CALCULUS



Eric Hutchinson
College of Southern Nevada



Calculus (Hutchinson)

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide and accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 06/10/2025

TABLE OF CONTENTS

Licensing

1: Functions and Graphs (Precalculus Review)

- 1.1: Review of Functions
- 1.2: Basic Classes of Functions
- 1.3: Trigonometric Functions
- 1.4: Inverse Functions
- 1.5: Exponential and Logarithmic Functions

2: Limits and Continuity

- 2.1: Limit of a Function and Limit Laws
- 2.2: The Precise Definition of a Limit
- 2.3: One-Sided Limits
- 2.4: Continuity
- 2.5: Limits Involving Infinity; Asymptotes of Graphs

3: Differentiation

- 3.1: Defining the Derivative
- 3.2: The Derivative as a Function
- 3.3: Differentiation Rules
- 3.4: The Derivative as a Rate of Change
- 3.5: Derivatives of Trigonometric Functions
- 3.6: The Chain Rule
- 3.7: Implicit Differentiation
- 3.8: Derivatives of Inverse and Logarithmic Functions
- 3.9: Inverse Trigonometric Functions
- 3.10: Related Rates
- 3.11: Linearization and Differentials

4: Applications of Derivatives

- 4.1: Maxima and Minima
- 4.2: The Mean Value Theorem
- 4.3: The First Derivative Test
- 4.4: The Second Derivative Test and Curve Sketching
- 4.5: Indeterminate Forms and L'Hôpital's Rule
- 4.6: Applied Optimization Problems
- 4.7: Newton's Method

5: Integration

- 5.1: Antiderivatives
- 5.2: Approximating Areas with Limits of Finite Sums
- 5.3: The Definite Integral
- 5.4: The Fundamental Theorem of Calculus
- 5.5: Substitution with Indefinite Integrals
- 5.6: Substitution with Definite Integrals

6: Applications of Definite Integrals

- 6.1: Volumes Using Cross-Sections
- 6.2: Volumes Using Cylindrical Shells
- 6.3: Arc Length
- 6.4: Areas of Surfaces of Revolution
- 6.5: Work
- 6.6: Moments and Centers of Mass

7: Integrals and Transcendental Functions

- 7.1: The Logarithm Defined as an Integral
- 7.2: Exponential Change and Separable Differential Equations
- 7.3: Hyperbolic Functions

8: Techniques of Integration

- 8.1: Integration by Parts
- 8.2: Trigonometric Integrals
- 8.3: Trigonometric Substitutions
- 8.4: Integration of Rational Functions by Partial Fractions
- 8.5: Integral Tables and Computer Algebra Systems
- 8.6: Numerical Integration
- 8.7: Improper Integrals

9: Infinite Sequence and Series

- 9.1: The Binomial Series and Applications of Taylor Series
- 9.2: Sequences
- 9.3: Infinite Series
- 9.4: The Integral Test
- 9.5: Comparison Tests
- 9.6: The Ratio and Root Tests
- 9.7: Alternating Series, Absolute and Conditional Convergence
- 9.8: Power Series
- 9.9: Taylor and Maclaurin Series
- 9.10: Convergence of Taylor Series

10: Parametric Equations and Polar Coordinates

- 10.1: Parametrizations of Plane Curves
- 10.2: Calculus with Parametric Curves
- 10.3: Polar Coordinates
- 10.4: Graphing in Polar Coordinates
- 10.5: Areas and Lengths in Polar Coordinates
- 10.6: Conics in Polar Coordinates

11: Vectors and the Geometry of Space

- 11.1: Three-Dimensional Coordinate Systems
- 11.2: Vectors
- 11.3: The Dot Product
- 11.4: The Cross Product
- 11.5: Lines and Planes in Space
- 11.6: Cylinders and Quadric Surfaces

12: Vector-Valued Functions and Motion in Space

- 12.1: Curves in Space and Their Tangents
- 12.2: Integrals of Vector Functions; Projectile Motion
- 12.3: Arc Length in Space
- 12.4: Curvature and Normal Vectors of a Curve
- 12.5: Tangential and Normal Components of Acceleration
- 12.6: Velocity and Acceleration in Polar Coordinates

13: Partial Derivatives

- 13.1: Functions of Several Variables
- 13.2: Limits and Continuity in Higher Dimensions
- 13.3: Partial Derivatives
- 13.4: The Chain Rule
- 13.5: Directional Derivatives and Gradient Vectors
- 13.6: Tangent Planes and Differentials
- 13.7: Extreme Values and Saddle Points
- 13.8: Lagrange Multipliers

14: Multiple Integrals

- 14.1: Double and Iterated Integrals over Rectangles
- 14.2: Double Integrals over General Regions
- 14.3: Area by Double Integration
- 14.4: Double Integrals in Polar Form
- 14.5: Triple Integrals in Rectangular Coordinates
- 14.6: Moments and Centers of Mass
- 14.7: Triple Integrals in Cylindrical and Spherical Coordinates
- 14.8: Substitutions in Multiple Integrals

15: Integration in Vector Fields

- 15.1: Line Integrals
- 15.2: Vector Fields and Line Integrals- Work, Circulation, and Flux
- 15.3: Path Independence, Conservative Fields, and Potential Functions
- 15.4: Green's Theorem in the Plane
- 15.5: Surfaces and Area
- 15.6: Surface Integrals
- 15.7: Stokes' Theorem
- 15.8: The Divergence Theorem and a Unified Theory

16: First-Order Differential Equations

- 16.1: Solutions, Slope Fields, and Euler's Method
- 16.2: First-Order Linear Equations
- 16.3: Applications
- 16.4: Graphical Solutions of Autonomous Equations
- 16.5: Systems of Equations and Phase Planes

17: Second-Order Differential Equations

- 17.1: Second-Order Linear Equations
- 17.2: Nonhomogeneous Linear Equations
- 17.3: Applications

- 17.4: Euler Equations
- 17.5: Power Series Solutions

Index

Glossary

Detailed Licensing

Detailed Licensing

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

CHAPTER OVERVIEW

1: Functions and Graphs (Precalculus Review)

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

1.1: Review of Functions
1.2: Basic Classes of Functions
1.3: Trigonometric Functions
1.4: Inverse Functions
1.5: Exponential and Logarithmic Functions

1: Functions and Graphs (Precalculus Review) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

1.1: Review of Functions

Learning Objectives

- Use functional notation to evaluate a function.
- Determine the domain and range of a function.
- Draw the graph of a line.
- Find function values from a given graph.
- Use a graph to determine where a function is increasing, decreasing, or constant.
- Make new functions from composing given functions.
- Find the difference quotient of a function.

In this section, we provide a formal definition of a function and examine several ways in which functions are represented—namely, through tables, formulas, and graphs. We study formal notation and terms related to functions. We also define composition of functions and symmetry properties. Most of this material will be a review for you, but it serves as a handy reference to remind you of some of the algebraic techniques useful for working with functions.

Functions

Given two sets A and B a set with elements that are ordered pairs (x, y) where x is an element of A and y is an element of B, is a relation from A to B. A relation from A to B defines a relationship between those two sets. A function is a special type of relation in which each element of the first set is related to exactly one element of the second set. The element of the first set is called the ***input***; the element of the second set is called the ***output***. Functions are used all the time in mathematics to describe relationships between two sets. For any function, when we know the input, the output is determined, so we say that the output is a function of the input. For example, the area of a square is determined by its side length, so we say that the area (the output) is a function of its side length (the input). The velocity of a ball thrown in the air can be described as a function of the amount of time the ball is in the air. The cost of mailing a package is a function of the weight of the package. Since functions have so many uses, it is important to have precise definitions and terminology to study them.



Figure 1.1.1: A function can be visualized as an input/output device

Definition: Functions

A *function* f consists of a set of inputs, a set of outputs, and a rule for assigning each input to exactly one output. The set of inputs is called the *domain* of the function. The set of outputs is called the *range* of the *function*.

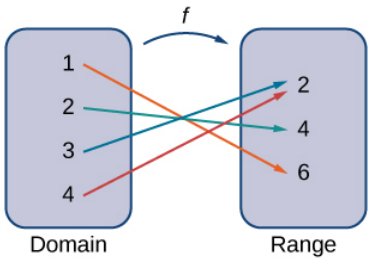


Figure 1.1.2: A function maps every element in the domain to exactly one element in the range. Although each input can be sent to only one output, two different inputs can be sent to the same output.

For example, consider the function f, where the domain is the set of all real numbers and the rule is to square the input. Then, the input $x = 3$ is assigned to the output $3^2 = 9$.

Since every nonnegative real number has a real-value square root, every nonnegative number is an element of the range of this function. Since there is no real number with a square that is negative, the negative real numbers are not elements of the range. We

conclude that the range is the set of nonnegative real numbers.

For a general function f with domain D, we often use x to denote the input and y to denote the output associated with x. When doing so, we refer to x as the **independent variable** and y as the **dependent variable**, because it depends on x. Using function notation, we write $y = f(x)$, and we read this equation as "y equals f of x." For the squaring function described earlier, we write $f(x) = x^2$.

The concept of a function can be visualized using Figures 1.1.1 - 1.1.3.

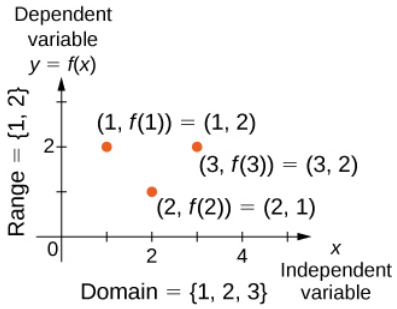


Figure 1.1.3: In this case, a graph of a function f has a domain of $\{1,2,3\}$ and a range of $\{1,2\}$. The independent variable is x and the dependent variable is y.

We can also visualize a function by plotting points (x, y) in the coordinate plane where $y = f(x)$. The graph of a function is the set of all these points. For example, consider the function f, where the domain is the set $D = \{1, 2, 3\}$ and the rule is $f(x) = 3 - x$. In Figure 1.1.4, we plot a graph of this function.

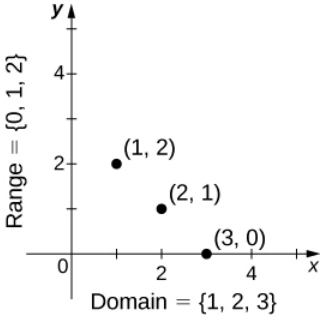


Figure 1.1.4: Here we see a graph of the function f with domain $\{1,2,3\}$ and rule $f(x) = 3 - x$. The graph consists of the points $(x, f(x))$ for all x in the domain.

Every function has a domain. However, sometimes a function is described by an equation, as in $f(x) = x^2$, with no specific domain given. In this case, the domain is taken to be the set of all real numbers x for which $f(x)$ is a real number. For example, since any real number can be squared, if no other domain is specified, we consider the domain of $f(x) = x^2$ to be the set of all real numbers. On the other hand, the square root function $f(x) = \sqrt{x}$ only gives a real output if x is nonnegative. Therefore, the domain of the function $f(x) = \sqrt{x}$ is the set of nonnegative real numbers, sometimes called the natural domain.

For the functions $f(x) = x^2$ and $f(x) = \sqrt{x}$, the domains are sets with an infinite number of elements. Clearly we cannot list all these elements. When describing a set with an infinite number of elements, it is often helpful to use set-builder or interval notation. When using set-builder notation to describe a subset of all real numbers, denoted R, we write

$$\{x \mid x \, has \, some \, property\}.$$

We read this as the set of real numbers x such that x has some property. For example, if we were interested in the set of real numbers that are greater than one but less than five, we could denote this set using set-builder notation by writing

$$\{x \mid 1 < x < 5\}.$$

A set such as this, which contains all numbers greater than a and less than b, can also be denoted using the *interval notation* (a, b). Therefore,

$$(1, 5) = \{x \mid 1 < x < 5\}.$$

The numbers 1 and 5 are called the endpoints of this set. If we want to consider the set that includes the endpoints, we would denote this set by writing

$$[1,5] = \{x \mid 1 \leq x \leq 5\}.$$

We can use similar notation if we want to include one of the endpoints, but not the other. To denote the set of nonnegative real numbers, we would use the set-builder notation

$$\{x \mid x \geq 0\}.$$

The smallest number in this set is zero, but this set does not have a largest number. Using interval notation, we would use the symbol ∞, which refers to positive infinity, and we would write the set as

$$[0,\infty) = \{x \mid x \geq 0\}.$$

It is important to note that ∞ is not a real number. It is used symbolically here to indicate that this set includes all real numbers greater than or equal to zero. Similarly, if we wanted to describe the set of all nonpositive numbers, we could write

$$(-\infty, 0] = \{x \mid x \leq 0\}.$$

Here, the notation $-\infty$ refers to negative infinity, and it indicates that we are including all numbers less than or equal to zero, no matter how small. The set

$$(-\infty,\infty) = \{x \mid x \text{ is any real number}\}$$

refers to the set of all real numbers. Some functions are defined using different equations for different parts of their domain. These types of functions are known as piecewise-defined functions. For example, suppose we want to define a function f with a domain that is the set of all real numbers such that $f(x) = 3x+1$ for $x \geq 2$ and $f(x) = x^2$ for $x < 2$. We denote this function by writing

$$f(x) = \begin{cases} 3x+1, & \text{if } x \geq 2 \\ x^2, & \text{if } x < 2 \end{cases}$$

When evaluating this function for an input x, the equation to use depends on whether $x \geq 2$ or $x < 2$. For example, since $5 > 2$, we use the fact that $f(x) = 3x+1$ for $x \geq 2$ and see that $f(5) = 3(5)+1 = 16$. On the other hand, for $x = -1$, we use the fact that $f(x) = x^2$ for $x < 2$ and see that $f(-1) = 1$.

Example 1.1.1: Evaluating Functions

For the function $f(x) = 3x^2 + 2x - 1$, evaluate:

a. $f(-2)$
b. $f(\sqrt{2})$
c. $f(a+h)$

Solution

Substitute the given value for x in the formula for $f(x)$.

a. $f(-2) = 3(-2)^2 + 2(-2) - 1 = 12 - 4 - 1 = 7$
b. $f(\sqrt{2}) = 3(\sqrt{2})^2 + 2\sqrt{2} - 1 = 6 + 2\sqrt{2} - 1 = 5 + 2\sqrt{2}$
c. $f(a+h) = 3(a+h)^2 + 2(a+h) - 1 = 3(a^2 + 2ah + h^2) + 2a + 2h - 1 = 3a^2 + 6ah + 3h^2 + 2a + 2h - 1$

Exercise 1.1.1

For $f(x) = x^2 - 3x + 5$, evaluate $f(1)$ and $f(a+h)$.

Hint

Substitute 1 and $a+h$ for x in the formula for $f(x)$.

Answer

$f(1) = 3$ and $f(a+h) = a^2 + 2ah + h^2 - 3a - 3h + 5$

Example 1.1.2: Finding Domain and Range

For each of the following functions, determine the i. domain and ii. range.

a. $f(x)=(x-4)^2+5$
b. $f(x)=\sqrt{3x+2}-1$
c. $f(x)=\dfrac{3}{x-2}$

Solution

a. Consider $f(x)=(x-4)^2+5$.

1.Since $f(x)=(x-4)^2+5$ is a real number for any real number x, the domain of f is the interval $(-\infty,\infty)$.

2. Since $(x-4)^2\geq 0$, we know $f(x)=(x-4)^2+5\geq 5$. Therefore, the range must be a subset of $\{y\,|\,y\geq 5\}$. To show that every element in this set is in the range, we need to show that for a given y in that set, there is a real number x such that $f(x)=(x-4)^2+5=y$. Solving this equation for x, we see that we need x such that

$$(x-4)^2=y-5.$$

This equation is satisfied as long as there exists a real number x such that

$$x-4=\pm\sqrt{y-5}$$

Since $y\geq 5$, the square root is well-defined. We conclude that for $x=4\pm\sqrt{y-5}$, $f(x)=y$, and therefore the range is $\{y\,|\,y\geq 5\}$.

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

1.To find the domain of f, we need the expression $3x+2\geq 0$. Solving this inequality, we conclude that the domain is $\{x\,|\,x\geq -2/3\}$.

2.To find the range of f, we note that since $\sqrt{3x+2}\geq 0$, $f(x)=\sqrt{3x+2}-1\geq -1$. Therefore, the range of f must be a subset of the set $\{y\,|\,y\geq -1\}$. To show that every element in this set is in the range of f, we need to show that for all y in this set, there exists a real number x in the domain such that $f(x)=y$. Let $y\geq -1$. Then, $f(x)=y$ if and only if

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

Solving this equation for x, we see that x must solve the equation

$$\sqrt{3x+2}=y+1.$$

Since $y\geq -1$, such an x could exist. Squaring both sides of this equation, we have $3x+2=(y+1)^2$.

Therefore, we need

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

which implies

$$x=\tfrac{1}{3}(y+1)^2-\tfrac{2}{3}.$$

We just need to verify that x is in the domain of f. Since the domain of f consists of all real numbers greater than or equal to $\frac{-2}{3}$, and

$$\tfrac{1}{3}(y+1)^2-\tfrac{2}{3}\geq -\tfrac{2}{3},$$

there does exist an x in the domain of f. We conclude that the range of f is $\{y\,|\,y\geq -1\}$.

c. Consider $f(x)=\dfrac{3}{x-2}$.

1.Since $3/(x-2)$ is defined when the denominator is nonzero, the domain is $\{x\,|\,x\neq 2\}$.

2.To find the range of f, we need to find the values of y such that there exists a real number x in the domain with the property that

$$\frac{3}{x-2} = y.$$

Solving this equation for x, we find that

$$x = \frac{3}{y} + 2.$$

Therefore, as long as $y \neq 0$, there exists a real number x in the domain such that $f(x) = y$. Thus, the range is $\{y \mid y \neq 0\}$.

? Exercise 1.1.2

Find the domain and range for $f(x) = \sqrt{4-2x} + 5$.

Hint

Use $4 - 2x \geq 0$.

Answer

Domain = $\{x \mid x \leq 2\}$ and range = $\{y \mid y \geq 5\}$

Finding Function Values from a Graph

Evaluating a function using a graph also requires finding the corresponding output value for a given input value, only in this case, we find the output value by looking at the graph. Solving a function equation using a graph requires finding all instances of the given output value on the graph and observing the corresponding input value(s).

✔ Example 1.1.3: Reading Function Values from a Graph

Given the graph in Figure 1.1.7,

a. Evaluate $f(2)$.
b. Solve $f(x) = 4$.

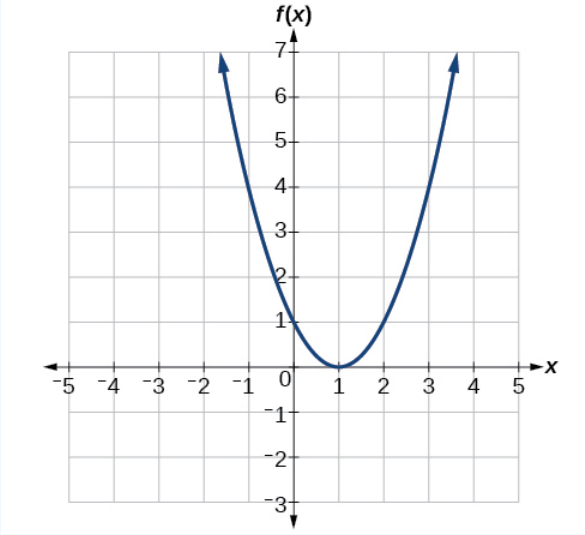
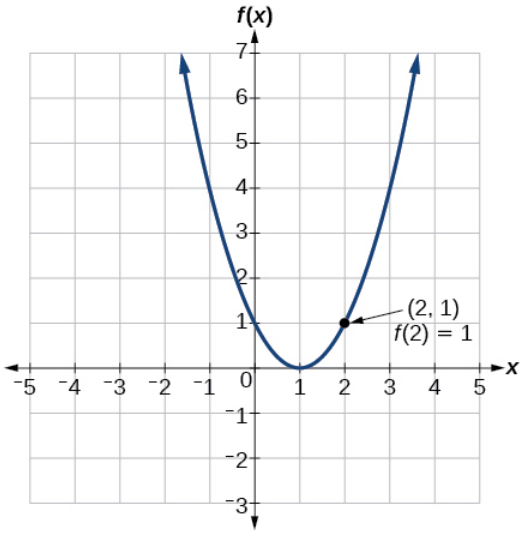


Figure 1.1.7: Graph of a positive parabola centered at $(1, 0)$.

Solution

To evaluate $f(2)$, locate the point on the curve where $x = 2$, then read the y-coordinate of that point. The point has coordinates $(2, 1)$, so $f(2) = 1$. See Figure 1.1.8.



1.1.8: Graph of a positive parabola centered at $(1,0)$ with the labeled point $(2,1)$ where $f(2)=1$.

To solve $f(x)=4$, we find the output value 4 on the vertical axis. Moving horizontally along the line $y=4$, we locate two points of the curve with output value 4: $(-1,4)$ and $(3,4)$. These points represent the two solutions to $f(x)=4$: -1 or 3. This means $f(-1)=4$ and $f(3)=4$, or when the input is -1 or 3, the output is 4. See Figure 1.1.9.

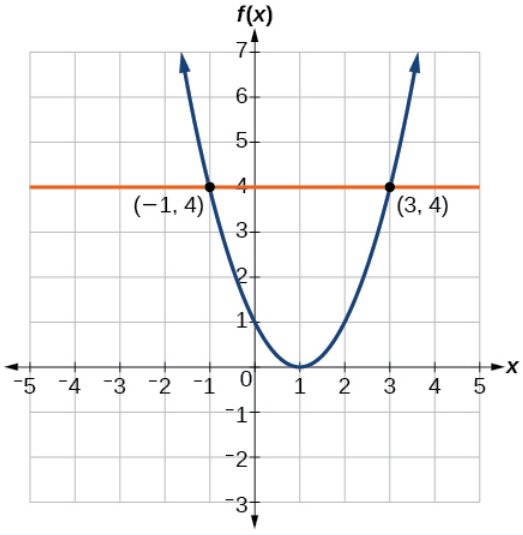


Figure 1.1.9: Graph of an upward-facing parabola with a vertex at $(0,1)$ and labeled points at $(-1,4)$ and $(3,4)$. A line at $y=4$ intersects the parabola at the labeled points.

? Exercise 1.1.3

Given the graph in Figure 1.1.7, solve $f(x)=1$.

Answer

$x=0$ or $x=2$

Using a Graph to Determine Where a Function is Increasing, Decreasing, or Constant

As part of exploring how functions change, we can identify intervals over which the function is changing in specific ways. We say that a function is increasing on an interval if the function values increase as the input values increase within that interval. Similarly, a function is decreasing on an interval if the function values decrease as the input values increase over that interval. The average rate of change of an increasing function is positive, and the average rate of change of a decreasing function is negative. Figure 1.1.3 shows examples of increasing and decreasing intervals on a function.

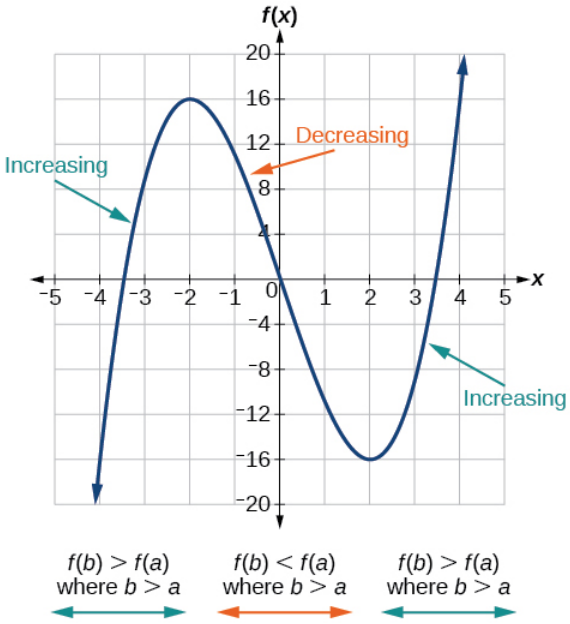


Figure 1.1.3: The function $f(x) = x^3 - 12x$ is increasing on $(-\infty, -2) \cup (2, \infty)$ and is decreasing on $(-2, 2)$.

While some functions are increasing (or decreasing) over their entire domain, many others are not. A value of the input where a function changes from increasing to decreasing (as we go from left to right, that is, as the input variable increases) is called a **local maximum**. If a function has more than one, we say it has local maxima. Similarly, a value of the input where a function changes from decreasing to increasing as the input variable increases is called a **local minimum**. The plural form is "local minima." Together, local maxima and minima are called **local extrema**, or local extreme values, of the function. (The singular form is "extremum.") Often, the term local is replaced by the term relative. In this text, we will use the term local.

Clearly, a function is neither increasing nor decreasing on an interval where it is constant. A function is also neither increasing nor decreasing at extrema. Note that we have to speak of local extrema, because any given local extremum as defined here is not necessarily the highest maximum or lowest minimum in the function's entire domain.

For the function whose graph is shown in Figure 1.1.4, the local maximum is 16, and it occurs at $x = -2$. The local minimum is -16 and it occurs at $x = 2$.

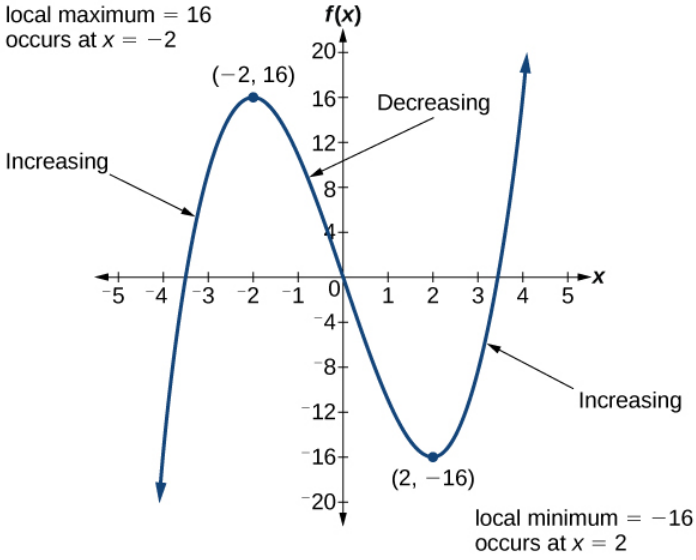


Figure 1.1.4: Graph of a polynomial that shows the increasing and decreasing intervals and local maximum.maximum

To locate the local maxima and minima from a graph, we need to observe the graph to determine where the graph attains its highest and lowest points, respectively, within an open interval. Like the summit of a roller coaster, the graph of a function is higher at a

local maximum than at nearby points on both sides. The graph will also be lower at a local minimum than at neighboring points. Figure 1.1.5 illustrates these ideas for a local maximum.

