\frac{1}{2},\frac{11\pi}{6}\right)$
15. local maxima at $\left(\frac{2\sqrt{2}}{3},\alpha\right)$ and $\left(-\frac{2\sqrt{2}}{3},2\pi-\alpha\right)$, local minima at $\left(-\frac{2\sqrt{2}}{3},\pi-\alpha\right)$ and $\left(\frac{2\sqrt{2}}{3},\pi+\alpha\right)$, where $\alpha=\tan^{-1}\sqrt{2}$ **17.** $\frac{\pi}{2}$

Chapter 8

Section 8.1 (p. 256)

1. $\frac{32}{3}$ **2.** $\frac{4}{3}$ **3.** $\frac{1}{12}$ **5.** $\frac{9}{2}$ **7.** $\frac{19}{3}$ **9.** $\frac{1}{12}$
11. 3π **13.** $25\left(\frac{\pi}{3}+1-\sqrt{3}\right)$
17. $\frac{6250\pi^3}{3}$ sq ft

Section 8.2 (p. 261)

1. 1 **3.** $\frac{4}{3}$ **5.** $\frac{2}{\pi}$ **7.** 0 **9.** $\frac{1}{2}\ln 3$

Section 8.3 (p. 270)

1. $\frac{8}{27}(10^{3/2})-\frac{1}{27}(13^{3/2})\approx 7.634$
2. $\frac{\sqrt{5}}{2}+\frac{1}{4}\sinh^{-1}2\approx 1.479$
3. $\frac{8}{27}(10^{3/2})-\frac{1}{27}(13^{3/2})\approx 7.634$
5. $\frac{3}{4}+\ln\sqrt{2}\approx 1.097$
7. $\sqrt{2}(e^{\pi}-1)\approx 31.312$
9. 8 **11.** 3 **13.** $\kappa(0)=0$, $\kappa(\frac{\pi}{2})=-1$
15. $\kappa(0)=\frac{a}{b^2}$ at $(a,0)$, $\kappa(\frac{\pi}{2})=\frac{b}{a^2}$ at $(0,b)$
17. -1 **23.** 13.27 ft

Section 8.4 (p. 276)

1. 4π **2.** $\frac{\pi}{2}(2+\sinh 2)$ **3.** $\frac{208\pi}{9}$
5. $\frac{\pi^2}{2}$ **7.** 2π **9.** $\frac{\pi}{10}$ **10.** $\frac{\pi}{6}$
13. $S=\pi r\sqrt{r^2+h^2}$, $V=\frac{1}{3}\pi r^2 h$

Section 8.5 (p. 283)

1. $\left(\frac{4}{5},\frac{2}{7}\right)$ **3.** $\left(\frac{3}{5},\frac{12}{35}\right)$ **5.** $\left(\frac{4r}{3\pi},\frac{4r}{3\pi}\right)$
7. $\left(0,\frac{11}{4\pi}\right)$ **9.** 0.192 Nm
11. $RT\left(\frac{1}{V_a^2}-\frac{1}{V_b^2}\right)$ **13.** 0.3486 **15.** $\left(1,\frac{1}{4}\right)$
18. Hint: Use Exercise 28 in Section 6.1.
20. Hint: Use the equation from Section 5.1 for free fall motion to write time as a function of height.

Chapter 9

Section 9.1 (p. 291)

1. Converges to 0 **2.** Converges to $\frac{1}{3}$
3. Converges to 0 **5.** Divergent
7. Divergent **9.** 6 **11.** 32 **13.** $\frac{113}{999}$
14. 1 **15.** $\frac{1}{4}$ **20.** $\frac{132}{7}$ ft **24.** No

Section 9.2 (p. 298)

6. Divergent **7.** Convergent
8. Divergent **9.** Convergent
10. Divergent **11.** Convergent
12. Convergent **13.** Convergent
14. Divergent **15.** Divergent
16. Convergent **17.** Convergent
18. $\frac{1}{6}$ **19.** $\frac{1}{10}$ **20.** $\frac{1}{6}$ **21.** $\frac{1}{2}$

Section 9.3 (p. 302)

1. Conditionally convergent
3. Conditionally convergent
5. Absolutely convergent
7. Answer to second question: Yes

Section 9.4 (p. 307)

1. $-1 \le x < 1$ **2.** $-2 < x < 2$ **3.** $1 < x < 3$
4. $-6 < x < -2$ **5.** $-\infty < x < \infty$ **6.** $x = 0$
7. $-1 < x < 1$ **10.** Hint: Use Example 9.19

Section 9.5 (p. 313)

1. $1 - \frac{(x-\frac{\pi}{2})^2}{2!} + \frac{(x-\frac{\pi}{2})^4}{4!} - \cdots$
2. $x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots$ **3.** $1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots$
4. $x + \frac{x^3}{3} + \frac{2x^5}{15} + \cdots$ **5.** $x - \frac{x^3}{3} + \frac{2x^5}{15} - \cdots$
6. $1 + \frac{x^2}{2} + \frac{5x^4}{24} + \cdots$ **7.** $1 - x^2 + x^4 - \cdots$
8. $\frac{1}{2} - \frac{(x-1)}{2} + \frac{(x-1)^2}{4} - \cdots$
9. $1 + \frac{x^2}{2} - \frac{x^4}{8} + \cdots$ **11.** $x^2 - \frac{x^4}{2} + \frac{3x^6}{8} - \cdots$
13. $x - \frac{x^5}{10} + \frac{x^9}{216} - \cdots$ **15.** $x + \frac{x^7}{14} - \frac{x^{13}}{104} + \cdots$
18. 0.68485 **19.** 97.18; −132.605

GNU Free Documentation License

Version 1.3, 3 November 2008
Copyright © 2000, 2001, 2002, 2007, 2008 Free Software Foundation, Inc.

<http://fsf.org/>

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "**Document**", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "**you**". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "**Modified Version**" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "**Secondary Section**" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "**Invariant Sections**" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "**Cover Texts**" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "**Transparent**" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "**Opaque**".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "**Title Page**" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "**publisher**" means any person or entity that distributes copies of the Document to the public.

A section "**Entitled XYZ**" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "**Acknowledgments**", "**Dedications**", "**Endorsements**", or "**History**".) To "**Preserve the Title**" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgments" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgments and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgments", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections.

You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled “Acknowledgments”, “Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See `http://www.gnu.org/copyleft/`.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License “or any later version” applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy’s public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright © YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with ... Texts." line with this:

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

History

This section contains the revision history of the book. For persons making modifications to the book, please record the pertinent information here, following the format in the first item below.

1. VERSION: 0.1

   Date: 2016-01-24

   Author(s): Michael Corral

   Title: Elementary Calculus

   Modification(s): Initial version

2. VERSION: 0.5

   Date: 2020-05-27

   Author(s): Michael Corral

   Title: Elementary Calculus

   Modification(s): Numerous corrections of typos and errors, as well as enhancements of some sections and more exercises.

3. VERSION: 0.5a

   Date: 2020-06-03

   Author(s): Michael Corral

   Title: Elementary Calculus

   Modification(s): Replaced the old Section 5.5 (Average Value of a Function) with the new Section 5.5 (Improper Integrals). Average Value of a Function will be moved to Chapter 8 in the second half of the book.

4. VERSION: 1.0

   Date: 2020-12-31

   Author(s): Michael Corral

   Title: Elementary Calculus

   Modification(s): Full version.

Index

D

E

I

J

L