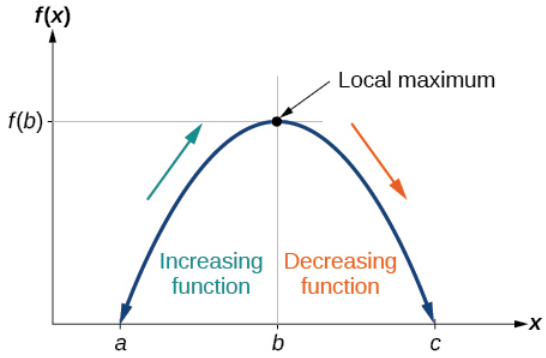


Figure 1.1.5: Definition of a local maximum

These observations lead us to a formal definition of local extrema.

Local Minima and Local Maxima

- A function f is an **increasing function** on an open interval if $f(b) > f(a)$ for every a, b interval where $b > a$.
- A function f is a **decreasing function** on an open interval if $f(b) < f(a)$ for every a, b interval where $b > a$.

A function f has a local maximum at a point b in an open interval (a, c) if $f(b)$ is greater than or equal to $f(x)$ for every point x (x does not equal b) in the interval. Likewise, f has a local minimum at a point b in (a, c) if $f(b)$ is less than or equal to $f(x)$ for every x (x does not equal b) in the interval.

Example 1.1.4 Finding Increasing and Decreasing Intervals on a Graph

Given the function $p(t)$ in Figure 1.1.6, identify the intervals on which the function appears to be increasing.

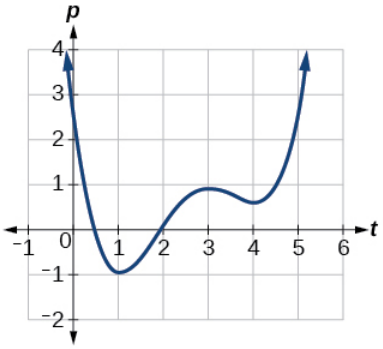


Figure 1.1.6: Graph of a polynomial.

Solution

We see that the function is not constant on any interval. The function is increasing where it slants upward as we move to the right and decreasing where it slants downward as we move to the right. The function appears to be increasing from $t = 1$ to $t = 3$ and from $t = 4$ on.

In interval notation, we would say the function appears to be increasing on the interval $(1, 3)$ and the interval $(4, \infty)$.

Analysis

Notice in this example that we used open intervals (intervals that do not include the endpoints), because the function is neither increasing nor decreasing at $t = 1$, $t = 3$, and $t = 4$. These points are the local extrema (two minima and a maximum).

Example 1.1.5: Finding Local Extrema from a Graph

Graph the function $f(x) = \frac{2}{x} + \frac{x}{3}$. Then use the graph to estimate the local extrema of the function and to determine the intervals on which the function is increasing.

Solution

Using technology, we find that the graph of the function looks like that in Figure 1.1.7. It appears there is a low point, or local minimum, between $x = 2$ and $x = 3$, and a mirror-image high point, or local maximum, somewhere between $x = -3$ and $x = -2$

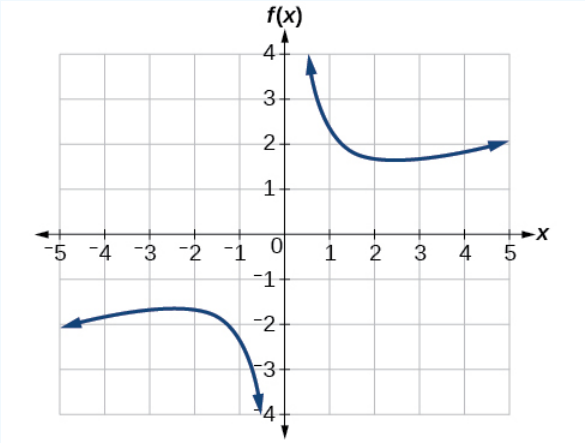


Figure 1.1.7: Graph of a reciprocal function.

Analysis

Most graphing calculators and graphing utilities can estimate the location of maxima and minima. Figure 1.1.8 provides screen images from two different technologies, showing the estimate for the local maximum and minimum.

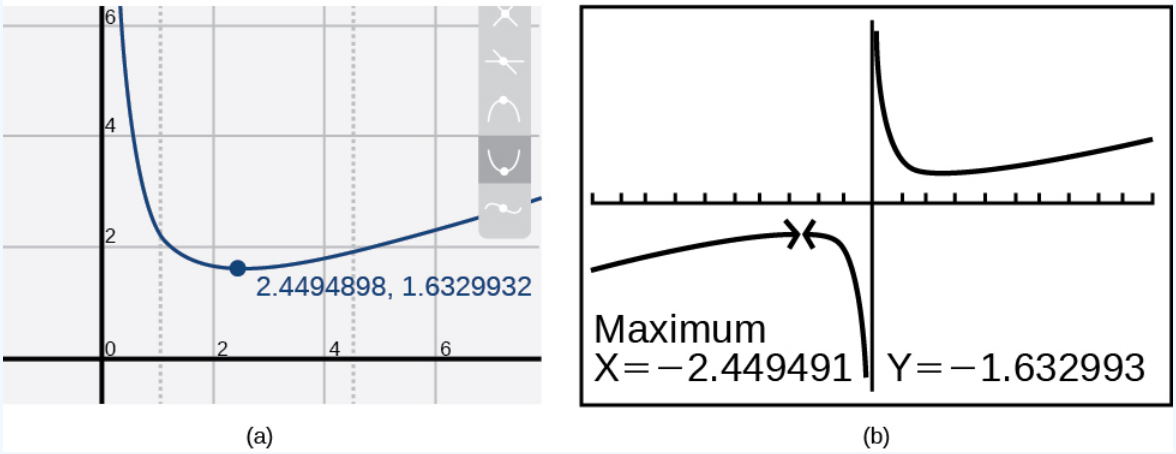


Figure 1.1.8: Graph of the reciprocal function on a graphing calculator.

Based on these estimates, the function is increasing on the interval $(-\infty, -2.449)$ and $(2.449, \infty)$. Notice that, while we expect the extrema to be symmetric, the two different technologies agree only up to four decimals due to the differing approximation algorithms used by each. (The exact location of the extrema is at $\pm\sqrt{6}$, but determining this requires calculus.)

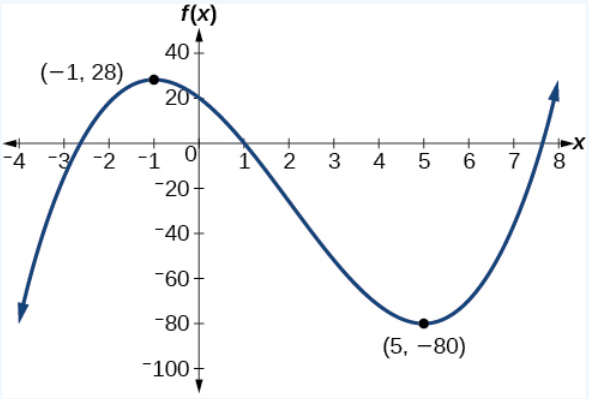
Exercise 1.1.5

Graph the function $f(x) = x^3 - 6x^2 - 15x + 20$ to estimate the local extrema of the function. Use these to determine the intervals on which the function is increasing and decreasing.

Solution

https://math.libretexts.org/@go/page/139016

The local maximum appears to occur at $(-1, 28)$, and the local minimum occurs at $(5, -80)$. The function is increasing on $(-\infty, -1) \cup (5, \infty)$ and decreasing on $(-1, 5)$.



Graph of a polynomial with a local maximum at (-1, 28) and local minimum at (5, -80).

Example 1.1.6: Finding Local Maxima and Minima from a Graph

For the function f whose graph is shown in Figure 1.1.9, find all local maxima and minima.

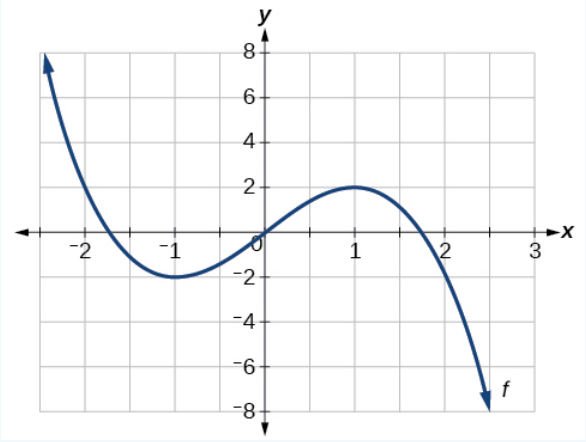


Figure 1.1.9: Graph of a polynomial.

Solution

Observe the graph of f. The graph attains a local maximum at $x = 1$ because it is the highest point in an open interval around $x = 1$.The local maximum is the y-coordinate at $x = 1$, which is 2.

The graph attains a local minimum at $x = -1$ because it is the lowest point in an open interval around $x = -1$. The local minimum is the y-coordinate at $x = -1$, which is −2.

Use A Graph to Locate the Absolute Maximum and Absolute Minimum

There is a difference between locating the highest and lowest points on a graph in a region around an open interval (locally) and locating the highest and lowest points on the graph for the entire domain. The y-coordinates (output) at the highest and lowest points are called the **absolute maximum** and **absolute minimum**, respectively. To locate absolute maxima and minima from a graph, we need to observe the graph to determine where the graph attains it highest and lowest points on the domain of the function (Figure 1.1.13).

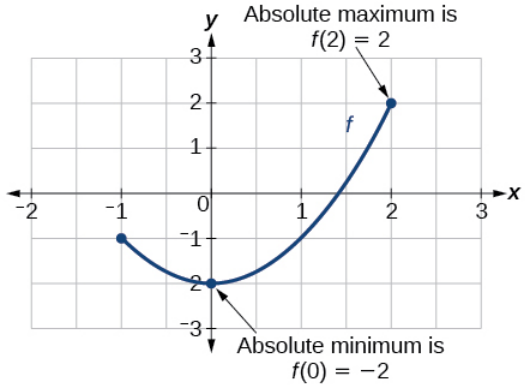


Figure 1.1.13: Graph of a segment of a parabola with an absolute minimum at (0, -2) and absolute maximum at (2, 2).

Not every function has an absolute maximum or minimum value. The toolkit function $f(x) = x^3$ is one such function.

Absolute Maxima and Minima

- The **absolute maximum** of f at $x = c$ is $f(c)$ where $f(c) \geq f(x)$ for all x in the domain of f.
- The **absolute minimum** of f at $x = d$ is $f(d)$ where $f(d) \leq f(x)$ for all x in the domain of f.

Example 1.1.7: Finding Absolute Maxima and Minima from a Graph

For the function f shown in Figure 1.1.14, find all absolute maxima and minima.

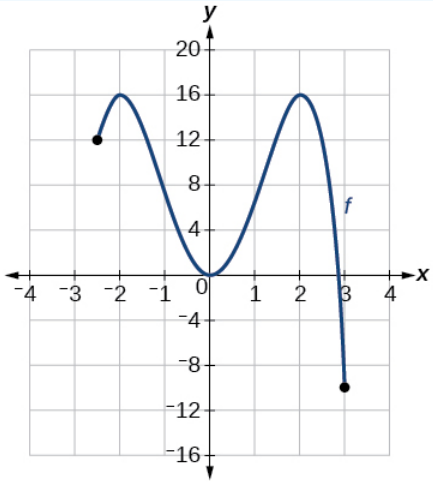


Figure 1.1.14: Graph of a polynomial.

Solution

Observe the graph of f. The graph attains an absolute maximum in two locations, $x = -2$ and $x = 2$, because at these locations, the graph attains its highest point on the domain of the function. The absolute maximum is the y-coordinate at $x = -2$ and $x = 2$, which is 16.

The graph attains an absolute minimum at x=3, because it is the lowest point on the domain of the function's graph. The absolute minimum is the y-coordinate at x=3,which is−10.

Function Composition

When we compose functions, we take a function of a function. For example, suppose the temperature T on a given day is described as a function of time t (measured in hours after midnight) as in Table 1.1.1. Suppose the cost C, to heat or cool a building for 1 hour, can be described as a function of the temperature T. Combining these two functions, we can describe the cost of heating or cooling a building as a function of time by evaluating $C(T(t))$. We have defined a new function, denoted $C \circ T$, which is defined such that $(C \circ T)(t) = C(T(t))$ for all t in the domain of T. This new function is called a composite function. We note that since

cost is a function of temperature and temperature is a function of time, it makes sense to define this new function $(C \circ T)(t)$. It does not make sense to consider $(T \circ C)(t)$, because temperature is not a function of cost.

Definition: Composite Functions

Consider the function f with domain A and range B, and the function g with domain D and range E. If B is a subset of D, then the composite function $(g \circ f)(x)$ is the function with domain A such that

$$(g \circ f)(x) = g(f(x))$$

A composite function $g \circ f$ can be viewed in two steps. First, the function f maps each input x in the domain of f to its output $f(x)$ in the range of f. Second, since the range of f is a subset of the domain of g, the output $f(x)$ is an element in the domain of g, and therefore it is mapped to an output $g(f(x))$ in the range of g. In Figure 1.1.11, we see a visual image of a composite function.

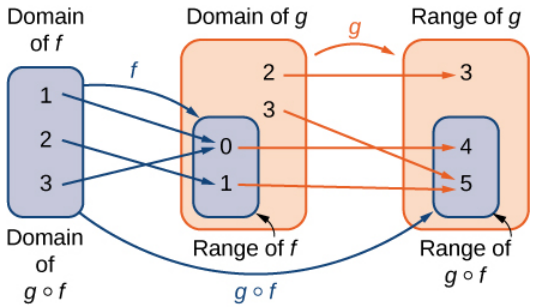


Figure 1.1.11: For the composite function $g \circ f$, we have $(g \circ f)(1) = 4$, $(g \circ f)(2) = 5$, and $(g \circ f)(3) = 4$.

Example 1.1.7: Compositions of Functions Defined by Formulas

Consider the functions $f(x) = x^2 + 1$ and $g(x) = 1/x$.

a. Find $(g \circ f)(x)$ and state its domain and range.
b. Evaluate $(g \circ f)(4)$, $(g \circ f)(-1/2)$.
c. Find $(f \circ g)(x)$ and state its domain and range.
d. Evaluate $(f \circ g)(4)$, $(f \circ g)(-1/2)$.

Solution

1. We can find the formula for $(g \circ f)(x)$ in two different ways. We could write

$$(g \circ f)(x) = g(f(x)) = g(x^2 + 1) = \frac{1}{x^2 + 1} \quad .$$

Alternatively, we could write

$$(g \circ f)(x) = g(f(x)) = \frac{1}{f(x)} = \frac{1}{x^2 + 1}.$$

Since $x^2 + 1 \neq 0$ for all real numbers x, the domain of $(g \circ f)(x)$ is the set of all real numbers. Since $0 < 1/(x^2 + 1) \leq 1$, the range is, at most, the interval $(0, 1]$. To show that the range is this entire interval, we let $y = 1/(x^2 + 1)$ and solve this equation for x to show that for all y in the interval $(0, 1]$, there exists a real number x such that $y = 1/(x^2 + 1)$. Solving this equation for x, we see that $x^2 + 1 = 1/y$, which implies that

$$x = \pm\sqrt{\tfrac{1}{y} - 1}$$

If y is in the interval $(0, 1]$, the expression under the radical is nonnegative, and therefore there exists a real number x such that $1/(x^2 + 1) = y$. We conclude that the range of $g \circ f$ is the interval $(0, 1]$.

2. $(g \circ f)(4) = g(f(4)) = g(4^2 + 1) = g(17) = \frac{1}{17}$

$(g \circ f)(-\frac{1}{2}) = g(f(-\frac{1}{2})) = g((-\frac{1}{2})^2 + 1) = g(\frac{5}{4}) = \frac{4}{5}$

3. We can find a formula for $(f \circ g)(x)$ in two ways. First, we could write

$$(f\circ g)(x)=f(g(x))=f(\tfrac{1}{x})=(\tfrac{1}{x})^2+1.$$

Alternatively, we could write

$$(f\circ g)(x)=f(g(x))=(g(x))^2+1=(\tfrac{1}{x})^2+1.$$

The domain of $f\circ g$ is the set of all real numbers x such that $x\neq 0$. To find the range of f, we need to find all values y for which there exists a real number $x\neq 0$ such that

$$\left(\frac{1}{x}\right)^2+1=y.$$

Solving this equation for x, we see that we need x to satisfy

$$\left(\frac{1}{x}\right)^2=y-1,$$

which simplifies to

$$\frac{1}{x}=\pm\sqrt{y-1}$$

Finally, we obtain

$$x=\pm\frac{1}{\sqrt{y-1}}.$$

Since $1/\sqrt{y-1}$ is a real number if and only if $y>1$, the range of f is the set $\{y\,|\,y\geq 1\}$.

4.$(f\circ g)(4)=f(g(4))=f(\tfrac{1}{4})=(\tfrac{1}{4})^2+1=\tfrac{17}{16}$

$(f\circ g)(-\tfrac{1}{2})=f(g(-\tfrac{1}{2}))=f(-2)=(-2)^2+1=5$

In Example 1.1.7, we can see that $(f\circ g)(x)\neq(g\circ f)(x)$. This tells us, in general terms, that the order in which we compose functions matters.

? Exercise 1.1.7

Let $f(x)=2-5x$. Let $g(x)=\sqrt{x}$. Find $(f\circ g)(x)$.

Solution

$(f\circ g)(x)=2-5\sqrt{x}$.

✓ Example 1.1.8: Composition of Functions Defined by Tables

Consider the functions f and g described by

x	-3	-2	-1	0	1	2	3	4
$f(x)$	0	4	2	4	-2	0	-2	4

x	-4	-2	0	2	4
$g(x)$	1	0	3	0	5

a. Evaluate $(g\circ f)(3)$,$(g\circ f)(0)$.
b. State the domain and range of $(g\circ f)(x)$.
c. Evaluate $(f\circ f)(3)$,$(f\circ f)(1)$.
d. State the domain and range of $(f\circ f)(x)$.

Solution:

1. $(g\circ f)(3)=g(f(3))=g(-2)=0$

$(g \circ f)(0) = g(4) = 5$

2.The domain of $g \circ f$ is the set $\{-3, -2, -1, 0, 1, 2, 3, 4\}$.Since the range of f is the set $\{-2, 0, 2, 4\}$,the range of $g \circ f$ is the set $\{0, 3, 5\}$.

3. $(f \circ f)(3) = f(f(3)) = f(-2) = 4$

$(f \circ f)(1) = f(f(1)) = f(-2) = 4$

4.The domain of $f \circ f$ is the set $\{-3, -2, -1, 0, 1, 2, 3, 4\}$.Since the range of f is the set $\{-2, 0, 2, 4\}$,the range of $f \circ f$ is the set $\{0, 4\}$.

Example 1.1.9: Application Involving a Composite Function

A store is advertising a sale of 20% off all merchandise. Caroline has a coupon that entitles her to an additional 15% off any item, including sale merchandise. If Caroline decides to purchase an item with an original price of x dollars, how much will she end up paying if she applies her coupon to the sale price? Solve this problem by using a composite function.

Solution

Since the sale price is 20% off the original price, if an item is x dollars, its sale price is given by $f(x) = 0.80x$. Since the coupon entitles an individual to 15% off the price of any item, if an item is y dollars, the price, after applying the coupon, is given by g(y)=0.85y. Therefore, if the price is originally x dollars, its sale price will be $f(x) = 0.80x$ and then its final price after the coupon will be $g(f(x)) = 0.85(0.80x) = 0.68x$

Exercise 1.1.9

If items are on sale for 10% off their original price, and a customer has a coupon for an additional 30% off, what will be the final price for an item that is originally x dollars, after applying the coupon to the sale price?

Hint

The sale price of an item with an original price of x dollars is $f(x) = 0.90x$. The coupon price for an item that is y dollars is $g(y) = 0.70y$.

Solution

$(g \circ f)(x) = 0.63x$

The Difference Quotient

Let's imagine a point on the curve of function f at x as shown in Figure 13. The coordinates of the point are $(x, f(x))$. Connect this point with a second point on the curve a little to the right of x, with an x-value increased by some small real number h. The coordinates of this second point are $(x+h, f(x+h))$ for some positive-value h.

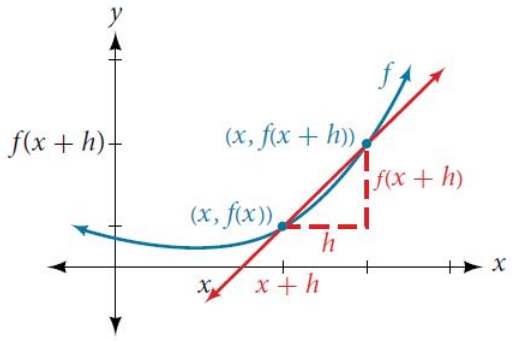


Figure 1.1.12: Connecting point a with a point just beyond allows us to measure a slope close to that of a tangent line at x.

We can calculate the slope of the line connecting the two points $(x, f(x))$ and $(x+h, f(x+h))$, called a **secant line**, by applying the slope formula,

$$\text{slope} = \frac{\text{change in } y}{\text{change in } x}$$

$$\text{slope} = \frac{\text{change in } y}{\text{change in } x}$$

We use the notation m_{sec} to represent the slope of the secant line connecting two points.

$$m_{\text{sec}} = \frac{f(x+h) - f(x)}{(x+h) - (x)} \tag{1.1.1}$$

$$= \frac{f(x+h) - f(x)}{h} \tag{1.1.2}$$

The slope m_{sec} equals the *difference quotient* between two points $(x, f(x))$ and $(x+h, f(x+h))$.

Definition: The Difference Quotient

The **difference quotient** between two points $(x, f(x))$ and $(x+h, f(x+h))$ on the curve of f is the slope of the line connecting the two points and is given by

$$\text{difference quotient} = \frac{f(x+h) - f(x)}{h}$$

Example 1.1.10: Finding the difference quotient (Quadratic function)

Find the difference quotient if $(x) = 2x^2 - x$.

Solution

We use the difference quotient formula.

$\frac{f(x+h) - f(x)}{h}$	Start with the difference quotient	(1.1.3)
$\frac{2(x+h)^2 - (x+h) - (2x^2 - x)}{h}$	Replace x with (x+h)	(1.1.4)
$\frac{2(x^2 + 2xh + h^2) - x - h - 2x^2 + x}{h}$	Distribute	(1.1.5)
$\frac{2x^2 + 4xh + 2h^2 - x - h - 2x^2 + x}{h}$	Distribute	(1.1.6)
$\frac{4xh + 2h^2 - h}{h}$	Simplify like terms	(1.1.7)
$4x + 2h - 1$	This is the difference quotient	(1.1.8)

Exercise 1.1.10

Find the difference quotient if $f(x) = x^2 + 2x - 8$.

Answer

$2x + h + 2$

Example 1.1.11: Finding the difference quotient (Rational function)

Find the difference quotient if $f(x) = \frac{1}{x}$.

Solution

We use the difference quotient formula.

$\frac{f(x+h)-f(x)}{h}$	Start with the difference quotient	(1.1.9)
$\frac{\frac{1}{x+h}-\frac{1}{x}}{h}$	Replace x with (x+h)	(1.1.10)
$\frac{\frac{x}{x(x+h)}-\frac{x+h}{x(x+h)}}{h}$	Get common denominators	(1.1.11)
$\frac{\frac{x-x-h}{x(x+h)}}{h}$	Subtract the fractions	(1.1.12)
$\frac{-h}{x(x+h)h}$	Flip and multiply	(1.1.13)
$\frac{-1}{x(x+h)}$	This is the difference quotient	(1.1.14)

? Exercise 1.1.11

Find the difference quotient if $f(x)=-\frac{2}{x}$.

Answer

$$\frac{2}{x(x+h)}$$

This page titled 1.1: Review of Functions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin "Jed" Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **1.1: Review of Functions** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **1.1: Functions and Function Notation** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.
- **1.3: Rates of Change and Behavior of Graphs** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

1.2: Basic Classes of Functions

Learning Objectives

- Calculate the slope of a linear function and interpret its meaning.
- Sketch the graph of a function that has been shifted, stretched, or reflected from its initial graph position.
- Evaluate a piecewise-defined function.
- Graph a quadratic function by finding the vertex and intercepts.

We have studied the general characteristics of functions, so now let's examine some specific classes of functions. We begin by reviewing the basic properties of linear and quadratic functions, and then generalize to include higher-degree polynomials. By combining root functions with polynomials, we can define general algebraic functions and distinguish them from the transcendental functions we examine later in this chapter. We finish the section with examples of piecewise-defined functions and take a look at how to sketch the graph of a function that has been shifted, stretched, or reflected from its initial form.

Linear Functions and Slope

The easiest type of function to consider is a **linear function**. Linear functions have the form $f(x) = ax + b$, where a and b are constants. In Figure 1.2.1, we see examples of linear functions when a is positive, negative, and zero. Note that if $a > 0$, the graph of the line rises as x increases. In other words, $f(x) = ax + b$ is increasing on $(-\infty, \infty)$. If $a < 0$, the graph of the line falls as x increases. In this case, $f(x) = ax + b$ is decreasing on $(-\infty, \infty)$. If $a = 0$, the line is horizontal.

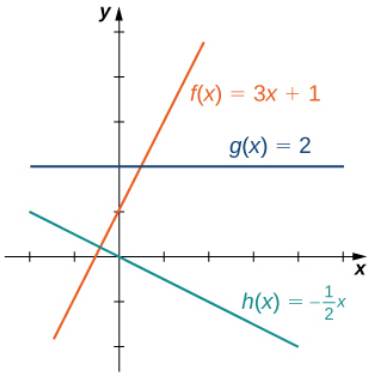


Figure 1.2.1: These linear functions are increasing or decreasing on (∞, ∞) and one function is a horizontal line.

As suggested by Figure 1.2.1, the graph of any linear function is a line. One of the distinguishing features of a line is its slope. The **slope** is the change in y for each unit change in x. The slope measures both the steepness and the direction of a line. If the slope is positive, the line points upward when moving from left to right. If the slope is negative, the line points downward when moving from left to right. If the slope is zero, the line is horizontal. To calculate the slope of a line, we need to determine the ratio of the change in y versus the change in x. To do so, we choose any two points (x_1, y_1) and (x_2, y_2) on the line and calculate $\dfrac{y_2 - y_1}{x_2 - x_1}$. In Figure 1.2.2, we see this ratio is independent of the points chosen.

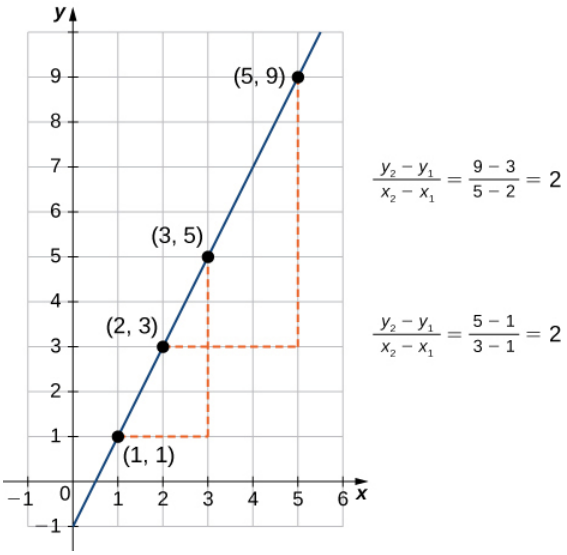


Figure 1.2.2: For any linear function, the slope $(y_2 - y_1)/(x_2 - x_1)$ is independent of the choice of points (x_1, y_1) and (x_2, y_2) on the line.

Definition: Slope of a Linear Function

Consider line L passing through points (x_1, y_1) and (x_2, y_2). Let $\Delta y = y_2 - y_1$ and $\Delta x = x_2 - x_1$ denote the changes in y and x,respectively. The slope of the line is

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x}$$

We now examine the relationship between slope and the formula for a linear function. Consider the linear function given by the formula $f(x) = ax + b$. As discussed earlier, we know the graph of a linear function is given by a line. We can use our definition of slope to calculate the slope of this line. As shown, we can determine the slope by calculating $(y_2 - y_1)/(x_2 - x_1)$ for any points (x_1, y_1) and (x_2, y_2) on the line. Evaluating the function f at $x = 0$, we see that $(0, b)$ is a point on this line. Evaluating this function at $x = 1$, we see that $(1, a + b)$ is also a point on this line. Therefore, the slope of this line is

$$\frac{(a+b) - b}{1 - 0} = a.$$

We have shown that the coefficient a is the slope of the line. We can conclude that the formula $f(x) = ax + b$ describes a line with slope a. Furthermore, because this line intersects the y-axis at the point $(0, b)$, we see that the y-intercept for this linear function is $(0, b)$. We conclude that the formula $f(x) = ax + b$ tells us the slope, a, and the y-intercept, $(0, b)$, for this line. Since we often use the symbol m to denote the slope of a line, we can write

$$\underbrace{f(x) = mx + b}_{\text{slope-intercept form}}$$

to denote the **slope-intercept form** of a linear function.

Sometimes it is convenient to express a linear function in different ways. For example, suppose the graph of a linear function passes through the point (x_1, y_1) and the slope of the line is m. Since any other point $(x, f(x))$ on the graph of f must satisfy the equation

$$m = \frac{f(x) - y_1}{x - x_1},$$

this linear function can be expressed by writing

$$\underbrace{f(x) - y_1 = m(x - x_1)}_{\text{point-slope equation}}.$$

We call this equation the **point-slope equation** for that linear function.

Since every nonvertical line is the graph of a linear function, the points on a nonvertical line can be described using the slope-intercept or point-slope equations. However, a vertical line does not represent the graph of a function and cannot be expressed in either of these forms. Instead, a vertical line is described by the equation $x = k$ for some constant k. Since neither the slope-intercept form nor the point-slope form allows for vertical lines, we use the notation

$$\underbrace{ax + by = c}_{\text{standard form}},$$

where a, b are both not zero, to denote the **standard form of a line.**

Definition: Point-Slope Equation, and the Slope-Intercept Form and Standard Form of the Equation of a Line

Consider a line passing through the point (x_1, y_1) with slope m. The equation

$$y - y_1 = m(x - x_1)$$

is the *point-slope equation* for that line.

Consider a line with slope m and y-intercept $(0, b)$. The equation

$$y = mx + b$$

is an equation for that line in *slope-intercept form.*

The *standard form of a line* is given by the equation

$$ax + by = c,$$

where a and b are both not zero. This form is more general because it allows for a vertical line, $x = k$.

Example 1.2.1: Finding the Slope and Equations of Lines

Consider the line passing through the points $(11, -4)$ and $(-4, 5)$, as shown in Figure 1.2.3.

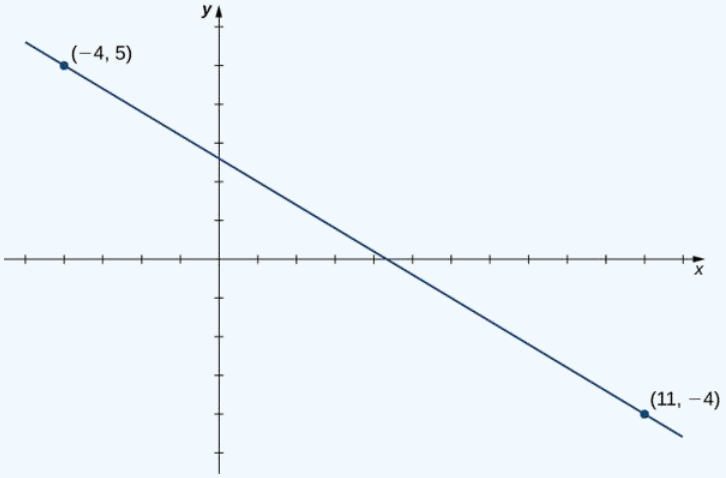


Figure 1.2.3: Finding the equation of a linear function with a graph that is a line between two given points.

1. Find the slope of the line.
2. Find an equation for this linear function in point-slope form.
3. Find an equation for this linear function in slope-intercept form.

Solution

1. The slope of the line is

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{5 - (-4)}{-4 - 11} = -\frac{9}{15} = -\frac{3}{5}.$$

2. To find an equation for the linear function in point-slope form, use the slope $m = -3/5$ and choose any point on the line. If we choose the point $(11, -4)$, we get the equation

$$f(x) + 4 = -\frac{3}{5}(x - 11).$$

3. To find an equation for the linear function in slope-intercept form, solve the equation in part b. for $f(x)$. When we do this, we get the equation

$$f(x) = -\frac{3}{5}x + \frac{13}{5}.$$

Exercise 1.2.1

Consider the line passing through points $(-3, 2)$ and $(1, 4)$.

a. Find the slope of the line.
b. Find an equation of that line in point-slope form.
c. Find an equation of that line in slope-intercept form.

Hint

The slope $m = \Delta y/\Delta x$.

Answer a

$m = 1/2$.

Answer b

The point-slope form is $y - 4 = \frac{1}{2}(x - 1)$.

Answer c

The slope-intercept form is $y = \frac{1}{2}x + \frac{7}{2}$.

Example 1.2.2:

Jessica leaves her house at 5:50 a.m. and goes for a 9-mile run. She returns to her house at 7:08 a.m. Answer the following questions, assuming Jessica runs at a constant pace.

a. Describe the distance D (in miles) Jessica runs as a linear function of her run time t (in minutes).
b. Sketch a graph of D.
c. Interpret the meaning of the slope.

Solution

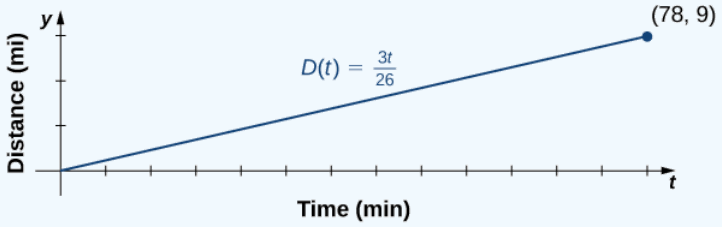
a. At time $t = 0$, Jessica is at her house, so $D(0) = 0$. At time $t = 78$ minutes, Jessica has finished running 9 mi, so $D(78) = 9$. The slope of the linear function is

$$m = \frac{9 - 0}{78 - 0} = \frac{3}{26}.$$

The y-intercept is $(0, 0)$, so the equation for this linear function is

$$D(t) = \frac{3}{26}t.$$

b. To graph D, use the fact that the graph passes through the origin and has slope $m = 3/26$.



c. The slope $m = 3/26 \approx 0.115$ describes the distance (in miles) Jessica runs per minute, or her average velocity.

Piecewise-Defined Functions

Sometimes a function is defined by different formulas on different parts of its domain. A function with this property is known as a piecewise-defined function. The absolute value function is an example of a piecewise-defined function because the formula changes with the sign of x:

$$f(x) = \begin{cases} -x, & \text{if } x < 0 \\ x, & \text{if } x \geq 0 \end{cases}.$$

Other piecewise-defined functions may be represented by completely different formulas, depending on the part of the domain in which a point falls. To graph a piecewise-defined function, we graph each part of the function in its respective domain, on the same coordinate system. If the formula for a function is different for $x < a$ and $x > a$, we need to pay special attention to what happens at $x = a$ when we graph the function. Sometimes the graph needs to include an open or closed circle to indicate the value of the function at $x = a$. We examine this in the next example.

Example 1.2.8: Graphing a Piecewise-Defined Function

Sketch a graph of the following piecewise-defined function:

$$f(x) = \begin{cases} x+3, & \text{if } x < 1 \\ (x-2)^2, & \text{if } x \geq 1 \end{cases}$$

Solution

Graph the linear function $y = x + 3$ on the interval $(-\infty, 1)$ and graph the quadratic function $y = (x-2)^2$ on the interval $[1, \infty)$. Since the value of the function at $x = 1$ is given by the formula $f(x) = (x-2)^2$, we see that $f(1) = 1$. To indicate this on the graph, we draw a closed circle at the point $(1, 1)$. The value of the function is given by $f(x) = x + 3$ for all $x < 1$, but not at $x = 1$. To indicate this on the graph, we draw an open circle at $(1, 4)$.

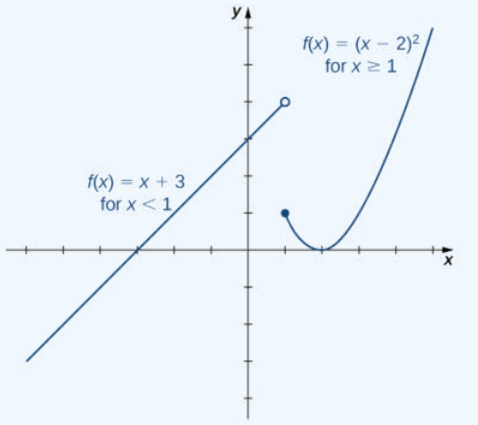
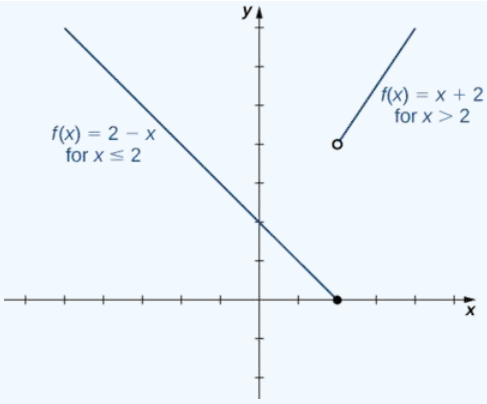


Figure 1.2.8: This piecewise-defined function is linear for $x < 1$ and quadratic for $x \geq 1$.

2) Sketch a graph of the function

$$f(x) = \begin{cases} 2-x, & \text{if } x \leq 2 \\ x+2, & \text{if } x > 2 \end{cases}.$$

Solution:



Example 1.2.9: Parking Fees Described by a Piecewise-Defined Function

In a big city, drivers are charged variable rates for parking in a parking garage. They are charged \$10 for the first hour or any part of the first hour and an additional \$2 for each hour or part thereof up to a maximum of \$30 for the day. The parking garage is open from 6 a.m. to 12 midnight.

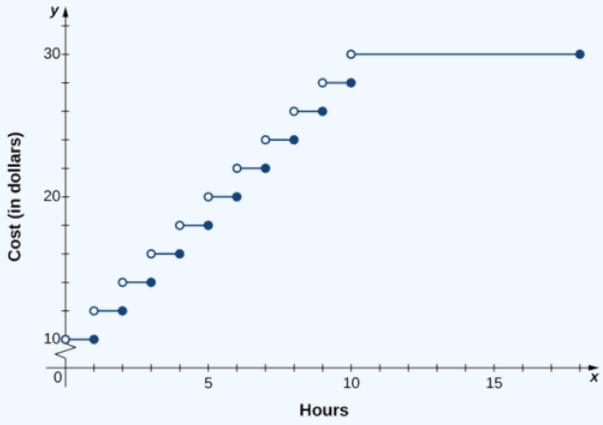
a. Write a piecewise-defined function that describes the cost C to park in the parking garage as a function of hours parked x.
b. Sketch a graph of this function $C(x)$.

Solution

1.Since the parking garage is open 18 hours each day, the domain for this function is $\{x \mid 0 < x \leq 18\}$. The cost to park a car at this parking garage can be described piecewise by the function

$$C(x) = \begin{cases} 10, & \text{for } 0 < x \leq 1 \\ 12, & \text{for } 1 < x \leq 2 \\ 14, & \text{for } 2 < x \leq 3 \\ 16, & \text{for } 3 < x \leq 4 \\ \vdots & \\ 30, & \text{for } 10 < x \leq 18 \end{cases}.$$

2.The graph of the function consists of several horizontal line segments.



Exercise 1.2.6

The cost of mailing a letter is a function of the weight of the letter. Suppose the cost of mailing a letter is $49¢$ for the first ounce and $21¢$ for each additional ounce. Write a piecewise-defined function describing the cost C as a function of the weight x for $0 < x \leq 3$, where C is measured in cents and x is measured in ounces.

Hint

The piecewise-defined function is constant on the intervals $(0,1]$, $(1,2]$, ...

Answer

$$C(x) = \begin{cases} 49, 0 < x \leq 1 \\ 70, 1 < x \leq 2 \\ 91, 2 < x \leq 3 \end{cases}$$

Transformations of Functions

We have seen several cases in which we have added, subtracted, or multiplied constants to form variations of simple functions. In the previous example, for instance, we subtracted 2 from the argument of the function $y = x^2$ to get the function $f(x) = (x-2)^2$. This subtraction represents a shift of the function $y = x^2$ two units to the right. A shift, horizontally or vertically, is a type of **transformation of a function**. Other transformations include horizontal and vertical scalings, and reflections about the axes.

A vertical shift of a function occurs if we add or subtract the same constant to each output y. For $c > 0$, the graph of $f(x) + c$ is a shift of the graph of $f(x)$ up c units, whereas the graph of $f(x) - c$ is a shift of the graph of $f(x)$ down c units. For example, the graph of the function $f(x) = x^3 + 4$ is the graph of $y = x^3$ shifted up 4 units; the graph of the function $f(x) = x^3 - 4$ is the graph of $y = x^3$ shifted down 4 units (Figure 1.2.9).

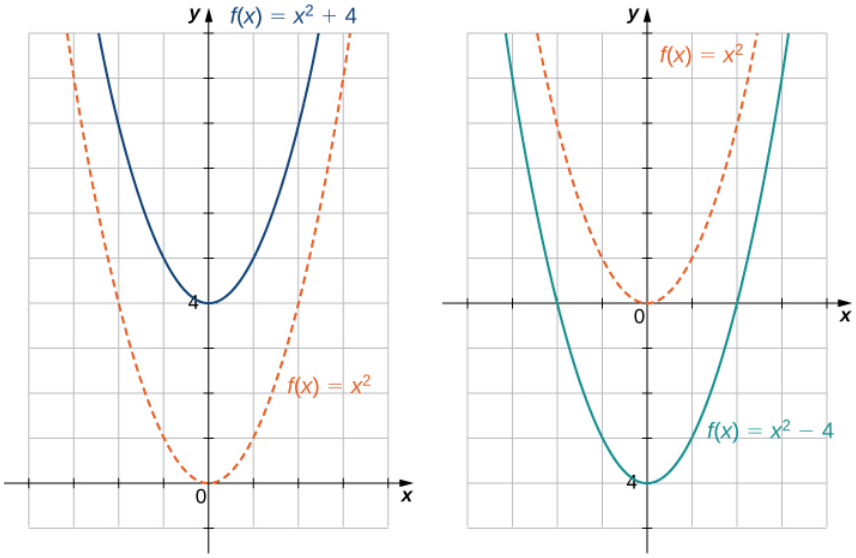


Figure 1.2.9: (a) For $c > 0$, the graph of $y = f(x) + c$ is a vertical shift up c units of the graph of $y = f(x)$. (b) For $c > 0$, the graph of $y = f(x) - c$ is a vertical shift down c units of the graph of $y = f(x)$.

A horizontal shift of a function occurs if we add or subtract the same constant to each input x. For $c > 0$, the graph of $f(x + c)$ is a shift of the graph of $f(x)$ to the left c units; the graph of $f(x - c)$ is a shift of the graph of $f(x)$ to the right c units. Why does the graph shift left when adding a constant and shift right when subtracting a constant? To answer this question, let's look at an example.

Consider the function $f(x) = |x + 3|$ and evaluate this function at $x - 3$. Since $f(x - 3) = |x|$ and $x - 3 < x$, the graph of $f(x) = |x + 3|$ is the graph of $y = |x|$ shifted left 3 units. Similarly, the graph of $f(x) = |x - 3|$ is the graph of $y = |x|$ shifted right 3 units (Figure 1.2.10).

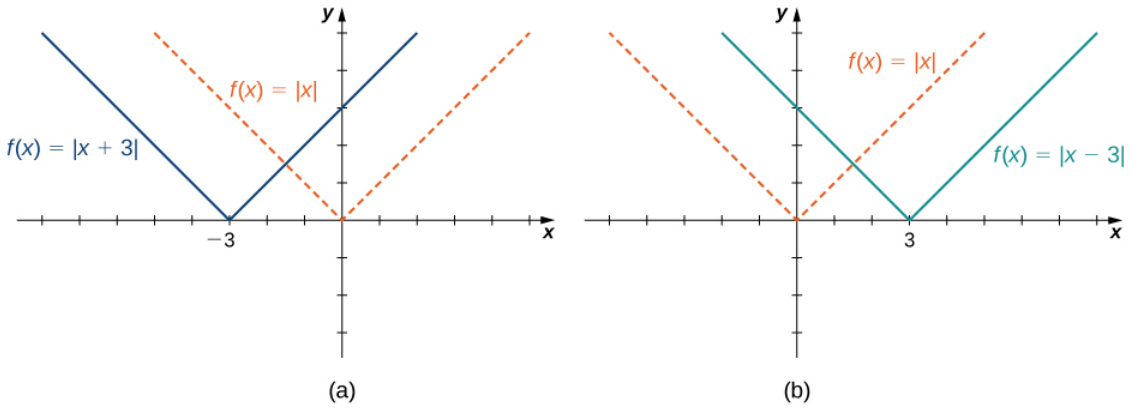


Figure 1.2.10: (a) For $c > 0$, the graph of $y = f(x + c)$ is a horizontal shift left c units of the graph of $y = f(x)$. (b) For $c > 0$, the graph of $y = f(x - c)$ is a horizontal shift right c units of the graph of $y = f(x)$.

A vertical scaling of a graph occurs if we multiply all outputs y of a function by the same positive constant. For $c > 0$, the graph of the function $cf(x)$ is the graph of $f(x)$ scaled vertically by a factor of c. If $c > 1$, the values of the outputs for the function $cf(x)$ are larger than the values of the outputs for the function $f(x)$; therefore, the graph has been stretched vertically. If $0 < c < 1$, then the outputs of the function $cf(x)$ are smaller, so the graph has been compressed. For example, the graph of the function $f(x) = 3x^2$ is the graph of $y = x^2$ stretched vertically by a factor of 3, whereas the graph of $f(x) = x^2/3$ is the graph of $y = x^2$ compressed vertically by a factor of 3 (Figure 1.2.11b).

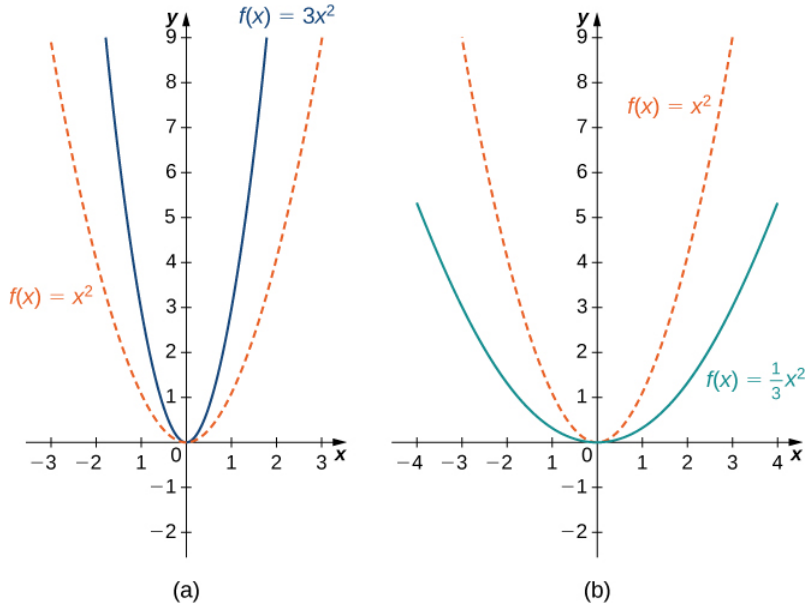


Figure 1.2.11: (a) If $c > 1$, the graph of $y = cf(x)$ is a vertical stretch of the graph of $y = f(x)$. (b) If $0 < c < 1$, the graph of $y = cf(x)$ is a vertical compression of the graph of $y = f(x)$.

The horizontal scaling of a function occurs if we multiply the inputs x by the same positive constant. For $c > 0$, the graph of the function $f(cx)$ is the graph of $f(x)$ scaled horizontally by a factor of c. If $c > 1$, the graph of $f(cx)$ is the graph of $f(x)$ compressed horizontally. If $0 < c < 1$, the graph of $f(cx)$ is the graph of $f(x)$ stretched horizontally. For example, consider the function $f(x) = \sqrt{2x}$ and evaluate f at $x/2$. Since $f(x/2) = \sqrt{x}$, the graph of $f(x) = \sqrt{2x}$ is the graph of $y = \sqrt{x}$ compressed horizontally. The graph of $y = \sqrt{x/2}$ is a horizontal stretch of the graph of $y = \sqrt{x}$ (Figure 1.2.12).

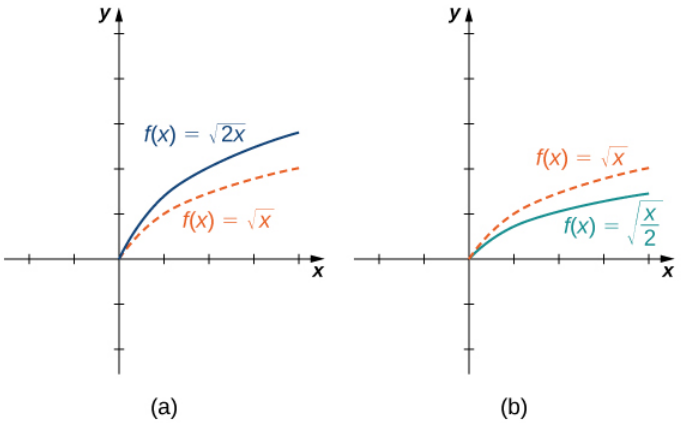


Figure 1.2.12: (a) If $c > 1$, the graph of $y = f(cx)$ is a horizontal compression of the graph of $y = f(x)$. (b) If $0 < c < 1$, the graph of $y = f(cx)$ is a horizontal stretch of the graph of $y = f(x)$.

We have explored what happens to the graph of a function f when we multiply f by a constant $c > 0$ to get a new function $cf(x)$. We have also discussed what happens to the graph of a function $f$when we multiply the independent variable x by $c > 0$ to get a new function $f(cx)$. However, we have not addressed what happens to the graph of the function if the constant c is negative. If we have a constant $c < 0$, we can write c as a positive number multiplied by -1; but, what kind of transformation do we get when we multiply the function or its argument by -1? When we multiply all the outputs by -1, we get a reflection about the x-axis. When we multiply all inputs by -1, we get a reflection about the y-axis. For example, the graph of $f(x) = -(x^3 + 1)$ is the graph of $y = (x^3 + 1)$ reflected about the x-axis. The graph of $f(x) = (-x)^3 + 1$ is the graph of $y = x^3 + 1$ reflected about the y-axis (Figure 1.2.13).

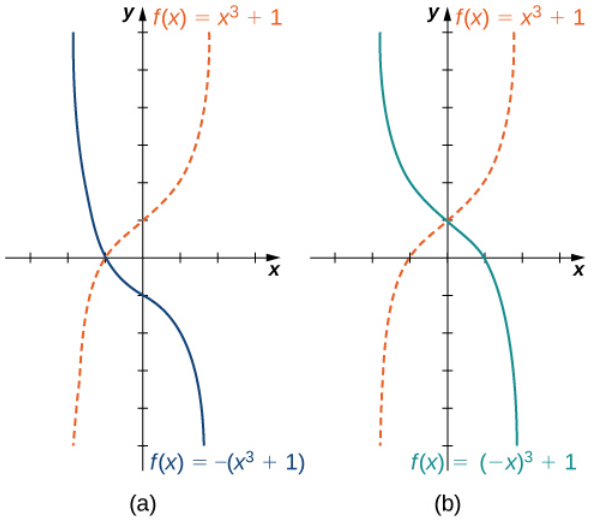


Figure 1.2.13: (a) The graph of $y = -f(x)$ is the graph of $y = f(x)$ reflected about the x-axis. (b) The graph of $y = f(-x)$ is the graph of $y = f(x)$ reflected about the y-axis.

If the graph of a function consists of more than one transformation of another graph, it is important to transform the graph in the correct order. Given a function $f(x)$, the graph of the related function $y = cf(a(x + b)) + d$ can be obtained from the graph of $y = f(x)$by performing the transformations in the following order.

- Horizontal shift of the graph of $y = f(x)$. If $b > 0$, shift left. If $b < 0$ shift right.
- Horizontal scaling of the graph of $y = f(x + b)$ by a factor of $|a|$. If $a < 0$, reflect the graph about the y-axis.
- Vertical scaling of the graph of $y = f(a(x + b))$ by a factor of $|c|$. If $c < 0$, reflect the graph about the x -axis.
- Vertical shift of the graph of $y = cf(a(x + b))$. If $d > 0$, shift up. If $d < 0$, shift down.

We can summarize the different transformations and their related effects on the graph of a function in the following table.

Transformation of $f(c > 0)$	**Effect of the graph of f**
$f(x) + c$	Vertical shift up c units

Transformation of $f(c>0)$	Effect of the graph of f
$f(x)-c$	Vertical shift down c units
$f(x+c)$	Shift left by c units
$f(x-c)$	Shift right by c units
$cf(x)$	Vertical stretch if $c>1$; vertical compression if $0<c<1$
$f(cx)$	Horizontal stretch if $0<c<1$; horizontal compression if $c>1$
$-f(x)$	Reflection about the x-axis
$f(-x)$	Reflection about the y-axis

Example 1.2.10: Transforming a Function

For each of the following functions, a. and b., sketch a graph by using a sequence of transformations of a well-known function.

a. $f(x)=-|x+2|-3$
b. $f(x)=3\sqrt{-x}+1$

Solution

1.Starting with the graph of $y=|x|$, shift 2 units to the left, reflect about the x-axis, and then shift down 3 units.

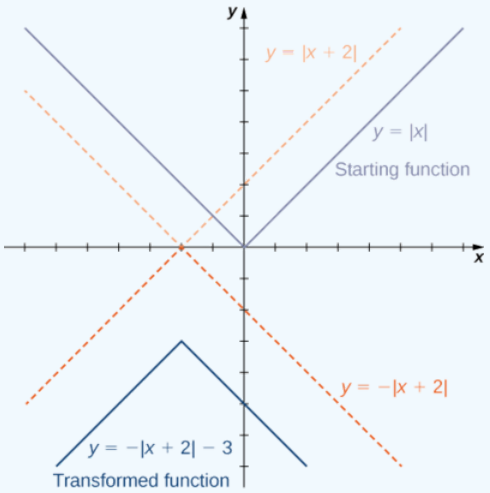


Figure 1.2.14: The function $f(x)=-|x+2|-3$ can be viewed as a sequence of three transformations of the function $y=|x|$.

2. Starting with the graph of $y=\sqrt{x}$, reflect about the y-axis, stretch the graph vertically by a factor of 3, and move up 1 unit.

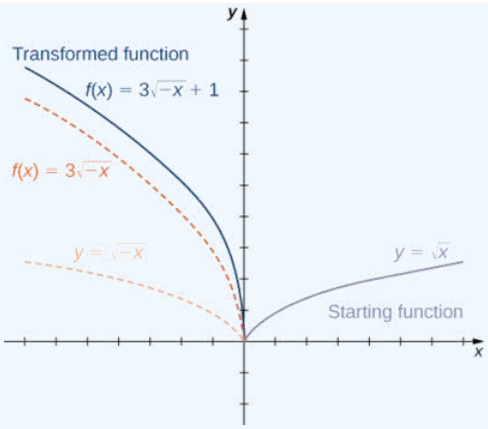


Figure 1.2.15: The function $f(x) = 3\sqrt{-x} + 1$ can be viewed as a sequence of three transformations of the function $y = \sqrt{x}$.

? Exercise 1.2.7

Describe how the function $f(x) = -(x+1)^2 - 4$ can be graphed using the graph of $y = x^2$ and a sequence of transformations

Answer

Shift the graph $y = x^2$ to the left 1 unit, reflect about the x-axis, then shift down 4 units.

Analyzing Parabolas

The graph of a quadratic function is a U-shaped curve called a parabola. One important feature of the graph is that it has an extreme point, called the **vertex**. If the parabola opens up, the vertex represents the lowest point on the graph, or the **minimum value** of the quadratic function. If the parabola opens down, the vertex represents the highest point on the graph, or the **maximum value**. In either case, the vertex is a turning point on the graph. The graph is also symmetric with a vertical line drawn through the vertex, called the **axis of symmetry**. These features are illustrated in Figure 1.2.2.

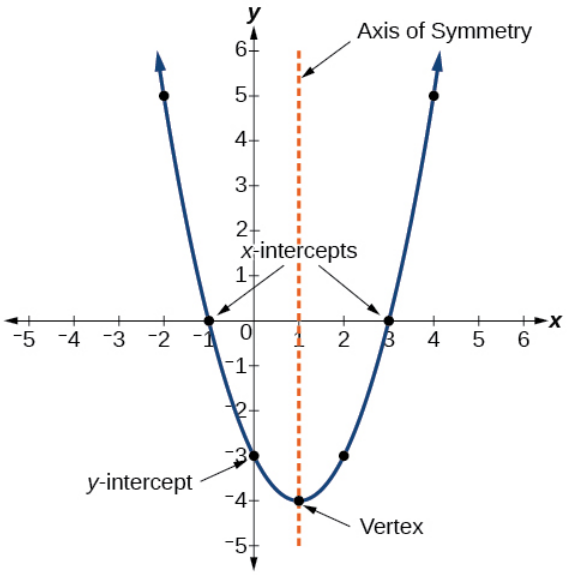


Figure 1.2.2: Graph of a parabola showing where the x and y intercepts, vertex, and axis of symmetry are.

The y-intercept is the point at which the parabola crosses the y-axis. The x-intercepts are the points at which the parabola crosses the x-axis. If they exist, the x-intercepts represent the **zeros**, or **roots**, of the quadratic function, the values of x at which $y = 0$.

Definitions: Forms of Quadratic Functions

A quadratic function is a function of degree two. The graph of a *quadratic function* is a parabola.

- The *general form* of a quadratic function is $f(x)=ax^2+bx+c$ where a, b, and c are real numbers and $a\neq 0$.
- The standard form of a quadratic function is $f(x)=a(x-h)^2+k$.
- The vertex (h,k) is located at

$$h=-\frac{b}{2a},\ k=f(h)=f(\frac{-b}{2a}). \tag{1.2.1}$$

How To: Given a quadratic function in general form, find the vertex of the parabola.

1. Identify a, b, and c.
2. Find h, the x-coordinate of the vertex, by substituting a and b into $h=-\frac{b}{2a}$.
3. Find k, the y-coordinate of the vertex, by evaluating $k=f(h)=f\left(-\frac{b}{2a}\right)$.

Example 1.2.8: Finding the Vertex of a Quadratic Function

Find the vertex of the quadratic function $f(x)=2x^2-6x+7$. Rewrite the quadratic in standard form (vertex form).

Solution

The horizontal coordinate of the vertex will be at

$$h=-\frac{b}{2a} \tag{1.2.2}$$
$$=-\frac{-6}{2(2)} \tag{1.2.3}$$
$$=\frac{6}{4} \tag{1.2.4}$$
$$=\frac{3}{2} \tag{1.2.5}$$

The vertical coordinate of the vertex will be at

$$k=f(h) \tag{1.2.6}$$
$$=f\left(\frac{3}{2}\right) \tag{1.2.7}$$
$$=2\left(\frac{3}{2}\right)^2-6\left(\frac{3}{2}\right)+7 \tag{1.2.8}$$
$$=\frac{5}{2} \tag{1.2.9}$$

Rewriting into standard form, the stretch factor will be the same as the a in the original quadratic.

$$f(x)=ax^2+bx+c \tag{1.2.10}$$
$$f(x)=2x^2-6x+7$$

Using the vertex to determine the shifts,

$$f(x)=2\left(x-\frac{3}{2}\right)^2+\frac{5}{2} \tag{1.2.11}$$

Analysis

One reason we may want to identify the vertex of the parabola is that this point will inform us what the maximum or minimum value of the function is, (k),and where it occurs, (h).

Exercise 1.2.8

Given the equation $g(x) = 13 + x^2 - 6x$, write the equation in general form and then in standard form.

Answer

$g(x) = x^2 - 6x + 13$ in general form; $g(x) = (x-3)^2 + 4$ in standard form.

Finding the Domain and Range of a Quadratic Function

Any number can be the input value of a quadratic function. Therefore, the domain of any quadratic function is all real numbers. Because parabolas have a maximum or a minimum point, the range is restricted. Since the vertex of a parabola will be either a maximum or a minimum, the range will consist of all y-values greater than or equal to the y-coordinate at the turning point or less than or equal to the y-coordinate at the turning point, depending on whether the parabola opens up or down.

Definition: Domain and Range of a Quadratic Function

The domain of any **quadratic function** is all real numbers.

The range of a quadratic function written in general form $f(x) = ax^2 + bx + c$ with a positive a value is $f(x) \geq f\left(-\frac{b}{2a}\right)$, or $[f(-\frac{b}{2a}), \infty)$; the range of a quadratic function written in general form with a negative a value is $f(x) \leq f(-\frac{b}{2a})$, or $(-\infty, f(-\frac{b}{2a})]$.

The range of a quadratic function written in standard form $f(x) = a(x-h)^2 + k$ with a positive a value is $f(x) \geq k$; the range of a quadratic function written in standard form with a negative a value is $f(x) \leq k$.

How To: Given a quadratic function, find the domain and range.

1. Identify the domain of any quadratic function as all real numbers.
2. Determine whether a is positive or negative. If a is positive, the parabola has a minimum. If a is negative, the parabola has a maximum.
3. Determine the maximum or minimum value of the parabola, k.
4. If the parabola has a minimum, the range is given by $f(x) \geq k$, or $[k, \infty)$. If the parabola has a maximum, the range is given by $f(x) \leq k$, or $(-\infty, k]$.

Example 1.2.9: Finding the Domain and Range of a Quadratic Function

Find the domain and range of $f(x) = -5x^2 + 9x - 1$.

Solution

As with any quadratic function, the domain is all real numbers.

Because a is negative, the parabola opens downward and has a maximum value. We need to determine the maximum value. We can begin by finding the x-value of the vertex.

$$h = -\frac{b}{2a} \tag{1.2.12}$$

$$= -\frac{9}{2(-5)} \tag{1.2.13}$$

$$= \frac{9}{10} \tag{1.2.14}$$

The maximum value is given by $f(h)$.

$$f(\frac{9}{10}) = 5(\frac{9}{10})^2 + 9(\frac{9}{10}) - 1 \tag{1.2.15}$$

$$= \frac{61}{20} \tag{1.2.16}$$

The range is $f(x) \leq \frac{61}{20}$, or $\left(-\infty, \frac{61}{20}\right]$.

Exercise 1.2.9

Find the domain and range of $f(x) = 2\left(x - \frac{4}{7}\right)^2 + \frac{8}{11}$.

Answer

The domain is all real numbers. The range is $f(x) \geq \frac{8}{11}$, or $\left[\frac{8}{11}, \infty\right)$.

Finding the x- and y-Intercepts of a Quadratic Function

Much as we did in the application problems above, we also need to find intercepts of quadratic equations for graphing parabolas. Recall that we find the y-intercept of a quadratic by evaluating the function at an input of zero, and we find the x-intercepts at locations where the output is zero. Notice in Figure 1.2.13 that the number of x-intercepts can vary depending upon the location of the graph.

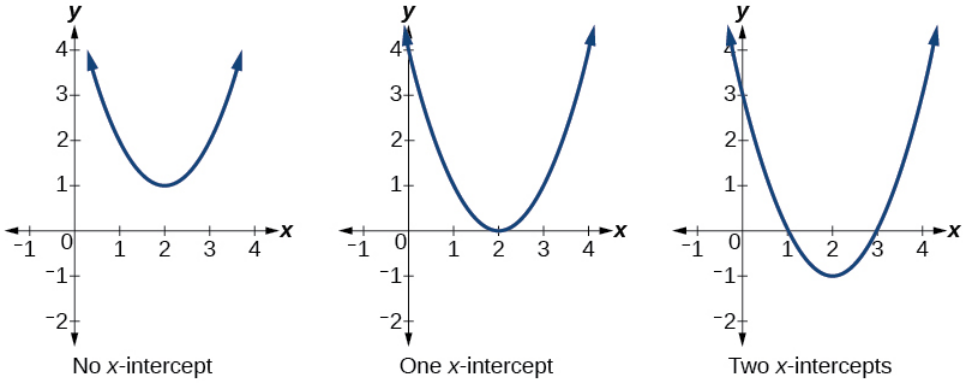


Figure 1.2.13: Number of x-intercepts of a parabola.

How To: Given a quadratic function $f(x)$, find the y- and x-intercepts.

1. Evaluate $f(0)$ to find the y-intercept.
2. Solve the quadratic equation $f(x) = 0$ to find the x-intercepts.

Example 1.2.10: Finding the y- and x-Intercepts of a Parabola

Find the y- and x-intercepts of the quadratic $f(x) = 3x^2 + 5x - 2$.

Solution

We find the y-intercept by evaluating $f(0)$.

$$f(0) = 3(0)^2 + 5(0) - 2 \tag{1.2.17}$$
$$= -2 \tag{1.2.18}$$

So the y-intercept is at $(0, -2)$.

For the x-intercepts, we find all solutions of $f(x) = 0$.

$$0 = 3x^2 + 5x - 2 \tag{1.2.19}$$

In this case, the quadratic can be factored easily, providing the simplest method for solution.

$$0 = (3x - 1)(x + 2) \tag{1.2.20}$$
$$0 = 3x - 1 \qquad 0 = x + 2 \tag{1.2.21}$$
$$x = \frac{1}{3} \quad \text{or} \quad x = -2 \tag{1.2.22}$$

So the x-intercepts are at $(\frac{1}{3}, 0)$ and $(-2, 0)$.

Analysis

By graphing the function, we can confirm that the graph crosses the y-axis at $(0, -2)$. We can also confirm that the graph crosses the x-axis at $\left(\frac{1}{3}, 0\right)$ and $(-2, 0)$. See Figure 1.2.14

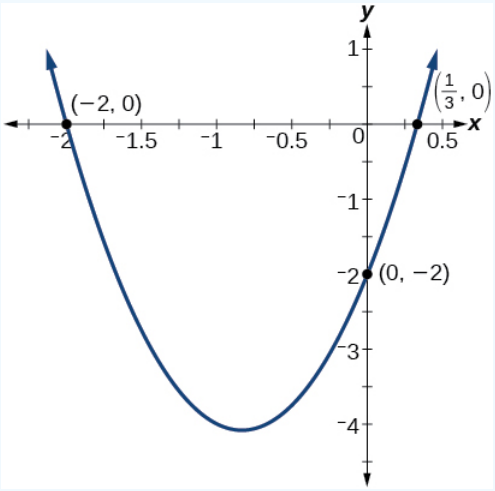


Figure 1.2.14: Graph of a parabola.

Rewriting Quadratics in Standard Form

In Example 1.2.7, the quadratic was easily solved by factoring. However, there are many quadratics that cannot be factored. We can solve these quadratics by first rewriting them in standard form.

How To: Given a quadratic function, find the x-intercepts by rewriting in standard form.

1. Substitute a and b into $h = -\frac{b}{2a}$.
2. Substitute $x = h$ into the general form of the quadratic function to find k.
3. Rewrite the quadratic in standard form using h and k.
4. Solve for when the output of the function will be zero to find the x-intercepts.

Example 1.2.11: Finding the x-Intercepts of a Parabola

Find the x-intercepts of the quadratic function $f(x) = 2x^2 + 4x - 4$.

Solution

We begin by solving for when the output will be zero.

$$0 = 2x^2 + 4x - 4$$

Because the quadratic is not easily factorable in this case, we solve for the intercepts by first rewriting the quadratic in standard form.

$$f(x) = a(x - h)^2 + k$$

We know that $a = 2$. Then we solve for h and k.

$$\begin{aligned} h &= -\frac{b}{2a} & k &= f(-1) \\ &= -\frac{4}{2(2)} & &= 2(-1)^2 + 4(-1) - 4 \\ &= -1 & &= -6 \end{aligned}$$

So now we can rewrite in standard form.

$$f(x) = 2(x+1)^2 - 6$$

We can now solve for when the output will be zero.

$$\begin{aligned} 0 &= 2(x+1)^2 - 6 \\ 6 &= 2(x+1)^2 \\ 3 &= (x+1)^2 \\ x+1 &= \pm\sqrt{3} \\ x &= -1 \pm \sqrt{3} \end{aligned}$$

The graph has x-intercepts at $(-1-\sqrt{3}, 0)$ and $(-1+\sqrt{3}, 0)$.

Analysis

We can check our work by graphing the given function on a graphing utility and observing the x-intercepts. See Figure 1.2.15.

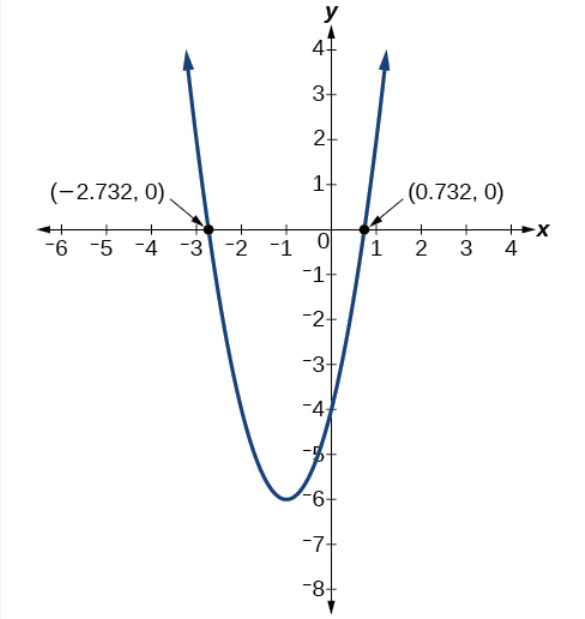


Figure 1.2.15: Graph of a parabola which has the following x-intercepts: $(-2.732, 0)$ and $(0.732, 0)$.

Exercise 1.2.11

In Try It 1.2.1, we found the standard and general form for the function $g(x) = 13 + x^2 - 6x$. Now find the y- and x-intercepts (if any).

Answer

y-intercept at $(0, 13)$, No x-intercepts

Example 1.2.12: Solving a Quadratic Equation with the Quadratic Formula

Solve $x^2 + x + 2 = 0$.

Solution

Let's begin by writing the quadratic formula: $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

When applying the quadratic formula, we identify the coefficients a, b and c. For the equation $x^2 + x + 2 = 0$, we have $a = 1$, $b = 1$, and $c = 2$. Substituting these values into the formula we have:

$$
\begin{aligned}
x &= \frac{-b \pm \sqrt{b^2-4ac}}{2a} \\
&= \frac{-1 \pm \sqrt{1^2-4 \cdot 1 \cdot (2)}}{2 \cdot 1} \\
&= \frac{-1 \pm \sqrt{1-8}}{2} \\
&= \frac{-1 \pm \sqrt{-7}}{2} \\
&= \frac{-1 \pm i\sqrt{7}}{2}
\end{aligned}
$$

The solutions to the equation are $x = \frac{-1+i\sqrt{7}}{2}$ and $x = \frac{-1-i\sqrt{7}}{2}$ or $x = -\frac{1}{2} + \frac{i\sqrt{7}}{2}$ and $x = \frac{-1}{2} - \frac{i\sqrt{7}}{2}$.

Example 1.2.13: Applying the Vertex and x-Intercepts of a Parabola

A ball is thrown upward from the top of a 40 foot high building at a speed of 80 feet per second. The ball's height above ground can be modeled by the equation $H(t) = -16t^2 + 80t + 40$.

When does the ball reach the maximum height?
What is the maximum height of the ball?
When does the ball hit the ground?

The ball reaches the maximum height at the vertex of the parabola.

$$
\begin{aligned}
h &= -\frac{80}{2(-16)} && (1.2.23)\\
&= \frac{80}{32} && (1.2.24)\\
&= \frac{5}{2} && (1.2.25)\\
&= 2.5 && (1.2.26)
\end{aligned}
$$

The ball reaches a maximum height after 2.5 seconds.

To find the maximum height, find the y-coordinate of the vertex of the parabola.

$$
\begin{aligned}
k &= H(-\frac{b}{2a}) && (1.2.27)\\
&= H(2.5) && (1.2.28)\\
&= -16(2.5)^2 + 80(2.5) + 40 && (1.2.29)\\
&= 140 && (1.2.30)
\end{aligned}
$$

The ball reaches a maximum height of 140 feet.

To find when the ball hits the ground, we need to determine when the height is zero, $H(t) = 0$.

We use the quadratic formula.

$$
\begin{aligned}
t &= \frac{-80 \pm \sqrt{80^2 - 4(-16)(40)}}{2(-16)} && (1.2.31)\\
&= \frac{-80 \pm \sqrt{8960}}{-32} && (1.2.32)
\end{aligned}
$$

Because the square root does not simplify nicely, we can use a calculator to approximate the values of the solutions.

$$
t = \frac{-80 - \sqrt{8960}}{-32} \approx 5.458 \text{ or } t = \frac{-80 + \sqrt{8960}}{-32} \approx -0.458 \quad (1.2.33)
$$

The second answer is outside the reasonable domain of our model, so we conclude the ball will hit the ground after about 5.458 seconds. See Figure 1.2.16.

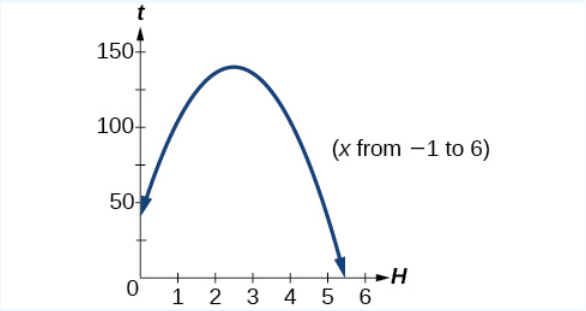


Figure 1.2.16

1.2.13: A rock is thrown upward from the top of a 112-foot high cliff overlooking the ocean at a speed of 96 feet per second. The rock's height above ocean can be modeled by the equation $H(t) = -16t^2 + 96t + 112$.

a. When does the rock reach the maximum height?
b. What is the maximum height of the rock?
c. When does the rock hit the ocean?

Solution

a. 3 seconds b. 256 feet c. 7 seconds

This page titled 1.2: Basic Classes of Functions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin "Jed" Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **1.2: Basic Classes of Functions** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **3.2: Quadratic Functions** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

1.3: Trigonometric Functions

Learning Objectives

- Convert angle measures between degrees and radians.
- Recognize the triangular and circular definitions of the basic trigonometric functions.
- Verify the fundamental trigonometric identities.
- Solve trigonometric equations.

Trigonometric functions are used to model many phenomena, including sound waves, vibrations of strings, alternating electrical current, and the motion of pendulums. In fact, almost any repetitive, or cyclical, motion can be modeled by some combination of trigonometric functions. In this section, we define the six basic trigonometric functions and look at some of the main identities involving these functions.

Radian Measure

To use trigonometric functions, we first must understand how to measure the angles. Although we can use both radians and degrees, **radians** are a more natural measurement because they are related directly to the unit circle, a circle with radius 1. The radian measure of an angle θ, let s be the length of the corresponding arc on the unit circle (Figure $1.3.1$). We say the angle corresponding to the arc of length 1 has radian measure 1.

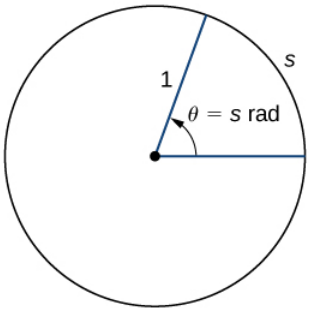


Figure $1.3.1$: The radian measure of an angle θ is the arc length s of the associated arc on the unit circle.

Since an angle of 360° corresponds to the circumference of a circle, or an arc of length 2π, we conclude that an angle with a degree measure of 360° has a radian measure of 2π. Similarly, we see that 180° is equivalent to π radians. Table $1.3.1$ shows the relationship between common degree and radian values.

Table $1.3.1$: Common Angles Expressed in Degrees and Radians

Degrees	Radians	Degrees	Radians
0	0	120	$2\pi/3$
30	$\pi/6$	135	$3\pi/4$
45	$\pi/4$	150	$5\pi/6$
60	$\pi/3$	180	π
90	$\pi/2$		

Converting between Radians and Degrees

a. Express 225° using radians.
b. Express $5\pi/3$ rad using degrees.

Solution

Use the fact that $180°$ is equivalent to π radians as a **conversion factor** (Table $1.3.1$):

$$1 = \frac{\pi \text{ rad}}{180^\circ} = \frac{180^\circ}{\pi \text{ rad}}.$$

a. $225^\circ = 225^\circ \cdot \left(\frac{\pi}{180^\circ}\right) = \left(\frac{5\pi}{4}\right)$ rad

b. $\frac{5\pi}{3}$ rad $= \frac{5\pi}{3} \cdot \frac{180^\circ}{\pi} = 300°$

Exercise 1.3.1

a. Express 210° using radians.

b. Express $11\pi/6$ rad using degrees.

Hint

π radians is equal to 180°

Answer

a. $7\pi/6$

b. 330°

The Six Basic Trigonometric Functions

Trigonometric functions allow us to use angle measures, in radians or degrees, to find the coordinates of a point on any circle—not only on a unit circle—or to find an angle given a point on a circle. They also define the relationship between the sides and angles of a triangle.

To define the trigonometric functions, first consider the unit circle centered at the origin and a point $P = (x, y)$ on the unit circle. Let θ be an angle with an initial side that lies along the positive x-axis and with a terminal side that is the line segment OP. An angle in this position is said to be in standard position (Figure 1.3.2). We can then define the values of the six trigonometric functions for θ in terms of the coordinates x and y.

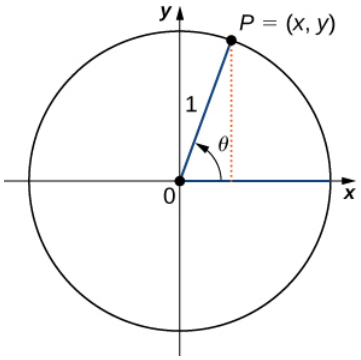


Figure 1.3.2: The angle θ is in standard position. The values of the trigonometric functions for θ are defined in terms of the coordinates x and y.

Definition: Trigonometric functions

Let $P = (x, y)$ be a point on the unit circle centered at the origin O. Let θ be an angle with an initial side along the positive x-axis and a terminal side given by the line segment OP. The trigonometric functions are then defined as

$\sin\theta = y$	$\csc\theta = \frac{1}{y}$
$\cos\theta = x$	$\sec\theta = \frac{1}{x}$
$\tan\theta = \frac{y}{x}$	$\cot\theta = \frac{x}{y}$

If $x = 0$, $\sec\theta$ and $\tan\theta$ are undefined. If $y = 0$, then $\cot\theta$ and $\csc\theta$ are undefined.

We can see that for a point $P = (x, y)$ on a circle of radius r with a corresponding angle θ, the coordinates x and y satisfy

$$\cos\theta = \frac{x}{r} \tag{1.3.1}$$

$$x = r\cos\theta \tag{1.3.2}$$

and

$$\sin\theta = \frac{y}{r} \tag{1.3.3}$$

$$y = r\sin\theta. \tag{1.3.4}$$

The values of the other trigonometric functions can be expressed in terms of x, y, and r (Figure 1.3.3).

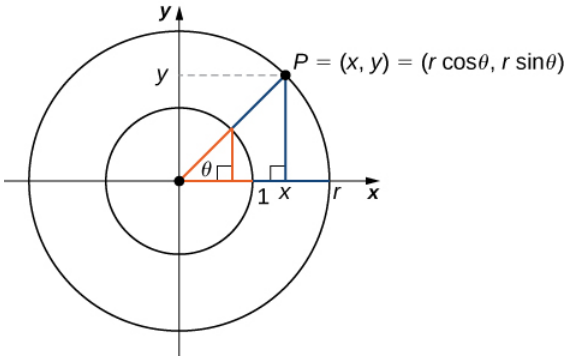


Figure 1.3.3: For a point $P = (x, y)$ on a circle of radius r, the coordinates x and y satisfy $x = r\cos\theta$ and $y = r\sin\theta$.

Table 1.3.2 shows the values of sine and cosine at the major angles in the first quadrant. From this table, we can determine the values of sine and cosine at the corresponding angles in the other quadrants. The values of the other trigonometric functions are calculated easily from the values of $\sin\theta$ and $\cos\theta$.

Table 1.3.2: Values of $\sin\theta$ and $\cos\theta$ at Major Angles θ in the First Quadrant

θ	$\sin\theta$	$\cos\theta$
0	0	1
$\frac{\pi}{6}$	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$
$\frac{\pi}{4}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$
$\frac{\pi}{3}$	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$
$\frac{\pi}{2}$	1	0

Example 1.3.2: Evaluating Trigonometric Functions

Evaluate each of the following expressions.

a. $\sin\left(\frac{2\pi}{3}\right)$

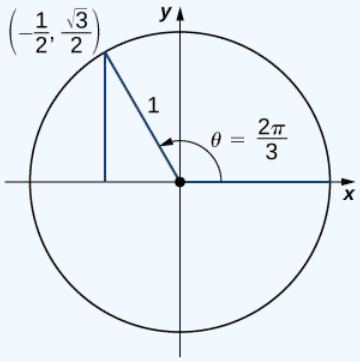
b. $\cos\left(-\frac{5\pi}{6}\right)$

c. $\tan\left(\frac{15\pi}{4}\right)$

Solution:

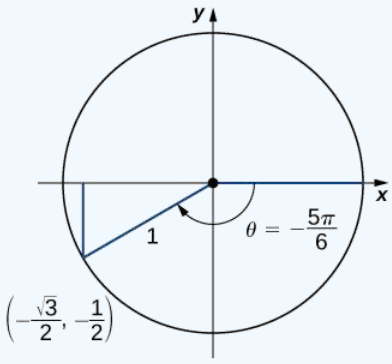
a) On the unit circle, the angle $\theta = \dfrac{2\pi}{3}$ corresponds to the point $\left(-\dfrac{1}{2}, \dfrac{\sqrt{3}}{2}\right)$. Therefore,

$$\sin\left(\frac{2\pi}{3}\right) = y = \left(\frac{\sqrt{3}}{2}\right).$$



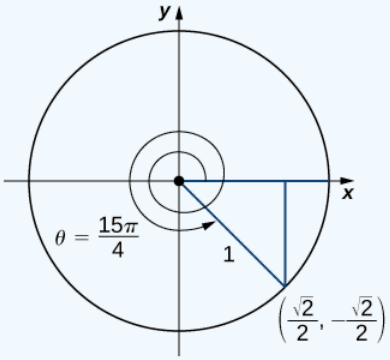
b) An angle $\theta = -\dfrac{5\pi}{6}$ corresponds to a revolution in the negative direction, as shown. Therefore,

$$\cos\left(-\frac{5\pi}{6}\right) = x = -\frac{\sqrt{3}}{2}.$$



c) An angle $\theta = \dfrac{15\pi}{4} = 2\pi + \dfrac{7\pi}{4}$. Therefore, this angle corresponds to more than one revolution, as shown. Knowing the fact that an angle of $\dfrac{7\pi}{4}$ corresponds to the point $\left(\dfrac{\sqrt{2}}{2}, -\dfrac{\sqrt{2}}{2}\right)$, we can conclude that

$$\tan\left(\frac{15\pi}{4}\right) = \frac{y}{x} = -1.$$



Exercise 1.3.2

Evaluate $\cos(3\pi/4)$ and $\sin(-\pi/6)$.

Hint

Look at angles on the unit circle.

Answer

$$\cos(3\pi/4) = -\sqrt{2}/2$$

$$\sin(-\pi/6) = -1/2$$

As mentioned earlier, the ratios of the side lengths of a right triangle can be expressed in terms of the trigonometric functions evaluated at either of the acute angles of the triangle. Let θ be one of the acute angles. Let A be the length of the adjacent leg, O be the length of the opposite leg, and H be the length of the hypotenuse. By inscribing the triangle into a circle of radius H, as shown in Figure 1.3.4, we see that A, H, and O satisfy the following relationships with θ:

$\sin\theta = \dfrac{O}{H}$	$\csc\theta = \dfrac{H}{O}$
$\cos\theta = \dfrac{A}{H}$	$\sec\theta = \dfrac{H}{A}$
$\tan\theta = \dfrac{O}{A}$	$\cot\theta = \dfrac{A}{O}$

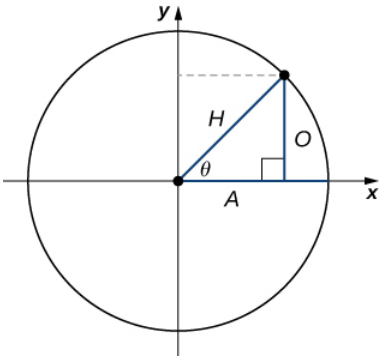


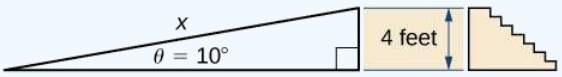
Figure 1.3.4: By inscribing a right triangle in a circle, we can express the ratios of the side lengths in terms of the trigonometric functions evaluated at θ.

Example 1.3.3: Constructing a Wooden Ramp

A wooden ramp is to be built with one end on the ground and the other end at the top of a short staircase. If the top of the staircase is 4 ft from the ground and the angle between the ground and the ramp is to be 10°, how long does the ramp need to be?

Solution

Let x denote the length of the ramp. In the following image, we see that x needs to satisfy the equation $\sin(10^\circ) = 4/x$. Solving this equation for x, we see that $x = 4/\sin(10^\circ) \approx 23.035$ ft.



Exercise 1.3.3

A house painter wants to lean a 20-ft ladder against a house. If the angle between the base of the ladder and the ground is to be 60°, how far from the house should she place the base of the ladder?

Hint

Draw a right triangle with hypotenuse 20.

Answer

10 ft

Verifying the Fundamental Trigonometric Identities

Identities enable us to simplify complicated expressions. They are the basic tools of trigonometry used in solving trigonometric equations, just as factoring, finding common denominators, and using special formulas are the basic tools of solving algebraic equations. In fact, we use algebraic techniques constantly to simplify trigonometric expressions. Basic properties and formulas of algebra, such as the difference of squares formula and the perfect squares formula, will simplify the work involved with trigonometric expressions and equations. We already know that all of the trigonometric functions are related because they all are defined in terms of the unit circle. Consequently, any trigonometric identity can be written in many ways.

To verify the trigonometric identities, we usually start with the more complicated side of the equation and essentially rewrite the expression until it has been transformed into the same expression as the other side of the equation. Sometimes we have to factor expressions, expand expressions, find common denominators, or use other algebraic strategies to obtain the desired result. In this first section, we will work with the fundamental identities: the **Pythagorean identities**, the even-odd identities, the reciprocal identities, and the quotient identities.

We will begin with the **Pythagorean identities** (Table 1.3.1), which are equations involving trigonometric functions based on the properties of a right triangle. We have already seen and used the first of these identifies, but now we will also use additional identities.

Table 1.3.1: Pythagorean Identities

$\sin^2\theta+\cos^2\theta=1$	$1+\cot^2\theta=\csc^2\theta$	$1+\tan^2\theta=\sec^2\theta$

The second and third identities can be obtained by manipulating the first. The identity $1+\cot^2\theta=\csc^2\theta$ is found by rewriting the left side of the equation in terms of sine and cosine.

Prove: $1+\cot^2\theta=\csc^2\theta$

$$\begin{aligned} 1+\cot^2\theta &= (1+\frac{\cos^2}{\sin^2}) && \text{Rewrite the left side} \\ &= \left(\frac{\sin^2}{\sin^2}\right)+\left(\frac{\cos^2}{\sin^2}\right) && \text{Write both terms with the common denominator} \\ &= \frac{\sin^2+\cos^2}{\sin^2} \\ &= \frac{1}{\sin^2} \\ &= \csc^2 \end{aligned}$$

Similarly,$1+\tan^2\theta=\sec^2\theta$ can be obtained by rewriting the left side of this identity in terms of sine and cosine. This gives

$$
\begin{aligned}
1+\tan^2\theta &= 1+\left(\frac{\sin\theta}{\cos\theta}\right)^2 && \text{Rewrite left side}\\
&= \left(\frac{\cos\theta}{\cos\theta}\right)^2+\left(\frac{\sin\theta}{\cos\theta}\right)^2 && \text{Write both terms with the common denominator}\\
&= \frac{\cos^2\theta+\sin^2\theta}{\cos^2\theta}\\
&= \frac{1}{\cos^2\theta}\\
&= \sec^2\theta
\end{aligned}
$$

Recall that we determined which trigonometric functions are odd and which are even. The next set of fundamental identities is the set of **even-odd identities.** The **even-odd identities** relate the value of a trigonometric function at a given angle to the value of the function at the opposite angle (Table 1.3.2).

Table 1.3.2: Even-Odd Identities

$\tan(-\theta)=-\tan\theta$	$\sin(-\theta)=-\sin\theta$	$\cos(-\theta)=\cos\theta$
$\cot(-\theta)=-\cot\theta$	$\csc(-\theta)=-\csc\theta$	$\sec(-\theta)=\sec\theta$

Recall that an **odd function** is one in which $f(-x)=-f(x)$ for all x in the domain off. f. The sine function is an odd function because $\sin(-\theta)=-\sin\theta$. The graph of an odd function is symmetric about the origin. For example, consider corresponding inputs of $\frac{\pi}{2}$ and $-\frac{\pi}{2}$. The output of $\sin\left(\frac{\pi}{2}\right)$ is opposite the output of $\sin\left(-\frac{\pi}{2}\right)$. Thus,

$$
\begin{aligned}
\sin\left(\frac{\pi}{2}\right) &= 1\\
\sin\left(-\frac{\pi}{2}\right) &= -\sin\left(\frac{\pi}{2}\right)\\
&= -1
\end{aligned}
$$

This is shown in Figure 1.3.2.

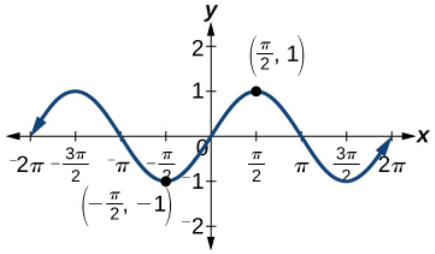


Figure 1.3.2: Graph of $y=\sin\theta$

Recall that an **even function** is one in which

$$f(-x)=f(x) \text{ for all } x \text{ in the domain of } f$$

The graph of an even function is symmetric about the *y*-axis. The cosine function is an even function because $\cos(-\theta)=\cos\theta$. For example, consider corresponding inputs $\frac{\pi}{4}$ and $-\frac{\pi}{4}$. The output of $\cos\left(\frac{\pi}{4}\right)$ is the same as the output of $\cos\left(-\frac{\pi}{4}\right)$. Thus,

$$
\begin{aligned}
\cos\left(-\frac{\pi}{4}\right) &= \cos\left(\frac{\pi}{4}\right)\\
&\approx 0.707
\end{aligned}
$$

See Figure 1.3.3.

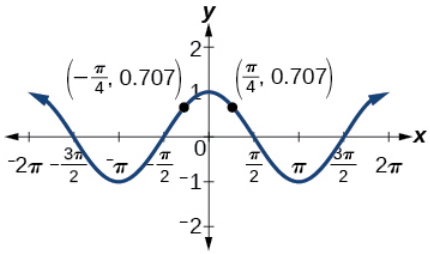


Figure 1.3.3: Graph of $y = \cos\theta$

For all θ in the domain of the sine and cosine functions, respectively, we can state the following:

- Since $\sin(-\theta) = -\sin\theta$,sine is an odd function.
- Since $\cos(-\theta) = \cos\theta$,cosine is an even function.

The other even-odd identities follow from the even and odd nature of the sine and cosine functions. For example, consider the tangent identity,$\tan(-\theta) = -\tan\theta$. We can interpret the tangent of a negative angle as

$$\tan(-\theta) = \frac{\sin(-\theta)}{\cos(-\theta)} = \frac{-\sin\theta}{\cos\theta} = -\tan\theta.$$

Tangent is therefore an odd function, which means that $\tan(-\theta) = -\tan(\theta)$ for all θ in the domain of the **tangent function.**

The cotangent identity, $\cot(-\theta) = -\cot\theta$,also follows from the sine and cosine identities. We can interpret the cotangent of a negative angle as

$$\cot(-\theta) = \frac{\cos(-\theta)}{\sin(-\theta)} = \frac{\cos\theta}{-\sin\theta} = -\cot\theta.$$

Cotangent is therefore an odd function, which means that $\cot(-\theta) = -\cot(\theta)$ for all θ in the domain of the **cotangent function.**

The **cosecant function** is the reciprocal of the sine function, which means that the cosecant of a negative angle will be interpreted as

$$\csc(-\theta) = \frac{1}{\sin(-\theta)} = \frac{1}{-\sin\theta} = -\csc\theta.$$

The cosecant function is therefore odd.

Finally, the secant function is the reciprocal of the cosine function, and the secant of a negative angle is interpreted as

$$\sec(-\theta) = \frac{1}{\cos(-\theta)} = \frac{1}{\cos\theta} = \sec\theta.$$

The secant function is therefore even.

To sum up, only two of the trigonometric functions, cosine and secant, are even. The other four functions are odd, verifying the even-odd identities.

The next set of fundamental identities is the set of reciprocal identities, which, as their name implies, relate trigonometric functions that are reciprocals of each other. (Table 1.3.3). Recall that we first encountered these identities when defining trigonometric functions from right angles in Right Angle Trigonometry.

Table 1.3.3: Reciprocal Identities

$\sin\theta = \frac{1}{\csc\theta}$	$\csc\theta = \frac{1}{\sin\theta}$
$\cos\theta = \frac{1}{\sec\theta}$	$\sec\theta = \frac{1}{\cos\theta}$
$\tan\theta = \frac{1}{\cot\theta}$	$\cot\theta = \frac{1}{\tan\theta}$

The final set of identities is the set of **quotient identities,** which define relationships among certain trigonometric functions and can be very helpful in verifying other identities (Table 1.3.4).

Table 1.3.4: Quotient Identities

$\tan\theta = \dfrac{\sin\theta}{\cos\theta}$	$\cot\theta = \dfrac{\cos\theta}{\sin\theta}$

The reciprocal and quotient identities are derived from the definitions of the basic trigonometric functions.

SUMMARIZING TRIGONOMETRIC IDENTITIES

The Pythagorean identities are based on the properties of a right triangle.

$$\cos^2\theta + \sin^2\theta = 1 \tag{1.3.5}$$

$$1 + \cot^2\theta = \csc^2\theta \tag{1.3.6}$$

$$1 + \tan^2\theta = \sec^2\theta \tag{1.3.7}$$

The even-odd identities relate the value of a trigonometric function at a given angle to the value of the function at the opposite angle.

$$\tan(-\theta) = -\tan\theta \tag{1.3.8}$$

$$\cot(-\theta) = -\cot\theta \tag{1.3.9}$$

$$\sin(-\theta) = -\sin\theta \tag{1.3.10}$$

$$\csc(-\theta) = -\csc\theta \tag{1.3.11}$$

$$\cos(-\theta) = \cos\theta \tag{1.3.12}$$

$$\sec(-\theta) = \sec\theta \tag{1.3.13}$$

The reciprocal identities define reciprocals of the trigonometric functions.

$$\sin\theta = \frac{1}{\csc\theta} \tag{1.3.14}$$

$$\cos\theta = \frac{1}{\sec\theta} \tag{1.3.15}$$

$$\tan\theta = \frac{1}{\cot\theta} \tag{1.3.16}$$

$$\csc\theta = \frac{1}{\sin\theta} \tag{1.3.17}$$

$$\sec\theta = \frac{1}{\cos\theta} \tag{1.3.18}$$

$$\cot\theta = \frac{1}{\tan\theta} \tag{1.3.19}$$

The **quotient identities** define the relationship among the trigonometric functions.

$$\tan\theta = \frac{\sin\theta}{\cos\theta} \tag{1.3.20}$$

$$\cot\theta = \frac{\cos\theta}{\sin\theta} \tag{1.3.21}$$

How to: Given a trigonometric identity, verify that it is true.

1. Work on one side of the equation. It is usually better to start with the more complex side, as it is easier to simplify than to build.
2. Look for opportunities to factor expressions, square a binomial, or add fractions.
3. Noting which functions are in the final expression, look for opportunities to use the identities and make the proper substitutions.
4. If these steps do not yield the desired result, try converting all terms to sines and cosines.

Example 1.3.2: Verifying a Trigonometric Identity

Verify $\tan\theta\cos\theta = \sin\theta$.

Solution

We will start on the left side, as it is the more complicated side:

$$\begin{aligned}\tan\theta\cos\theta &= \left(\frac{\sin\theta}{\cos\theta}\right)\cos\theta \\ &= \sin\theta\end{aligned}$$

Analysis

This identity was fairly simple to verify, as it only required writing $\tan\theta$ in terms of $\sin\theta$ and $\cos\theta$.

Exercise 1.3.1

Verify the identity $\csc\theta\cos\theta\tan\theta = 1$.

Answer

$$\begin{aligned}\csc\theta\cos\theta\tan\theta = \left(\frac{1}{\sin\theta}\right)\cos\theta\left(\frac{\sin\theta}{\cos\theta}\right) \\ &= \frac{\cos\theta}{\sin\theta}\left(\frac{\sin\theta}{\cos\theta}\right) \\ &= \frac{\sin\theta\cos\theta}{\sin\theta\cos\theta} \\ &= 1\end{aligned}$$

Example 1.3.3A: Verifying the Equivalency Using the Even-Odd Identities

Verify the following equivalency using the even-odd identities:

$$(1+\sin x)[1+\sin(-x)] = \cos^2 x$$

Solution

Working on the left side of the equation, we have

$$(1+\sin x)[1+\sin(-x)] = (1+\sin x)(1-\sin x)$$

Since

$$\begin{aligned}\sin(-x) &= -\sin x \\ [5pt] &= 1-\sin^2 x \qquad \text{Difference of squares} \\ [5pt] &= \cos^2 x \\ \cos^2 x &= 1-\sin^2 x\end{aligned}$$

Example 1.3.3B: Verifying a Trigonometric Identity Involving $\sec^2\theta$

Verify the identity $\dfrac{\sec^2\theta-1}{\sec^2\theta}=\sin^2\theta$

Solution

As the left side is more complicated, let's begin there.

$$\begin{aligned}\frac{\sec^2\theta-1}{\sec^2\theta} &= \frac{(\tan^2\theta+1)-1}{\sec^2\theta}\\ \sec^2\theta &= \tan^2\theta+1\\ &= \frac{\tan^2\theta}{\sec^2\theta}\\ &= \tan^2\theta\left(\frac{1}{\sec^2\theta}\right)\\ &= \tan^2\theta\left(\cos^2\theta\right)\\ \cos^2\theta &= \frac{1}{\sec^2\theta}\\ &= \left(\frac{\sin^2\theta}{\cos^2\theta}\right)\\ \tan^2\theta &= \frac{\sin^2\theta}{\cos^2\theta}\\ &= \sin^2\theta\end{aligned}$$

There is more than one way to verify an identity. Here is another possibility. Again, we can start with the left side.

$$\begin{aligned}\frac{\sec^2\theta-1}{\sec^2\theta} &= \frac{\sec^2\theta}{\sec^2\theta}-\frac{1}{\sec^2\theta}\\ &= 1-\cos^2\theta\\ &= \sin^2\theta\end{aligned}$$

Analysis

In the first method, we used the identity $\sec^2\theta=\tan^2\theta+1$ and continued to simplify. In the second method, we split the fraction, putting both terms in the numerator over the common denominator. This problem illustrates that there are multiple ways we can verify an identity. Employing some creativity can sometimes simplify a procedure. As long as the substitutions are correct, the answer will be the same.

Exercise 1.3.2

Show that $\dfrac{\cot\theta}{\csc\theta}=\cos\theta$.

Answer

$$\begin{aligned}\frac{\cot\theta}{\csc\theta} &= \frac{\frac{\cos\theta}{\sin\theta}}{\frac{1}{\sin\theta}}\\ &= \frac{\cos\theta}{\sin\theta}\cdot\frac{\sin\theta}{1}\\ &= \cos\theta\end{aligned}$$

Example 1.3.4: Creating and Verifying an Identity

Create an identity for the expression $2\tan\theta\sec\theta$ by rewriting strictly in terms of sine.

Solution

There are a number of ways to begin, but here we will use the quotient and reciprocal identities to rewrite the expression:

$$\begin{aligned} 2\tan\theta\sec\theta &= 2\left(\frac{\sin\theta}{\cos\theta}\right)\left(\frac{1}{\cos\theta}\right) \\ &= \frac{2\sin\theta}{\cos^2\theta} \\ &= \frac{2\sin\theta}{1-\sin^2\theta} \qquad \text{Substitute } 1-\sin^2\theta \text{ for } \cos^2\theta \end{aligned}$$

Thus,

$$2\tan\theta\sec\theta = \frac{2\sin\theta}{1-\sin^2\theta}$$

Example 1.3.5: Verifying an Identity Using Algebra and Even/Odd Identities

Verify the identity:

$$\frac{\sin^2(-\theta)-\cos^2(-\theta)}{\sin(-\theta)-\cos(-\theta)} = \cos\theta-\sin\theta$$

Solution

Let's start with the left side and simplify:

$$\begin{aligned} \frac{\sin^2(-\theta)-\cos^2(-\theta)}{\sin(-\theta)-\cos(-\theta)} &= \frac{[\sin(-\theta)]^2-[\cos(-\theta)]^2}{\sin(-\theta)-\cos(-\theta)} \\ &= \frac{(-\sin\theta)^2-(\cos\theta)^2}{-\sin\theta-\cos\theta} \qquad ,\sin(-x)=-\sin\ x \text{ and } \cos(-x)=\cos\ x \\ &= \frac{(\sin\theta)^2-(\cos\theta)^2}{-\sin\theta-\cos\theta} \qquad \text{Difference of squares} \\ &= \frac{(\sin\theta-\cos\theta)(\sin\theta+\cos\theta)}{-(\sin\theta+\cos\theta)} \\ &= \cos\theta-\sin\theta \end{aligned}$$

Exercise 1.3.3

Verify the identity $\dfrac{\sin^2\theta-1}{\tan\theta\sin\theta-\tan\theta} = \dfrac{\sin\theta+1}{\tan\theta}$.

Answer

$$\begin{aligned} \frac{\sin^2\theta-1}{\tan\theta\sin\theta-\tan\theta} &= \frac{(\sin\theta+1)(\sin\theta-1)}{\tan\theta(\sin\theta-1)} \\ &= \frac{\sin\theta+1}{\tan\theta} \end{aligned}$$

Example 1.3.6: Verifying an Identity Involving Cosines and Cotangents

Verify the identity: $(1-\cos^2x)(1+\cot^2x)=1$.

Solution

$$
\begin{aligned}
(1-\cos^2 x)(1+\cot^2 x) &= (1-\cos^2 x)\left(1+\frac{\cos^2 x}{\sin^2 x}\right) \\
&= (1-\cos^2 x)\left(\frac{\sin^2 x}{\sin^2 x}+\frac{\cos^2 x}{\sin^2 x}\right) \qquad \text{Find the common denominator} \\
&= (1-\cos^2 x)\left(\frac{\sin^2 x+\cos^2 x}{\sin^2 x}\right) \\
&= (\sin^2 x)\left(\frac{1}{\sin^2 x}\right) \\
&= 1
\end{aligned}
$$

Solving Equations Involving a Single Trigonometric Function

When we are given equations that involve only one of the six trigonometric functions, their solutions involve using algebraic techniques and the unit circle (see [link]). We need to make several considerations when the equation involves trigonometric functions other than sine and cosine. Problems involving the reciprocals of the primary trigonometric functions need to be viewed from an algebraic perspective. In other words, we will write the reciprocal function, and solve for the angles using the function. Also, an equation involving the tangent function is slightly different from one containing a sine or cosine function. First, as we know, the period of tangent is π,not 2π. Further, the domain of tangent is all real numbers with the exception of odd integer multiples of $\frac{\pi}{2}$,unless, of course, a problem places its own restrictions on the domain.

Example 1.3.7A: Solving a Problem Involving a Single Trigonometric Function

Solve the problem exactly: $2\sin^2\theta-1=0$, $0\le\theta<2\pi$.

Solution

As this problem is not easily factored, we will solve using the square root property. First, we use algebra to isolate $\sin\theta$. Then we will find the angles.

$$
\begin{aligned}
2\sin^2\theta-1 &= 0 \\
2\sin^2\theta &= 1 \\
\sin^2\theta &= \frac{1}{2} \\
\sqrt{\sin^2\theta} &= \pm\sqrt{\frac{1}{2}} \\
\sin\theta &= \pm\frac{1}{\sqrt{2}} \\
&= \pm\frac{\sqrt{2}}{2} \\
\theta &= \frac{\pi}{4},\ \frac{3\pi}{4},\ \frac{5\pi}{4},\ \frac{7\pi}{4}
\end{aligned}
$$

Example 1.3.7B: Solving a Trigonometric Equation Involving Cosecant

Solve the following equation exactly: $\csc\theta=-2$, $0\le\theta<4\pi$.

Solution

We want all values of θ for which $\csc\theta=-2$ over the interval $0\le\theta<4\pi$.

$$\csc\theta = -2$$
$$\frac{1}{\sin\theta} = -2$$
$$\sin\theta = -\frac{1}{2}$$
$$\theta = \frac{7\pi}{6},\ \frac{11\pi}{6},\ \frac{19\pi}{6},\ \frac{23\pi}{6}$$

Analysis

As $\sin\theta = -\frac{1}{2}$, notice that all four solutions are in the third and fourth quadrants.

Example 1.3.7C: Solving an Equation Involving Tangent

Solve the equation exactly: $\tan\left(\theta - \frac{\pi}{2}\right) = 1$, $0 \le \theta < 2\pi$.

Solution

Recall that the tangent function has a period of π. On the interval $[0, \pi)$,and at the angle of $\frac{\pi}{4}$,the tangent has a value of 1. However, the angle we want is $\left(\theta - \frac{\pi}{2}\right)$. Thus, if $\tan\left(\frac{\pi}{4}\right) = 1$,then

$$\theta - \frac{\pi}{2} = \frac{\pi}{4}$$
$$\theta = \frac{3\pi}{4} \pm k\pi$$

Over the interval $[0, 2\pi)$,we have two solutions:

$\theta = \frac{3\pi}{4}$ and $\theta = \frac{3\pi}{4} + \pi = \frac{7\pi}{4}$

Exercise 1.3.4

Find all solutions for $\tan x = \sqrt{3}$.

Answer

$\frac{\pi}{3} \pm \pi k$

Example 1.3.8: Identify all Solutions to the Equation Involving Tangent

Identify all exact solutions to the equation $2(\tan x + 3) = 5 + \tan x$, $0 \le x < 2\pi$.

Solution

We can solve this equation using only algebra. Isolate the expression $\tan x$ on the left side of the equals sign.

$$\begin{aligned} 2(\tan x) + 2(3) &= 5 + \tan x \\ 2\tan x + 6 &= 5 + \tan x \\ 2\tan x - \tan x &= 5 - 6 \\ \tan x &= -1 \end{aligned}$$

There are two angles on the unit circle that have a tangent value of -1: $\theta = \frac{3\pi}{4}$ and $\theta = \frac{7\pi}{4}$.

Solving Trigonometric Equations in Quadratic Form

Solving a *quadratic equation* may be more complicated, but once again, we can use algebra as we would for any quadratic equation. Look at the pattern of the equation. Is there more than one trigonometric function in the equation, or is there only one?

Which trigonometric function is squared? If there is only one function represented and one of the terms is squared, think about the standard form of a quadratic. Replace the trigonometric function with a variable such as x or u. If substitution makes the equation look like a quadratic equation, then we can use the same methods for solving quadratics to solve the trigonometric equations.

Example 1.3.9A: Solving a Trigonometric Equation in Quadratic Form

Solve the equation exactly: $\cos^2\theta + 3\cos\theta - 1 = 0$, $0 \le \theta < 2\pi$.

Solution

We begin by using substitution and replacing $\cos\theta$ with x. It is not necessary to use substitution, but it may make the problem easier to solve visually. Let $\cos\theta = x$. We have

$$x^2 + 3x - 1 = 0$$

The equation cannot be factored, so we will use the **quadratic formula:** $x = \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

$$\begin{aligned} x &= \frac{-3 \pm \sqrt{(-3)^2 - 4(1)(-1)}}{2} \\ &= \frac{-3 \pm \sqrt{13}}{2} \end{aligned}$$

Replace x with $\cos\theta$ and solve.

$$\begin{aligned} \cos\theta &= \frac{-3 \pm \sqrt{13}}{2} \\ \theta &= \cos^{-1}\left(\frac{-3 + \sqrt{13}}{2}\right) \end{aligned}$$

Note that only the + sign is used. This is because we get an error when we solve $\theta = \cos^{-1}\left(\dfrac{-3 - \sqrt{13}}{2}\right)$ on a calculator, since the domain of the inverse cosine function is $[-1, 1]$. However, there is a second solution:

$$\begin{aligned} \theta &= \cos^{-1}\left(\frac{-3 + \sqrt{13}}{2}\right) \\ &\approx 1.26 \end{aligned}$$

This terminal side of the angle lies in quadrant I. Since cosine is also positive in quadrant IV, the second solution is

$$\begin{aligned} \theta &= 2\pi - \cos^{-1}\left(\frac{-3 + \sqrt{13}}{2}\right) \\ &\approx 5.02 \end{aligned}$$

Example 1.3.9B: Solving a Trigonometric Equation in Quadratic Form by Factoring

Solve the equation exactly: $2\sin^2\theta - 5\sin\theta + 3 = 0$, $0 \le \theta \le 2\pi$.

Solution

Using grouping, this quadratic can be factored. Either make the real substitution, $\sin\theta = u$,or imagine it, as we factor:

$$\begin{aligned} 2\sin^2\theta - 5\sin\theta + 3 &= 0 \\ (2\sin\theta - 3)(\sin\theta - 1) &= 0 \qquad \text{Now set each factor equal to zero.} \\ 2\sin\theta - 3 &= 0 \\ 2\sin\theta &= 3 \\ \sin\theta &= \frac{3}{2} \\ \sin\theta - 1 &= 0 \\ \sin\theta &= 1 \end{aligned}$$

Next solve for θ: $\sin\theta \neq \frac{3}{2}$, as the range of the sine function is $[-1, 1]$. However, $\sin\theta = 1$, giving the solution $\theta = \frac{\pi}{2}$.

Analysis

Make sure to check all solutions on the given domain as some factors have no solution.

Exercise 1.3.5

Solve $\sin^2\theta = 2\cos\theta + 2$, $0 \leq \theta \leq 2\pi$. [Hint: Make a substitution to express the equation only in terms of cosine.]

Answer

$\cos\theta = -1$, $\theta = \pi$

Example 1.3.10A: Solving a Trigonometric Equation Using Algebra

Solve exactly: $2\sin^2\theta + \sin\theta = 0;\ 0 \leq \theta < 2\pi$

Solution

This problem should appear familiar as it is similar to a quadratic. Let $\sin\theta = x$. The equation becomes $2x^2 + x = 0$. We begin by factoring:

$$\begin{aligned} 2x^2 + x &= 0 \\ x(2x+1) &= 0 \qquad \text{Set each factor equal to zero.} \\ x &= 0 \\ 2x + 1 &= 0 \\ x &= -\frac{1}{2} \end{aligned}$$

Then, substitute back into the equation the original expression $\sin\theta$ for x. Thus,

$$\begin{aligned} \sin\theta &= 0 \\ \theta &= 0, \pi \\ \sin\theta &= -\frac{1}{2} \\ \theta &= \frac{7\pi}{6}, \frac{11\pi}{6} \end{aligned}$$

The solutions within the domain $0 \leq \theta < 2\pi$ are $\theta = 0, \pi, \frac{7\pi}{6}, \frac{11\pi}{6}$.

If we prefer not to substitute, we can solve the equation by following the same pattern of factoring and setting each factor equal to zero.

$$\begin{aligned} \sin^2\theta + \sin\theta &= 0 \\ \sin\theta(2\sin\theta + 1) &= 0 \\ \sin\theta &= 0 \\ \theta &= 0, \pi \\ 2\sin\theta + 1 &= 0 \\ 2\sin\theta &= -1 \\ \sin\theta &= -\frac{1}{2} \\ \theta &= \frac{7\pi}{6}, \frac{11\pi}{6} \end{aligned}$$

Analysis

We can see the solutions on the graph in Figure 1.3.3. On the interval $0 \le \theta < 2\pi$,the graph crosses the x-axis four times, at the solutions noted. Notice that trigonometric equations that are in quadratic form can yield up to four solutions instead of the expected two that are found with quadratic equations. In this example, each solution (angle) corresponding to a positive sine value will yield two angles that would result in that value.

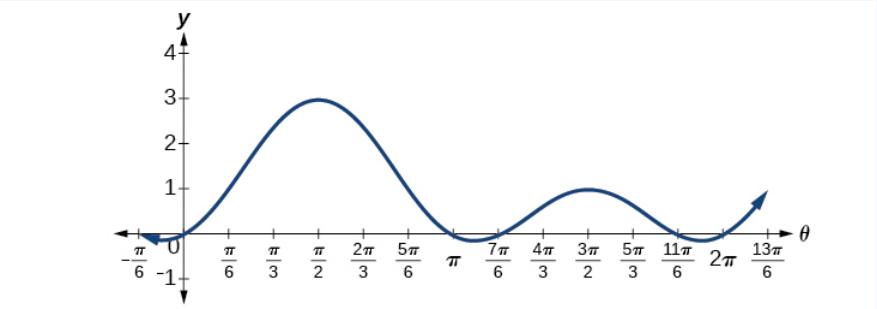


Figure 1.3.3

We can verify the solutions on the unit circle in via the result in the section on Sum and Difference Identities as well.

Example 1.3.10B: Solving a Trigonometric Equation Quadratic in Form

Solve the equation quadratic in form exactly: $2\sin^2\theta - 3\sin\theta + 1 = 0$, $0 \le \theta < 2\pi$.

Solution

We can factor using grouping. Solution values of θ can be found on the unit circle.

$$\begin{aligned}(2\sin\theta - 1)(\sin\theta - 1) &= 0\\ 2\sin\theta - 1 &= 0\\ \sin\theta &= \frac{1}{2}\\ \theta &= \frac{\pi}{6}, \frac{5\pi}{6}\\ \sin\theta &= 1\\ \theta &= \frac{\pi}{2}\end{aligned}$$

Exercise 1.3.6

Solve the quadratic equation $2\cos^2\theta + \cos\theta = 0$.

Answer

$\frac{\pi}{2}, \frac{2\pi}{3}, \frac{4\pi}{3}, \frac{3\pi}{2}$

This page titled 1.3: Trigonometric Functions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **1.3: Trigonometric Functions** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **9.1: Solving Trigonometric Equations with Identities** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.
- **9.5: Solving Trigonometric Equations** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

1.4: Inverse Functions

Learning Objectives

- Determine the conditions for when a function has an inverse.
- Use the horizontal line test to recognize when a function is one-to-one.
- Find the inverse of a given function.
- Draw the graph of an inverse function.
- Evaluate inverse trigonometric functions.

An inverse function reverses the operation done by a particular function. In other words, whatever a function does, the inverse function undoes it. In this section, we define an inverse function formally and state the necessary conditions for an inverse function to exist. We examine how to find an inverse function and study the relationship between the graph of a function and the graph of its inverse. Then we apply these ideas to define and discuss properties of the inverse trigonometric functions.

Existence of an Inverse Function

We begin with an example. Given a function f and an output $y = f(x)$, we are often interested in finding what value or values x were mapped to y by f. For example, consider the function $f(x) = x^3 + 4$. Since any output $y = x^3 + 4$, we can solve this equation for x to find that the input is $x = \sqrt[3]{y-4}$. This equation defines x as a function of y. Denoting this function as f^{-1}, and writing $x = f^{-1}(y) = \sqrt[3]{y-4}$, we see that for any x in the domain of f, $f^{-1}f(x)) = f^{-1}(x^3+4) = x$. Thus, this new function, f^{-1}, "undid" what the original function f did. A function with this property is called the inverse function of the original function.

Definition: Inverse Functions

Given a function f with domain D and range R, its *inverse function* (if it exists) is the function f^{-1} with domain R and range D such that $f^{-1}(y) = x$ if and only if $f(x) = y$. In other words, for a function f and its inverse f^{-1},

$$f^{-1}(f(x)) = x$$

for all x in D and

$$f(f^{-1}(y)) = y$$

for all y in R.

Note that f^{-1} is read as "f inverse." Here, the -1 is not used as an exponent so

$$f^{-1}(x) \neq \frac{1}{f(x)}.$$

Figure 1.4.1shows the relationship between the domain and range of f and the domain and range of f^{-1}.

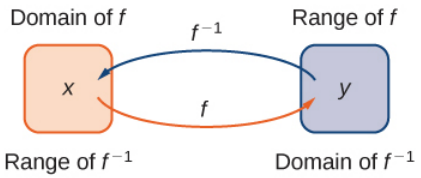


Figure 1.4.1: Given a function f and its inverse f^{-1}, $f^{-1}(y) = x$ if and only if $f(x) = y$. The range of f becomes the domain of f^{-1} and the domain of f becomes the range of f^{-1}.

Recall that a function has exactly one output for each input. Therefore, to define an inverse function, we need to map each input to exactly one output. For example, let's try to find the inverse function for $f(x) = x^2$. Solving the equation $y = x^2$ for x, we arrive at the equation $x = \pm\sqrt{y}$. This equation does not describe x as a function of y because there are two solutions to this equation for every $y > 0$. The problem with trying to find an inverse function for $f(x) = x^2$ is that two inputs are sent to the same output for each output $y > 0$. The function $f(x) = x^3 + 4$ discussed earlier did not have this problem. For that function, each input was sent to a different output. A function that sends each input to a different output is called a one-to-one function.

Definition: One-to-One functions

We say a function f is a one-to-one function if $f(x_1) \neq f(x_2)$ when $x_1 \neq x_2$.

One way to determine whether a function is one-to-one is by looking at its graph. If a function is one-to-one, then no two inputs can be sent to the same output. Therefore, if we draw a horizontal line anywhere in the xy-plane, according to the **horizontal line test**, it cannot intersect the graph more than once. We note that the horizontal line test is different from the vertical line test. The vertical line test determines whether a graph is the graph of a function. The horizontal line test determines whether a function is one-to-one (Figure 1.4.2).

Horizontal Line Test

A function f is one-to-one if and only if every horizontal line intersects the graph of f no more than once.

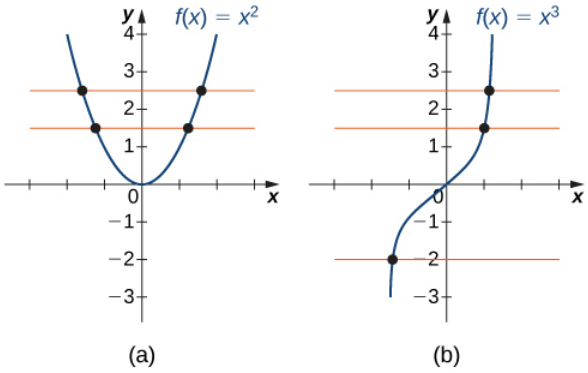
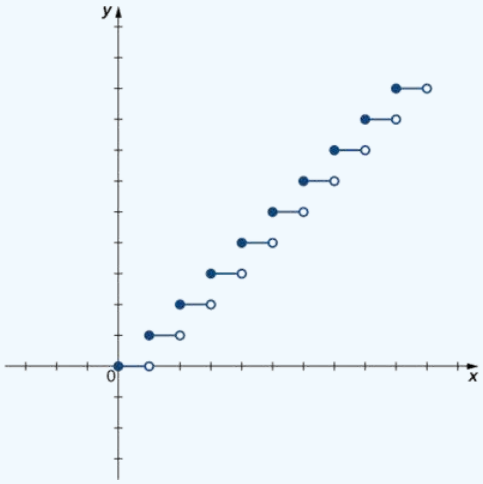


Figure 1.4.2: (a) The function $f(x) = x^2$ is not one-to-one because it fails the horizontal line test. (b) The function $f(x) = x^3$ is one-to-one because it passes the horizontal line test.

Example 1.4.1: Determining Whether a Function Is One-to-One

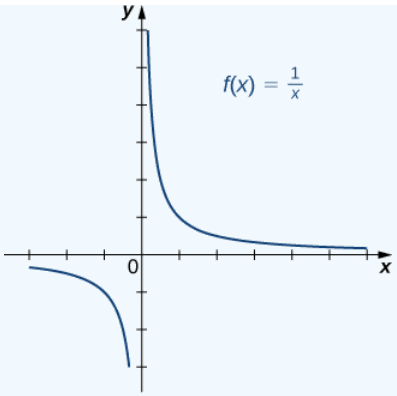
For each of the following functions, use the horizontal line test to determine whether it is one-to-one.

a)



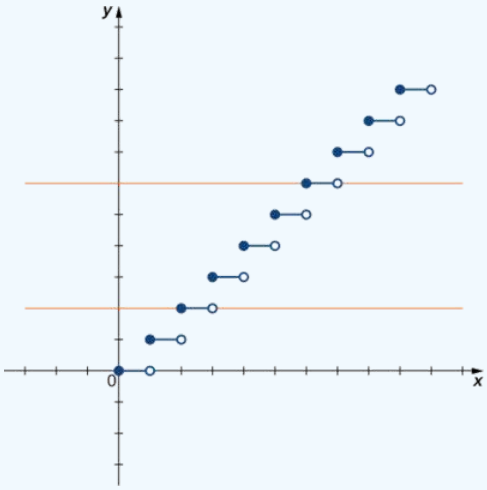
b)

https://math.libretexts.org/@go/page/139042

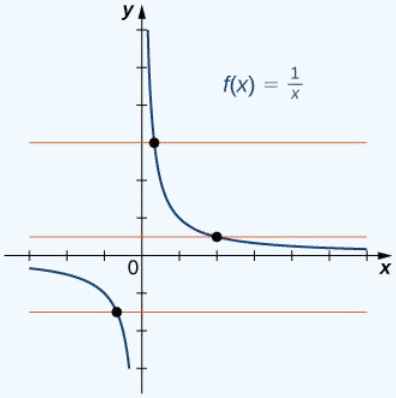


Solution

a) Since the horizontal line $y = n$ for any integer $n \geq 0$ intersects the graph more than once, this function is not one-to-one.

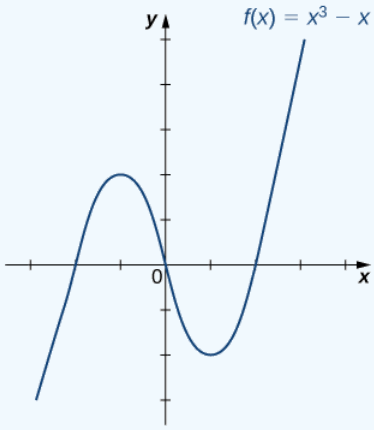


b) Since every horizontal line intersects the graph once (at most), this function is one-to-one.



? Exercise 1.4.1

Is the function f graphed in the following image one-to-one?



Solution

Use the horizontal line test.

Answer

No

Finding a Function's Inverse

We can now consider one-to-one functions and show how to find their inverses. Recall that a function maps elements in the domain of f to elements in the range of f. The inverse function maps each element from the range of f back to its corresponding element from the domain of f. Therefore, to find the inverse function of a one-to-one function f, given any y in the range of f, we need to determine which x in the domain of f satisfies $f(x) = y$. Since f is one-to-one, there is exactly one such value x. We can find that value x by solving the equation $f(x) = y$ for x. Doing so, we are able to write x as a function of y where the domain of this function is the range of f and the range of this new function is the domain of f. Consequently, this function is the inverse of f, and we write $x = f^{-1}(y)$. Since we typically use the variable x to denote the independent variable and y to denote the dependent variable, we often interchange the roles of x and y, and write $y = f^{-1}(x)$. Representing the inverse function in this way is also helpful later when we graph a function f and its inverse f^{-1} on the same axes.

Problem-Solving Strategy: Finding an Inverse Function

1. Solve the equation $y = f(x)$ for x.
2. Interchange the variables x and y and write $y = f^{-1}(x)$.

Example 1.4.2: Finding an Inverse Function

Find the inverse for the function $f(x) = 3x - 4$. State the domain and range of the inverse function. Verify that $f^{-1}(f(x)) = x$.

Solution

Follow the steps outlined in the strategy.

Step 1. If $y = 3x - 4$, then $3x = y + 4$ and $x = \frac{1}{3}y + \frac{4}{3}$.

Step 2. Rewrite as $y = \frac{1}{3}x + \frac{4}{3}$ and let $y = f^{-1}(x)$.Therefore, $f^{-1}(x) = \frac{1}{3}x + \frac{4}{3}$.

Since the domain of f is $(-\infty, \infty)$, the range of f^{-1} is $(-\infty, \infty)$. Since the range of f is $(-\infty, \infty)$, the domain of f^{-1} is $(-\infty, \infty)$.

You can verify that $f^{-1}(f(x)) = x$ by writing

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

Note that for $f^{-1}(x)$ to be the inverse of $f(x)$, both $f^{-1}(f(x)) = x$ and $f(f^{-1}(x)) = x$ for all x in the domain of the inside function.

Exercise 1.4.2

Find the inverse of the function $f(x) = 3x/(x-2)$. State the domain and range of the inverse function.

Hint

Use the Problem-Solving Strategy for finding inverse functions.

Answer

$f^{-1}(x) = \dfrac{2x}{x-3}$. The domain of f^{-1} is $\{x \mid x \neq 3\}$. The range of f^{-1} is $\{y \mid y \neq 2\}$.

Graphing Inverse Functions

Let's consider the relationship between the graph of a function f and the graph of its inverse. Consider the graph of f shown in Figure 1.4.3 and a point (a, b) on the graph. Since $b = f(a)$, then $f^{-1}(b) = a$. Therefore, when we graph f^{-1}, the point (b, a) is on the graph. As a result, the graph of f^{-1} is a reflection of the graph of f about the line $y = x$.

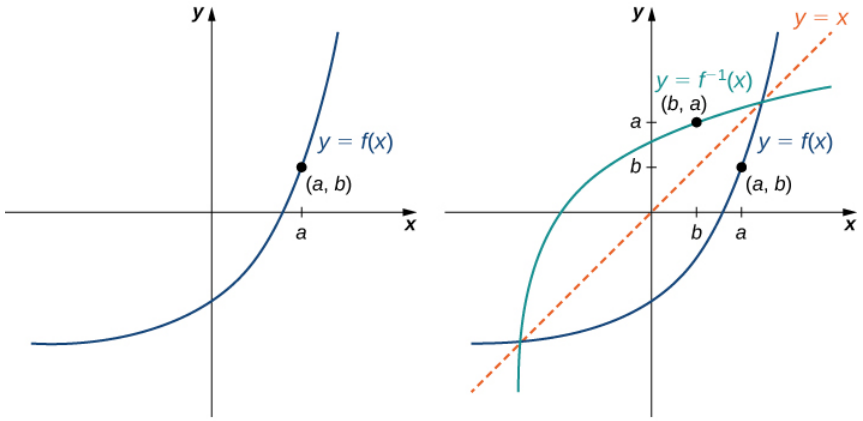
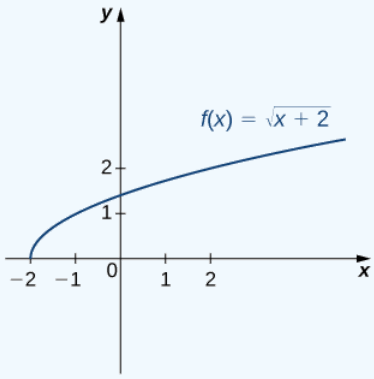


Figure 1.4.3: (a) The graph of this function f shows point (a, b) on the graph of f. (b) Since (a, b) is on the graph of f, the point (b, a) is on the graph of f^{-1}. The graph of f^{-1} is a reflection of the graph of f about the line $y = x$.

Example 1.4.3: Sketching Graphs of Inverse Functions

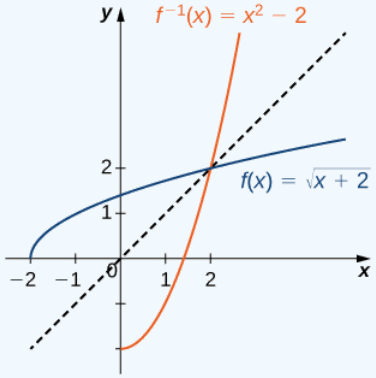
For the graph of f in the following image, sketch a graph of f^{-1} by sketching the line $y = x$ and using symmetry. Identify the domain and range of f^{-1}.



Solution

https://math.libretexts.org/@go/page/139042

Reflect the graph about the line $y = x$. The domain of f^{-1} is $[0, \infty)$. The range of f^{-1} is $[-2, \infty)$. By using the preceding strategy for finding inverse functions, we can verify that the inverse function is $f^{-1}(x) = x^2 - 2$, as shown in the graph.



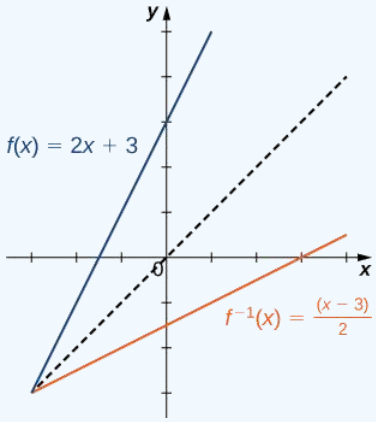
? Exercise 1.4.3

Sketch the graph of $f(x) = 2x + 3$ and the graph of its inverse using the symmetry property of inverse functions.

Hint

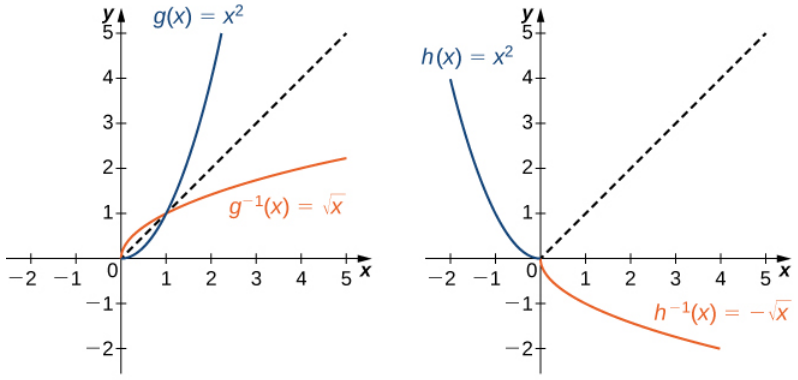
The graphs are symmetric about the line $y = x$

Answer



Restricting Domains

As we have seen, $f(x) = x^2$ does not have an inverse function because it is not one-to-one. However, we can choose a subset of the domain of f such that the function is one-to-one. This subset is called a **restricted domain**. By restricting the domain of f, we can define a new function g such that the domain of g is the restricted domain of f and $g(x) = f(x)$ for all x in the domain of g. Then we can define an inverse function for g on that domain. For example, since $f(x) = x^2$ is one-to-one on the interval $[0, \infty)$, we can define a new function g such that the domain of g is $[0, \infty)$ and $g(x) = x^2$ for all x in its domain. Since g is a one-to-one function, it has an inverse function, given by the formula $g^{-1}(x) = \sqrt{x}$. On the other hand, the function $f(x) = x^2$ is also one-to-one on the domain $(-\infty, 0]$. Therefore, we could also define a new function h such that the domain of h is $(-\infty, 0]$ and $h(x) = x^2$ for all x in the domain of h. Then h is a one-to-one function and must also have an inverse. Its inverse is given by the formula $h^{-1}(x) = -\sqrt{x}$ (Figure 1.4.4).



(a) (b)

Figure 1.4.4: (a) For $g(x) = x^2$ restricted to $[0, \infty)$, $g^{-1}(x) = \sqrt{x}$. (b) For $h(x) = x^2$ restricted to $(-\infty, 0]$, $h^{-1}(x) = -\sqrt{x}$.

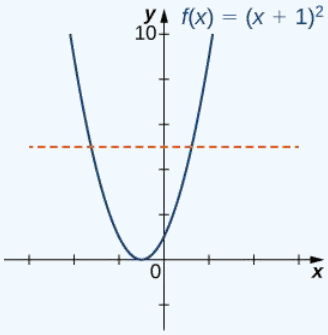
Example 1.4.4: Restricting the Domain

Consider the function $f(x) = (x+1)^2$.

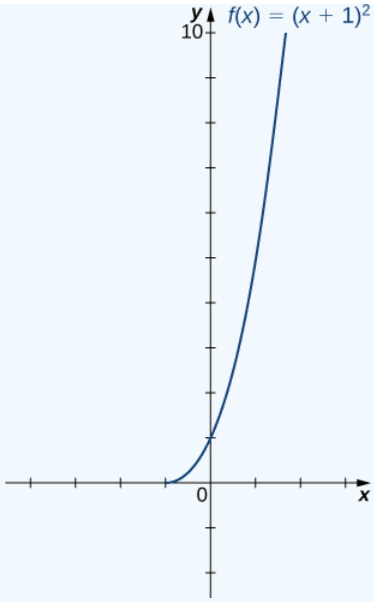
a. Sketch the graph of f and use the horizontal line test to show that f is not one-to-one.
b. Show that f is one-to-one on the restricted domain $[-1, \infty)$. Determine the domain and range for the inverse of f on this restricted domain and find a formula for f^{-1}.

Solution

a) The graph of f is the graph of $y = x^2$ shifted left 1 unit. Since there exists a horizontal line intersecting the graph more than once, f is not one-to-one.



b) On the interval $[-1, \infty)$, f is one-to-one.



The domain and range of f^{-1} are given by the range and domain of f, respectively. Therefore, the domain of f^{-1} is $[0,\infty)$ and the range of f^{-1} is $[-1,\infty)$. To find a formula for f^{-1}, solve the equation $y=(x+1)^2$ for x. If $y=(x+1)^2$, then $x=-1\pm\sqrt{y}$. Since we are restricting the domain to the interval where $x\geq -1$, we need $\pm\sqrt{y}\geq 0$. Therefore, $x=-1+\sqrt{y}$. Interchanging x and y, we write $y=-1+\sqrt{x}$ and conclude that $f^{-1}(x)=-1+\sqrt{x}$.

? Exercise 1.4.4

Consider $f(x)=1/x^2$ restricted to the domain $(-\infty,0)$. Verify that f is one-to-one on this domain. Determine the domain and range of the inverse of f and find a formula for f^{-1}.

Hint

The domain and range of f^{-1} is given by the range and domain of f, respectively. To find f^{-1}, solve $y=1/x^2$ for x.

Answer

The domain of f^{-1} is $(0,\infty)$. The range of f^{-1} is $(-\infty,0)$. The inverse function is given by the formula $f^{-1}(x)=-1/\sqrt{x}$.

Inverse Trigonometric Functions

The six basic trigonometric functions are periodic, and therefore they are not one-to-one. However, if we restrict the domain of a trigonometric function to an interval where it is one-to-one, we can define its inverse. Consider the sine function. The sine function is one-to-one on an infinite number of intervals, but the standard convention is to restrict the domain to the interval $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$. By doing so, we define the inverse sine function on the domain $[-1,1]$ such that for any x in the interval $[-1,1]$, the inverse sine function tells us which angle θ in the interval $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$ satisfies $\sin\theta=x$. Similarly, we can restrict the domains of the other trigonometric functions to define **inverse trigonometric functions**, which are functions that tell us which angle in a certain interval has a specified trigonometric value.

Definition: inverse trigonometric functions

The inverse sine function, denoted $\sin^{-1}$ or arcsin, and the inverse cosine function, denoted $\cos^{-1}$ or arccos, are defined on the domain $D=\{x|-1\leq x\leq 1\}$ as follows:

$$\sin^{-1}(x)=y$$

- if and only if $\sin(y)=x$ and $-\frac{\pi}{2}\leq y\leq\frac{\pi}{2}$;

$$\cos^{-1}(x) = y$$

- if and only if $\cos(y) = x$ and $0 \le y \le \pi$.

The inverse tangent function, denoted $\tan^{-1}$ or $\arctan$, and inverse cotangent function, denoted $\cot^{-1}$ or arccot, are defined on the domain $D = \{x | -\infty < x < \infty\}$ as follows:

$$\tan^{-1}(x) = y$$

- if and only if $\tan(y) = x$ and $-\frac{\pi}{2} < y < \frac{\pi}{2}$;

$$\cot^{-1}(x) = y$$

- if and only if $\cot(y) = x$ and $0 < y < \pi$.

The inverse cosecant function, denoted $\csc^{-1}$ or arccsc, and inverse secant function, denoted $\sec^{-1}$ or arcsec, are defined on the domain $D = \{x \,|\, |x| \ge 1\}$ as follows:

$$\csc^{-1}(x) = y$$

- if and only if $\csc(y) = x$ and $-\frac{\pi}{2} \le y \le \frac{\pi}{2}$, $y \ne 0$;

$$\sec^{-1}(x) = y$$

- if and only if $\sec(y) = x$ and$0 \le y \le \pi$, $y \ne \pi/2$.

To graph the inverse trigonometric functions, we use the graphs of the trigonometric functions restricted to the domains defined earlier and reflect the graphs about the line $y = x$ (Figure 1.4.5).

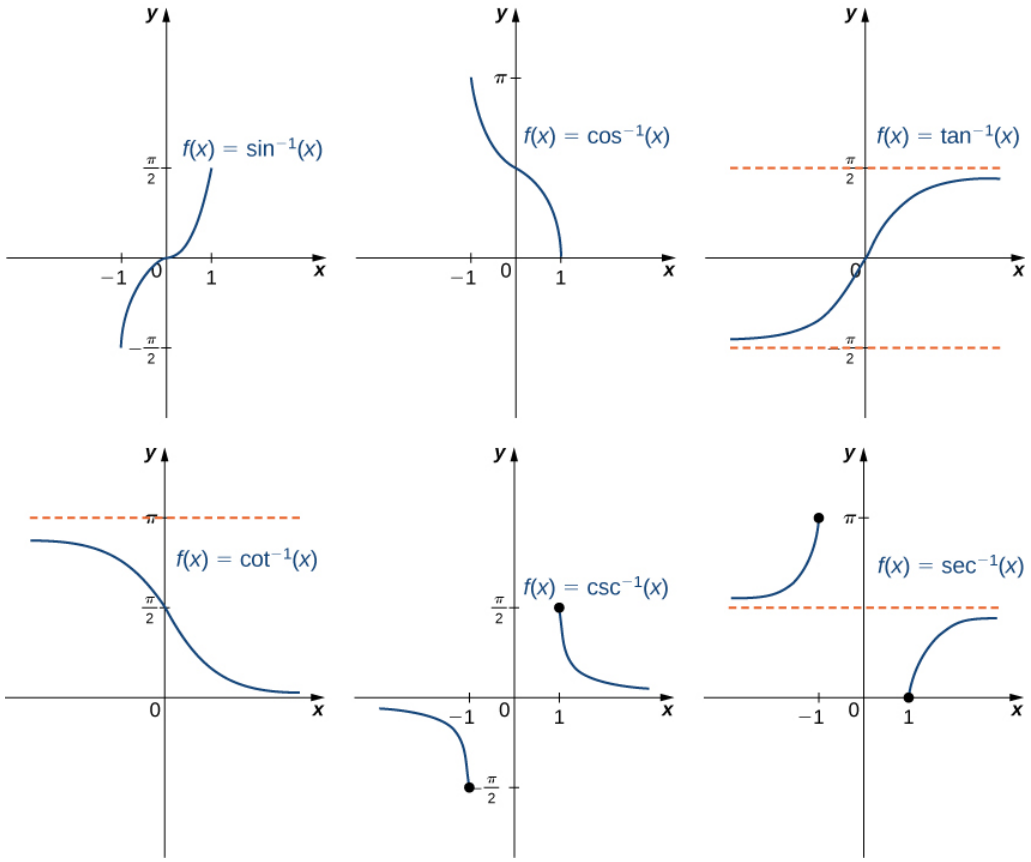


Figure 1.4.5: The graph of each of the inverse trigonometric functions is a reflection about the line $y = x$ of the corresponding restricted trigonometric function.

When evaluating an inverse trigonometric function, the output is an angle. For example, to evaluate $\cos^{-1}\left(\frac{1}{2}\right)$, we need to find an angle θ such that $\cos\theta = \frac{1}{2}$. Clearly, many angles have this property. However, given the definition of $\cos^{-1}$, we need the angle θ that not only solves this equation, but also lies in the interval $[0, \pi]$. We conclude that $\cos^{-1}\left(\frac{1}{2}\right) = \frac{\pi}{3}$.

We now consider a composition of a trigonometric function and its inverse. For example, consider the two expressions $\sin\left(\sin^{-1}\left(\frac{\sqrt{2}}{2}\right)\right)$ and $\sin^{-1}(\sin(\pi))$.

For the first one, we simplify as follows:

$$\sin\left(\sin^{-1}\left(\frac{\sqrt{2}}{2}\right)\right)=\sin\left(\frac{\pi}{4}\right)=\frac{\sqrt{2}}{2}.$$

For the second one, we have

$$\sin^{-1}(\sin(\pi))=\sin^{-1}(0)=0.$$

The inverse function is supposed to "undo" the original function, so why isn't $\sin^{-1}(\sin(\pi))=\pi$? Recalling our definition of inverse functions, a function f and its inverse f^{-1} satisfy the conditions $f(f^{-1}(y))=y$ for all y in the domain of f^{-1} and $f^{-1}(f(x))=x$ for all x in the domain of f, so what happened here? The issue is that the inverse sine function, $\sin^{-1}$, is the inverse of the restricted sine function defined on the domain $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$. Therefore, for x in the interval $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$, it is true that $\sin^{-1}(\sin x)=x$. However, for values of x outside this interval, the equation does not hold, even though $\sin^{-1}(\sin x)$ is defined for all real numbers x.

What about $\sin(\sin^{-1}y)$? Does that have a similar issue? The answer is no. Since the domain of $\sin^{-1}$ is the interval $[-1,1]$, we conclude that $\sin\left(\sin^{-1}y\right)=y$ if $-1\le y\le 1$ and the expression is not defined for other values of y. To summarize,

$$\sin(\sin^{-1}y)=y \text{ if } -1\le y\le 1$$

and

$$\sin^{-1}(\sin x)=x \text{ if } -\tfrac{\pi}{2}\le x\le\tfrac{\pi}{2}.$$

Similarly, for the cosine function,

$$\cos(\cos^{-1}y)=y \text{ if } -1\le y\le 1$$

and

$$\cos^{-1}(\cos x)=x \text{ if } 0\le x\le\pi.$$

Similar properties hold for the other trigonometric *functions and their inverses.*

Example 1.4.5: Evaluating Expressions Involving Inverse Trigonometric Functions

Evaluate each of the following expressions.

a. $\sin^{-1}\left(-\frac{\sqrt{3}}{2}\right)$
b. $\tan\left(\tan^{-1}\left(-\frac{1}{\sqrt{3}}\right)\right)$
c. $\cos^{-1}\left(\cos\left(\frac{5\pi}{4}\right)\right)$
d. $\sin^{-1}\left(\cos\left(\frac{2\pi}{3}\right)\right)$

Solution

a. Evaluating $\sin^{-1}(-\sqrt{3}/2)$ is equivalent to finding the angle θ such that $\sin\theta=-\sqrt{3}/2$ and $-\pi/2\le\theta\le\pi/2$. The angle $\theta=-\pi/3$ satisfies these two conditions. Therefore, $\sin^{-1}(-\sqrt{3}/2)=-\pi/3$.
b. First we use the fact that $\tan^{-1}(-1/\sqrt{3})=-\pi/6$. Then $\tan(-\pi/6)=-1/\sqrt{3}$. Therefore, $\tan(\tan^{-1}(-1/\sqrt{3}))=-1/\sqrt{3}$.
c. To evaluate $\cos^{-1}(\cos(5\pi/4))$, first use the fact that $\cos(5\pi/4)=-\sqrt{2}/2$. Then we need to find the angle θ such that $\cos(\theta)=-\sqrt{2}/2$ and $0\le\theta\le\pi$. Since $3\pi/4$ satisfies both these conditions, we have $\cos^{-1}(\cos(5\pi/4))=\cos^{-1}(-\sqrt{2}/2))=3\pi/4$.
d. Since $\cos(2\pi/3)=-1/2$, we need to evaluate $\sin^{-1}(-1/2)$. That is, we need to find the angle θ such that $\sin(\theta)=-1/2$ and $-\pi/2\le\theta\le\pi/2$. Since $-\pi/6$ satisfies both these conditions, we can conclude that $\sin^{-1}(\cos(2\pi/3))=\sin^{-1}(-1/2)=-\pi/6$.

The Maximum Value of a Function

In many areas of science, engineering, and mathematics, it is useful to know the maximum value a function can obtain, even if we don't know its exact value at a given instant. For instance, if we have a function describing the strength of a roof beam, we would want to know the maximum weight the beam can support without breaking. If we have a function that describes the speed of a train, we would want to know its maximum speed before it jumps off the rails. Safe design often depends on knowing maximum values.

This project describes a simple example of a function with a maximum value that depends on two equation coefficients. We will see that maximum values can depend on several factors other than the independent variable x.

1. Consider the graph in Figure 1.4.6 of the function $y = \sin x + \cos x$. Describe its overall shape. Is it periodic? How do you know?

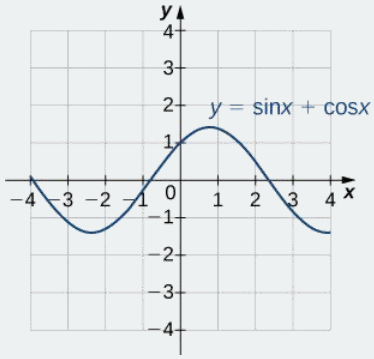


Figure 1.4.6: The graph of $y = \sin x + \cos x$.

Using a graphing calculator or other graphing device, estimate the x- and y-values of the maximum point for the graph (the first such point where $x > 0$). It may be helpful to express the x-value as a multiple of π.

2. Now consider other graphs of the form $y = A\sin x + B\cos x$ for various values of A and B. Sketch the graph when $A = 2$ and $B = 1$, and find the x- and y-values for the maximum point. (Remember to express the x-value as a multiple of π, if possible.) Has it moved?

3. Repeat for $A = 1$, $B = 2$. Is there any relationship to what you found in part (2)?

4. Complete the following table, adding a few choices of your own for A and B :

A	B	x	y	A	B	x	y
0	1			3	4		
1	0			4	3		
1	1			$\sqrt{3}$	1		
1	2			1	$\sqrt{3}$		
2	1			12	5		
2	2			5	12		

5. Try to figure out the formula for the y-values.

6. The formula for the x-values is a little harder. The most helpful points from the table are $(1, 1)$, $(1, \sqrt{3})$, $(\sqrt{3}, 1)$. (Hint: Consider inverse trigonometric functions.)

7. If you found formulas for parts (5) and (6), show that they work together. That is, substitute the x-value formula you found into $y = A\sin x + B\cos x$ and simplify it to arrive at the y-value formula you found.

Key Concepts

- For a function to have an inverse, the function must be one-to-one. Given the graph of a function, we can determine whether the function is one-to-one by using the horizontal line test.
- If a function is not one-to-one, we can restrict the domain to a smaller domain where the function is one-to-one and then define the inverse of the function on the smaller domain.
- For a function f and its inverse f^{-1}, $f(f^{-1}(x)) = x$ for all x in the domain of f^{-1} and $f^{-1}(f(x)) = x$ for all x in the domain of f.
- Since the trigonometric functions are periodic, we need to restrict their domains to define the inverse trigonometric functions.
- The graph of a function f and its inverse f^{-1} are symmetric about the line $y = x$.

Key Equations

- **Inverse function**

  $f^{-1}(f(x)) = x$ for all x in D, and $f(f^{-1}(y)) = y$ for all y in R.

Glossary

horizontal line test
a function f is one-to-one if and only if every horizontal line intersects the graph of f, at most, once

inverse function
for a function f, the inverse function f^{-1} satisfies $f^{-1}(y) = x$ if $f(x) = y$

inverse trigonometric functions
the inverses of the trigonometric functions are defined on restricted domains where they are one-to-one functions

one-to-one function
a function f is one-to-one if $f(x_1) \neq f(x_2)$ if $x_1 \neq x_2$

restricted domain
a subset of the domain of a function f

This page titled 1.4: Inverse Functions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **1.4: Inverse Functions** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

1.5: Exponential and Logarithmic Functions

Learning Objectives

- Identify the form of an exponential function.
- Explain the difference between the graphs of x^b and b^x.
- Recognize the significance of the number e.
- Identify the form of a logarithmic function.
- Explain the relationship between exponential and logarithmic functions.
- Describe how to calculate a logarithm to a different base.

Graphing Exponential Functions

For any base $b > 0$, $b \neq 1$, the exponential function $f(x) = b^x$ is defined for all real numbers x and $b^x > 0$. Therefore, the domain of $f(x) = b^x$ is $(-\infty, \infty)$ and the range is $(0, \infty)$. To graph b^x, we note that for $b > 1$, b^x is increasing on $(-\infty, \infty)$ and $b^x \to \infty$ as $x \to \infty$, whereas $b^x \to 0$ as $x \to -\infty$. On the other hand, if $0 < b < 1$, $f(x) = b^x$ is decreasing on $(-\infty, \infty)$ and $b^x \to 0$ as $x \to \infty$ whereas $b^x \to \infty$ as $x \to -\infty$ (Figure 1.5.2).

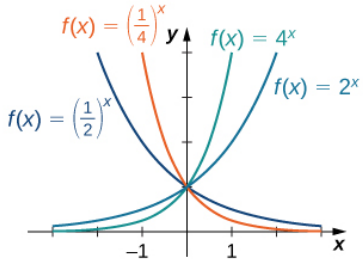


Figure 1.5.2: If $b > 1$, then b^x is increasing on $(-\infty, \infty)$. If $0 < b < 1$, then b^x is decreasing on $(-\infty, \infty)$.

Note that exponential functions satisfy the general laws of exponents. To remind you of these laws, we state them as rules.

Laws of Exponents

For any constants $a > 0$, $b > 0$, and for all x and y,

1. $b^x \cdot b^y = b^{x+y}$
2. $\dfrac{b^x}{b^y} = b^{x-y}$
3. $(b^x)^y = b^{xy}$
4. $(ab)^x = a^x b^x$
5. $\dfrac{a^x}{b^x} = \left(\dfrac{a}{b}\right)^x$

Example 1.5.2: Using the Laws of Exponents

Use the laws of exponents to simplify each of the following expressions.

a. $\dfrac{(2x^{2/3})^3}{(4x^{-1/3})^2}$

b. $\dfrac{(x^3y^{-1})^2}{(xy^2)^{-2}}$

Soution

a. We can simplify as follows:

$$\frac{(2x^{2/3})^3}{(4x^{-1/3})^2} = \frac{2^3(x^{2/3})^3}{4^2(x^{-1/3})^2} = \frac{8x^2}{16x^{-2/3}} = \frac{x^2x^{2/3}}{2} = \frac{x^{8/3}}{2}.$$

b. We can simplify as follows:

$$\frac{(x^3y^{-1})^2}{(xy^2)^{-2}} = \frac{(x^3)^2(y^{-1})^2}{x^{-2}(y^2)^{-2}} = \frac{x^6y^{-2}}{x^{-2}y^{-4}} = x^6x^2y^{-2}y^4 = x^8y^2.$$

? Exercise 1.5.2

Use the laws of exponents to simplify $\dfrac{6x^{-3}y^2}{12x^{-4}y^5}$.

Hint

$x^a/x^b = x^{a-b}$

Answer

$x/(2y^3)$

The Number e

A special type of exponential function appears frequently in real-world applications. To describe it, consider the following example of exponential growth, which arises from **compounding interest** in a savings account. Suppose a person invests P dollars in a savings account with an annual interest rate r, compounded annually. The amount of money after 1 year is

$$A(1) = P + rP = P(1+r) \ .$$

The amount of money after 2 years is

$$A(2) = A(1) + rA(1) = P(1+r) + rP(1+r) = P(1+r)^2 \quad .$$

More generally, the amount after t years is

$$A(t) = P(1+r)^t \ .$$

If the money is compounded 2 times per year, the amount of money after half a year is

$$A\left(\frac{1}{2}\right) = P + \left(\frac{r}{2}\right)P = P\left(1 + \left(\frac{r}{2}\right)\right) \ .$$

The amount of money after 1 year is

$$A(1) = A\left(\frac{1}{2}\right) + \left(\frac{r}{2}\right)A\left(\frac{1}{2}\right) = P\left(1+\frac{r}{2}\right) + \frac{r}{2}\left(\left(P(1+\frac{r}{2})\right)\right) = P\left(1+\frac{r}{2}\right)^2.$$

After t years, the amount of money in the account is

$$A(t) = P\left(1+\frac{r}{2}\right)^{2t}.$$

More generally, if the money is compounded n times per year, the amount of money in the account after t years is given by the function

$$A(t) = P\left(1+\frac{r}{n}\right)^{nt}.$$

What happens as $n \to \infty$? To answer this question, we let $m = n/r$ and write

$$\left(1+\frac{r}{n}\right)^{nt} = \left(1+\frac{1}{m}\right)^{mrt},$$

and examine the behavior of $(1+1/m)^m$ as $m \to \infty$, using a table of values (Table 1.5.3).

Table 1.5.3: Values of $\left(1+\frac{1}{m}\right)^m$ as $m \to \infty$

m	10	100	1000	10,000	100,000	1,000,000
$\left(1+\frac{1}{m}\right)^m$	2.5937	2.7048	2.71692	2.71815	2.718268	2.718280

Looking at this table, it appears that $(1+1/m)^m$ is approaching a number between 2.7 and 2.8 as $m \to \infty$. In fact, $(1+1/m)^m$ does approach some number as $m \to \infty$. We call this number e. To six decimal places of accuracy,

$$e \approx 2.718282.$$

Leonhard Euler

The letter e was first used to represent this number by the Swiss mathematician Leonhard Euler during the 1720s. Although Euler did not discover the number, he showed many important connections between e and logarithmic functions. We still use the notation e today to honor Euler's work because it appears in many areas of mathematics and because we can use it in many practical applications.

Returning to our savings account example, we can conclude that if a person puts P dollars in an account at an annual interest rate r, compounded continuously, then $A(t) = Pe^{rt}$. This function may be familiar. Since functions involving base e arise often in applications, we call the function $f(x) = e^x$ the **natural exponential function**. Not only is this function interesting because of the definition of the number e, but also, as discussed next, its graph has an important property.

Since $e > 1$, we know $f(x) = e^x$ is increasing on $(-\infty, \infty)$. In Figure 1.5.3, we show a graph of $f(x) = e^x$ along with a tangent line to the graph of f at $x = 0$. We give a precise definition of tangent line in the next chapter; but, informally, we say a tangent line to a graph of f at $x = a$ is a line that passes through the point $(a, f(a))$ and has the same "slope" as f at that point . The function $f(x) = e^x$ is the only exponential function b^x with tangent line at $x = 0$ that has a slope of 1. As we see later in the text, having this property makes the natural exponential function the most simple exponential function to use in many instances.

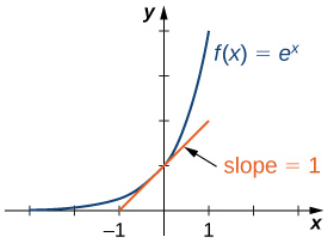


Figure 1.5.3: The graph of $f(x) = e^x$ has a tangent line with slope 1 at $x = 0$.

Example 1.5.3: Compounding Interest

Suppose $500 is invested in an account at an annual interest rate of $r = 5.5\%$, compounded continuously.

a. Let t denote the number of years after the initial investment and $A(t)$ denote the amount of money in the account at time t. Find a formula for $A(t)$.
b. Find the amount of money in the account after 10 years and after 20 years.

Solution

a. If P dollars are invested in an account at an annual interest rate r, compounded continuously, then $A(t) = Pe^{rt}$. Here $P = \$500$ and $r = 0.055$. Therefore, $A(t) = 500e^{0.055t}$.

b. After 10 years, the amount of money in the account is

$$A(10) = 500e^{0.055 \cdot 10} = 500e^{0.55} \approx \$866.63.$$

After 20 years, the amount of money in the account is

$$A(20) = 500e^{0.055 \cdot 20} = 500e^{1.1} \approx \$1,502.08.$$

? Exercise 1.5.3

If $\$750$ is invested in an account at an annual interest rate of 4%, compounded continuously, find a formula for the amount of money in the account after t years. Find the amount of money after 30 years.

Hint

$A(t) = Pe^{rt}$

Answer

$A(t) = 750e^{0.04t}$. After 30 years, there will be approximately $\$2,490.09$

Logarithmic Functions

Using our understanding of exponential functions, we can discuss their inverses, which are the logarithmic functions. These come in handy when we need to consider any phenomenon that varies over a wide range of values, such as the pH scale in chemistry or decibels in sound levels.

The exponential function $f(x) = b^x$ is *one-to-one*, with domain $(-\infty, \infty)$ and range $(0, \infty)$. Therefore, it has an inverse function, called the logarithmic function with base b. For any $b > 0$, $b \neq 1$, the logarithmic function with base b, denoted $\log_b$, has domain $(0, \infty)$ and range $(-\infty, \infty)$,and satisfies

$$\log_b(x) = y$$

if and only if $b^y = x$.

For example,

$$\log_2(8) = 3$$

since $2^3 = 8$,

$$\log_{10}\left(\frac{1}{100}\right) = -2$$

since $10^{-2} = \dfrac{1}{10^2} = \dfrac{1}{100}$,

$$\log_b(1) = 0$$

since $b^0 = 1$ for any base $b > 0$.

Furthermore, since $y = \log_b(x)$ and $y = b^x$ are inverse functions,

$$\log_b(b^x) = x$$

and

$$b^{\log_b(x)} = x.$$

The most commonly used logarithmic function is the function $\log_e$. Since this function uses natural e as its base, it is called the **natural logarithm**. Here we use the notation $\ln(x)$ or $\ln x$ to mean $\log_e(x)$. For example,

$$\begin{aligned} \ln(e) &= \log_e(e) = 1 \\ \ln(e^3) &= \log_e(e^3) = 3 \\ \ln(1) &= \log_e(1) = 0. \end{aligned}$$

Since the functions $f(x) = e^x$ and $g(x) = \ln(x)$ are inverses of each other,

$$\ln(e^x) = x \text{ and } e^{\ln x} = x,$$

and their graphs are symmetric about the line $y = x$ (Figure 1.5.4).

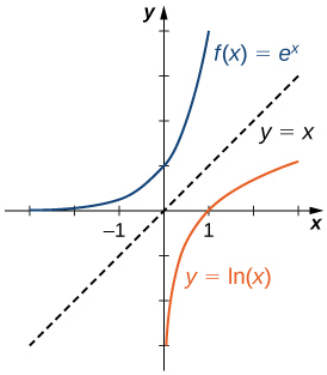


Figure 1.5.4: The functions $y = e^x$ and $y = \ln(x)$ are inverses of each other, so their graphs are symmetric about the line $y = x$.

In general, for any base $b > 0$, $b \neq 1$, the function $g(x) = \log_b(x)$ is symmetric about the line $y = x$ with the function $f(x) = b^x$. Using this fact and the graphs of the exponential functions, we graph functions $\log_b$ for several values of $b > 1$ (Figure 1.5.5).

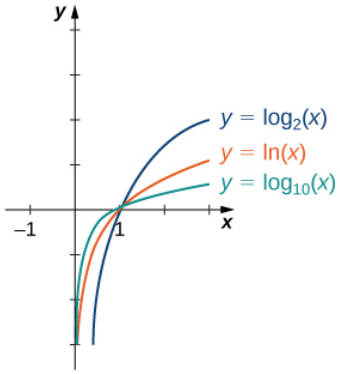


Figure 1.5.5: Graphs of $y = \log_b(x)$ are depicted for $b = 2,\ e,\ 10$.

Before solving some equations involving exponential and logarithmic functions, let's review the basic properties of logarithms.

Properties of Logarithms

If $M,\ N,\ b > 0,\ b \neq 1$, and r is any real number, then

- **Product property**

$$\log_b(MN) = \log_b(M) + \log_b(N) \tag{1.5.1}$$

- **Quotient property**

$$\log_b\left(\frac{M}{N}\right) = \log_b(M) - \log_b(N) \tag{1.5.2}$$

- **Power property**

$$\log_b(M^r) = r\log_b(M) \tag{1.5.3}$$

Example 1.5.4: Solving Equations Involving Exponential Functions

Solve each of the following equations for x.

a. $5^x = 2$
b. $e^x + 6e^{-x} = 5$

Solution

a. Applying the natural logarithm function to both sides of the equation, we have

$$\ln 5^x = \ln 2\,.$$

Using the power property of logarithms,

$$x \ln 5 = \ln 2.$$

Therefore,

$$x = \frac{\ln 2}{\ln 5}.$$

b. Multiplying both sides of the equation by e^x,we arrive at the equation

$$e^{2x} + 6 = 5e^x \, .$$

Rewriting this equation as

$$e^{2x} - 5e^x + 6 = 0 \, ,$$

we can then rewrite it as a quadratic equation in e^x:

$$(e^x)^2 - 5(e^x) + 6 = 0.$$

Now we can solve the quadratic equation. Factoring this equation, we obtain

$$(e^x - 3)(e^x - 2) = 0.$$

Therefore, the solutions satisfy $e^x = 3$ and $e^x = 2$. Taking the natural logarithm of both sides gives us the solutions $x = \ln 3, \ln 2$.

Exercise 1.5.4

Solve

$$e^{2x}/(3 + e^{2x}) = 1/2.$$

Hint

First solve the equation for e^{2x}

Answer

$$x = \frac{\ln 3}{2} \, .$$

Example 1.5.5: Solving Equations Involving Logarithmic Functions

Solve each of the following equations for x.

a. $\ln\left(\frac{1}{x}\right) = 4$
b. $\log_{10} \sqrt{x} + \log_{10} x = 2$
c. $\ln(2x) - 3\ln(x^2) = 0$

Solution

a. By the definition of the natural logarithm function,

$$\ln\left(\frac{1}{x}\right) = 4$$

- if and only if $e^4 = \frac{1}{x}$.

Therefore, the solution is $x = 1/e^4$.

b. Using the product (Equation 1.5.1) and power (Equation 1.5.3) properties of logarithmic functions, rewrite the left-hand side of the equation as

$$\begin{aligned}\log_{10}\sqrt{x}+\log_{10}x &= \log_{10}x\sqrt{x}\\ &= \log_{10}x^{3/2}\\ &= \frac{3}{2}\log_{10}x.\end{aligned}$$

Therefore, the equation can be rewritten as

$$\frac{3}{2}\log_{10}x = 2$$

or

$$\log_{10}x = \frac{4}{3}.$$

The solution is $x = 10^{4/3} = 10\sqrt[3]{10}$.

c. Using the power property (Equation 1.5.3) of logarithmic functions, we can rewrite the equation as $\ln(2x) - \ln(x^6) = 0$.

Using the quotient property (Equation 1.5.2), this becomes

$$\ln\left(\frac{2}{x^5}\right) = 0$$

Therefore, $2/x^5 = 1$, which implies $x = \sqrt[5]{2}$. We should then check for any extraneous solutions.

Exercise 1.5.5

Solve $\ln(x^3) - 4\ln(x) = 1$.

Hint

First use the power property, then use the product property of logarithms.

Answer

$$x = \frac{1}{e}$$

When evaluating a logarithmic function with a calculator, you may have noticed that the only options are $\log_{10}$ or log, called the **common logarithm**, or ln, which is the natural logarithm. However, exponential functions and logarithm functions can be expressed in terms of any desired base b. If you need to use a calculator to evaluate an expression with a different base, you can apply the change-of-base formulas first. Using this change of base, we typically write a given exponential or logarithmic function in terms of the natural exponential and natural logarithmic functions.

Rule: Change-of-Base Formulas

Let $a > 0$, $b > 0$, and $a \neq 1$, $b \neq 1$.

1. $a^x = b^{x\log_b a}$ for any real number x.

   If $b = e$, this equation reduces to $a^x = e^{x\log_e a} = e^{x\ln a}$.

2. $\log_a x = \dfrac{\log_b x}{\log_b a}$ for any real number $x > 0$.

   If $b = e$, this equation reduces to $\log_a x = \dfrac{\ln x}{\ln a}$.

Proof

For the first change-of-base formula, we begin by making use of the power property of logarithmic functions. We know that for any base $b > 0,\ b \neq 1$, $\log_b(a^x) = x \log_b a$. Therefore,

$$b^{\log_b(a^x)} = b^{x \log_b a}.$$

In addition, we know that b^x and $\log_b(x)$ are inverse functions. Therefore,

$$b^{\log_b(a^x)} = a^x.$$

Combining these last two equalities, we conclude that $a^x = b^{x \log_b a}$.

To prove the second property, we show that

$$(\log_b a) \cdot (\log_a x) = \log_b x.$$

Let $u = \log_b a, v = \log_a x$, and $w = \log_b x$. We will show that $u \cdot v = w$. By the definition of logarithmic functions, we know that $b^u = a,\ a^v = x$, and $b^w = x$. From the previous equations, we see that

$$b^{uv} = (b^u)^v = a^v = x = b^w.$$

Therefore, $b^{uv} = b^w$. Since exponential functions are one-to-one, we can conclude that $u \cdot v = w$.

□

Example 1.5.6: Changing Bases

Use a calculating utility to evaluate $\log_3 7$ with the change-of-base formula presented earlier.

Solution

Use the second equation with $a = 3$ and $b = e$: $\log_3 7 = \dfrac{\ln 7}{\ln 3} \approx 1.77124.$

Exercise 1.5.6

Use the change-of-base formula and a calculating utility to evaluate $\log_4 6$.

Hint

Use the change of base to rewrite this expression in terms of expressions involving the natural logarithm function.

Answer

$$\log_4 6 = \frac{\ln 6}{\ln 4} \approx 1.29248$$

Example 1.5.7: The Richter Scale for Earthquakes

In 1935, Charles Richter developed a scale (now known as the Richter scale) to measure the magnitude of an **earthquake**. The scale is a base-10 logarithmic scale, and it can be described as follows: Consider one earthquake with magnitude R_1 on the Richter scale and a second earthquake with magnitude R_2 on the Richter scale. Suppose $R_1 > R_2$, which means the earthquake of magnitude R_1 is stronger, but how much stronger is it than the other earthquake?



Figure 1.5.6: (credit: modification of work by Robb Hannawacker, NPS)

A way of measuring the intensity of an earthquake is by using a seismograph to measure the amplitude of the earthquake waves. If A_1 is the amplitude measured for the first earthquake and A_2 is the amplitude measured for the second earthquake, then the amplitudes and magnitudes of the two earthquakes satisfy the following equation:

$$R_1 - R_2 = \log_{10}\left(\frac{A1}{A2}\right).$$

Consider an earthquake that measures 8 on the Richter scale and an earthquake that measures 7 on the Richter scale. Then,

$$8 - 7 = \log_{10}\left(\frac{A1}{A2}\right).$$

Therefore,

$$\log_{10}\left(\frac{A1}{A2}\right) = 1,$$

which implies $A_1/A_2 = 10$ or $A_1 = 10A_2$. Since A_1 is 10 times the size of A_2, we say that the first earthquake is 10 times as intense as the second earthquake. On the other hand, if one earthquake measures 8 on the Richter scale and another measures 6, then the relative intensity of the two earthquakes satisfies the equation

$$\log_{10}\left(\frac{A1}{A2}\right) = 8 - 6 = 2.$$

Therefore, $A_1 = 100A_2$.That is, the first earthquake is 100 times more intense than the second earthquake.

How can we use logarithmic functions to compare the relative severity of the magnitude 9 earthquake in Japan in 2011 with the magnitude 7.3 earthquake in Haiti in 2010?

Solution

To compare the Japan and Haiti earthquakes, we can use an equation presented earlier:

$$9 - 7.3 = \log_{10}\left(\frac{A1}{A2}\right).$$

Therefore, $A_1/A_2 = 10^{1.7}$, and we conclude that the earthquake in Japan was approximately 50 times more intense than the earthquake in Haiti.

? Exercise 1.5.7

Compare the relative severity of a magnitude 8.4 earthquake with a magnitude 7.4 earthquake.

Hint

$R_1 - R_2 = \log_{10}(A1/A2)$.

Answer

The magnitude 8.4 earthquake is roughly 10 times as severe as the magnitude 7.4 earthquake.

This page titled 1.5: Exponential and Logarithmic Functions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **1.5: Exponential and Logarithmic Functions** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

CHAPTER OVERVIEW

2: Limits and Continuity

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

2.1: Limit of a Function and Limit Laws
2.2: The Precise Definition of a Limit
2.3: One-Sided Limits
2.4: Continuity
2.5: Limits Involving Infinity; Asymptotes of Graphs

2: Limits and Continuity is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

2.1: Limit of a Function and Limit Laws

Learning Objectives

- Using correct notation, describe the limit of a function.
- Use a table of values to estimate the limit of a function or to identify when the limit does not exist.
- Use a graph to estimate the limit of a function or to identify when the limit does not exist.
- Recognize the basic limit laws.
- Use the limit laws to evaluate the limit of a function.
- Evaluate the limit of a function by factoring.
- Use the limit laws to evaluate the limit of a polynomial or rational function.
- Evaluate the limit of a function by factoring or by using conjugates.
- Evaluate the limit of a function by using the squeeze theorem.

The concept of a limit or limiting process, essential to the understanding of calculus, has been around for thousands of years. In fact, early mathematicians used a limiting process to obtain better and better approximations of areas of circles. Yet, the formal definition of a limit—as we know and understand it today—did not appear until the late 19th century. We therefore begin our quest to understand limits, as our mathematical ancestors did, by using an intuitive approach. At the end of this chapter, armed with a conceptual understanding of limits, we examine the formal definition of a limit.

We begin our exploration of limits by taking a look at the graphs of the functions

- $f(x)=\dfrac{x^2-4}{x-2}$,
- $g(x)=\dfrac{|x-2|}{x-2}$, and
- $h(x)=\dfrac{1}{(x-2)^2}$,

which are shown in Figure 2.1.1. In particular, let's focus our attention on the behavior of each graph at and around $x=2$.

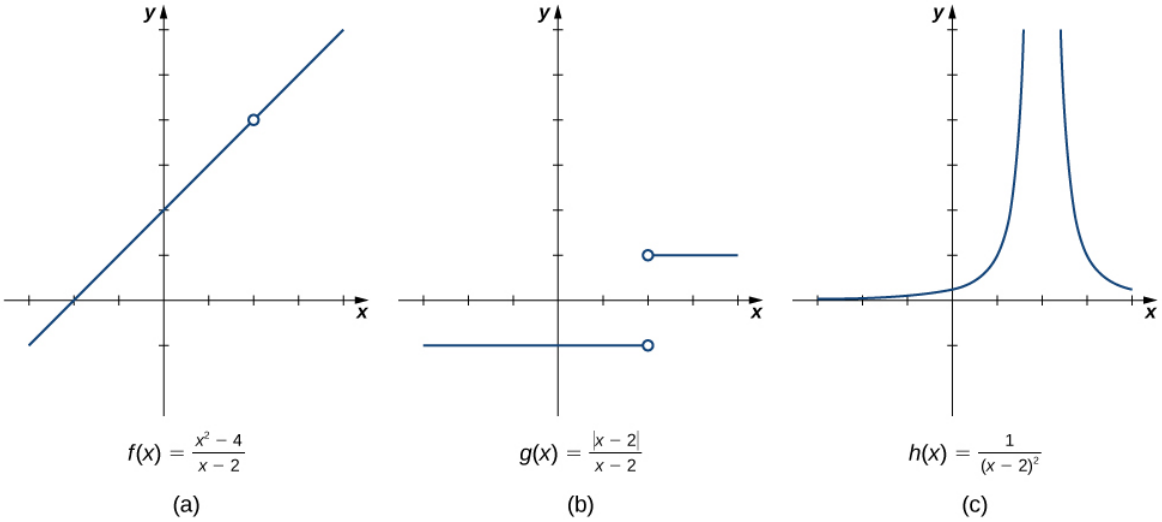


Figure 2.1.1: These graphs show the behavior of three different functions around $x=2$.

Each of the three functions is undefined at $x=2$, but if we make this statement and no other, we give a very incomplete picture of how each function behaves in the vicinity of $x=2$. To express the behavior of each graph in the vicinity of 2 more completely, we need to introduce the concept of a limit.

Intuitive Definition of a Limit

Let's first take a closer look at how the function $f(x)=(x^2-4)/(x-2)$ behaves around $x=2$ in Figure 2.1.1. As the values of x approach 2 from either side of 2, the values of $y=f(x)$ approach 4. Mathematically, we say that the limit of $f(x)$ as x

approaches 2 is 4. Symbolically, we express this limit as

$$\lim_{x\to 2} f(x) = 4$$

From this very brief informal look at one limit, let's start to develop an *intuitive definition of the limit*. We can think of the limit of a function at a number a as being the one real number L that the functional values approach as the x-values approach a, provided such a real number L exists. Stated more carefully, we have the following definition:

Definition (Intuitive): Limit

Let $f(x)$ be a function defined at all values in an open interval containing a, with the possible exception of a itself, and let L be a real number. If all values of the function $f(x)$ approach the real number L as the values of $x(\neq a)$ approach the number a, then we say that the limit of $f(x)$ as x approaches a is L. (More succinct, as x gets closer to a, $f(x)$ gets closer and stays close to L.) Symbolically, we express this idea as

$$\lim_{x\to a} f(x) = L. \tag{2.1.1}$$

We can estimate limits by constructing tables of functional values and by looking at their graphs. This process is described in the following Problem-Solving Strategy.

Problem-Solving Strategy: Evaluating a Limit Using a Table of Functional Values

1. To evaluate $\lim_{x\to a} f(x)$, we begin by completing a table of functional values. We should choose two sets of x-values—one set of values approaching a and less than a, and another set of values approaching a and greater than a. Table 2.1.1 demonstrates what your tables might look like.

Table 2.1.1

x	$f(x)$	x	$f(x)$
$a-0.1$	$f(a-0.1)$	$a+0.1$	$f(a+0.1)$
$a-0.01$	$f(a-0.01)$	$a+0.01$	$f(a+0.01)$
$a-0.001$	$f(a-0.001)$	$a+0.001$	$f(a+0.001)$
$a-0.0001$	$f(a-0.0001)$	$a+0.0001$	$f(a+0.0001)$
Use additional values as necessary.		Use additional values as necessary.	

2. Next, let's look at the values in each of the $f(x)$ columns and determine whether the values seem to be approaching a single value as we move down each column. In our columns, we look at the sequence $f(a-0.1)$, $f(a-0.01)$, $f(a-0.001)$, $f(a-0.0001)$, and so on, and $f(a+0.1)$, $f(a+0.01)$, $f(a+0.001)$, $f(a+0.0001)$, and so on. (Note: Although we have chosen the x-values $a\pm 0.1$, $a\pm 0.01$, $a\pm 0.001$, $a\pm 0.0001$, and so forth, and these values will probably work nearly every time, on very rare occasions we may need to modify our choices.)

3. If both columns approach a common y-value L, we state $\lim_{x\to a} f(x) = L$. We can use the following strategy to confirm the result obtained from the table or as an alternative method for estimating a limit.

4. Using a graphing calculator or computer software that allows us graph functions, we can plot the function $f(x)$, making sure the functional values of $f(x)$ for x-values near a are in our window. We can use the trace feature to move along the graph of the function and watch the y-value readout as the x-values approach a. If the y-values approach L as our x-values approach a from both directions, then $\lim_{x\to a} f(x) = L$. We may need to zoom in on our graph and repeat this process several times.

We apply this Problem-Solving Strategy to compute a limit in Examples 2.1.1*A* and 2.1.1*B*.

Example 2.1.1A: Evaluating a Limit Using a Table of Functional Values

Evaluate $\lim_{x\to 0}\dfrac{\sin x}{x}$ using a table of functional values.

Solution

We have calculated the values of $f(x)=\dfrac{\sin x}{x}$ for the values of x listed in Table 2.1.2.

Table 2.1.2

x	$\frac{\sin x}{x}$	x	$\frac{\sin x}{x}$
-0.1	0.998334166468	0.1	0.998334166468
-0.01	0.999983333417	0.01	0.999983333417
-0.001	0.999999833333	0.001	0.999999833333
-0.0001	0.999999998333	0.0001	0.999999998333

Note: The values in this table were obtained using a calculator and using all the places given in the calculator output.

As we read down each $\dfrac{\sin x}{x}$ column, we see that the values in each column appear to be approaching one. Thus, it is fairly reasonable to conclude that $\lim_{x\to 0}\dfrac{\sin x}{x}=1$. A calculator-or computer-generated graph of $f(x)=\dfrac{\sin x}{x}$ would be similar to that shown in Figure 2.1.2, and it confirms our estimate.

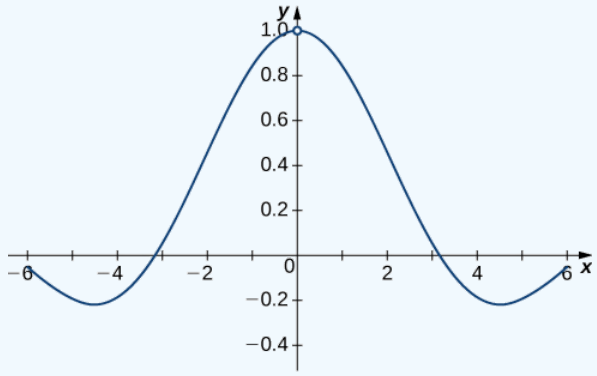


Figure 2.1.2: The graph of $f(x)=(\sin x)/x$ confirms the estimate from Table 2.1.2.

Example 2.1.1B: Evaluating a Limit Using a Table of Functional Values

Evaluate $\lim_{x\to 4}\dfrac{\sqrt{x}-2}{x-4}$ using a table of functional values.

Solution

As before, we use a table—in this case, Table 2.1.3—to list the values of the function for the given values of x.

Table 2.1.3

x	$\frac{\sqrt{x}-2}{x-4}$	x	$\frac{\sqrt{x}-2}{x-4}$
3.9	0.251582341869	4.1	0.248456731317
3.99	0.25015644562	4.01	0.24984394501
3.999	0.250015627	4.001	0.249984377
3.9999	0.250001563	4.0001	0.249998438
3.99999	0.25000016	4.00001	0.24999984

After inspecting this table, we see that the functional values less than 4 appear to be decreasing toward 0.25 whereas the functional values greater than 4 appear to be increasing toward 0.25. We conclude that $\lim\limits_{x\to 4}\dfrac{\sqrt{x}-2}{x-4}=0.25$. We confirm this estimate using the graph of $f(x)=\dfrac{\sqrt{x}-2}{x-4}$ shown in Figure 2.1.3.

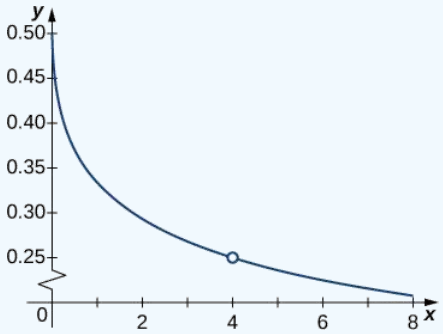


Figure 2.1.3: The graph of $\frac{\sqrt{x}-2}{x-4}$ confirms the estimate from Table 2.1.3.

? Exercise 2.1.1

Estimate $\lim\limits_{x\to 1}\dfrac{\frac{1}{x}-1}{x-1}$ using a table of functional values. Use a graph to confirm your estimate.

Hint

Use 0.9, 0.99, 0.999, 0.9999, 0.99999 and 1.1, 1.01, 1.001, 1.0001, 1.00001 as your table values.

Answer

$$\lim_{x\to 1}\frac{\frac{1}{x}-1}{x-1}=-1$$

At this point, we see from Examples 2.1.1A and 2.1.1b that it may be just as easy, if not easier, to estimate a limit of a function by inspecting its graph as it is to estimate the limit by using a table of functional values. In Example 2.1.2, we evaluate a limit exclusively by looking at a graph rather than by using a table of functional values.

✓ Example 2.1.2: Evaluating a Limit Using a Graph

For $g(x)$ shown in Figure 2.1.4, evaluate $\lim\limits_{x\to -1} g(x)$.

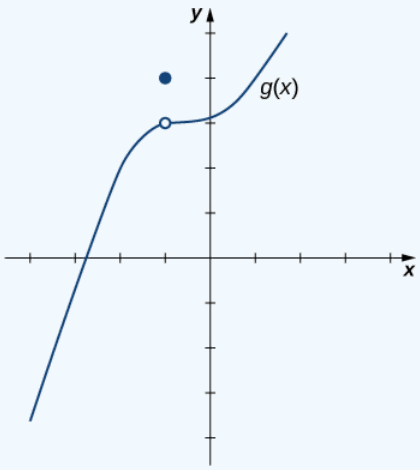


Figure 2.1.4: The graph of $g(x)$ includes one value not on a smooth curve.

Solution:

Despite the fact that $g(-1)=4$, as the x-values approach -1 from either side, the $g(x)$ values approach 3. Therefore, $\lim\limits_{x\to -1} g(x)=3$. Note that we can determine this limit without even knowing the algebraic expression of the function.

Based on Example 2.1.2, we make the following observation: It is possible for the limit of a function to exist at a point, and for the function to be defined at this point, but the limit of the function and the value of the function at the point may be different.

Exercise 2.1.2

Use the graph of $h(x)$ in Figure 2.1.5 to evaluate $\lim\limits_{x\to 2} h(x)$, if possible.

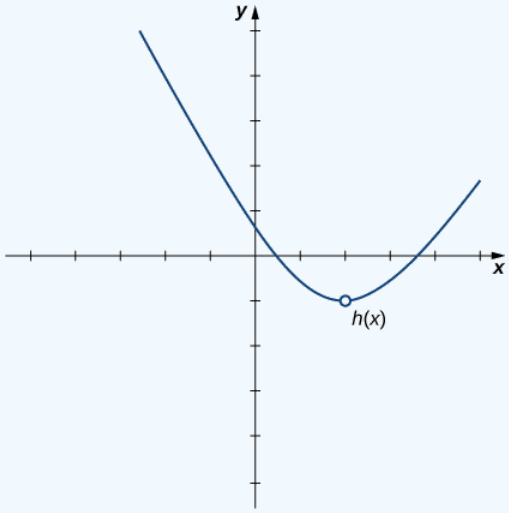


Figure 2.1.5:

Hint

What y-value does the function approach as the x-values approach 2?

Solution

$\lim\limits_{x\to 2} h(x)=-1$.

Looking at a table of functional values or looking at the graph of a function provides us with useful insight into the value of the limit of a function at a given point. However, these techniques rely too much on guesswork. We eventually need to develop alternative methods of evaluating limits. These new methods are more algebraic in nature and we explore them in the next section; however, at this point we introduce two special limits that are foundational to the techniques to come.

Two Important Limits

Let a be a real number and c be a constant.

i. $\lim\limits_{x\to a} x=a$
ii. $\lim\limits_{x\to a} c=c$

We can make the following observations about these two limits.

i. For the first limit, observe that as x approaches a, so does $f(x)$, because $f(x)=x$. Consequently, $\lim\limits_{x\to a} x=a$.
ii. For the second limit, consider Table 2.1.4.

Table 2.1.4

x	$f(x)=c$	x	$f(x)=c$

x	$f(x)=c$	x	$f(x)=c$
$a-0.1$	c	$a+0.1$	c
$a-0.01$	c	$a+0.01$	c
$a-0.001$	c	$a+0.001$	c
$a-0.0001$	c	$a+0.0001$	c

Observe that for all values of x (regardless of whether they are approaching a), the values $f(x)$ remain constant at c. We have no choice but to conclude $\lim\limits_{x\to a} c = c$.

The Existence of a Limit

As we consider the limit in the next example, keep in mind that for the limit of a function to exist at a point, the functional values must approach a single real-number value at that point. If the functional values do not approach a single value, then the limit does not exist.

Example 2.1.3: Evaluating a Limit That Fails to Exist

Evaluate $\lim\limits_{x\to 0} \sin(1/x)$ using a table of values.

Solution

Table 2.1.5 lists values for the function $\sin(1/x)$ for the given values of x.

Table 2.1.5

x	$\sin(1/x)$	x	$\sin(1/x)$
-0.1	0.544021110889	0.1	−0.544021110889
-0.01	0.50636564111	0.01	−0.50636564111
-0.001	−0.8268795405312	0.001	0.8268795405312
-0.0001	0.305614388888	0.0001	−0.305614388888
-0.00001	−0.035748797987	0.00001	0.035748797987
-0.000001	0.349993504187	0.000001	−0.349993504187

After examining the table of functional values, we can see that the y-values do not seem to approach any one single value. It appears the limit does not exist. Before drawing this conclusion, let's take a more systematic approach. Take the following sequence of x-values approaching 0:

$$\frac{2}{\pi}, \frac{2}{3\pi}, \frac{2}{5\pi}, \frac{2}{7\pi}, \frac{2}{9\pi}, \frac{2}{11\pi}, \ldots.$$

The corresponding y-values are

$$1, -1, 1, -1, 1, -1, \ldots.$$

At this point we can indeed conclude that $\lim\limits_{x\to 0} \sin(1/x)$ does not exist. (Mathematicians frequently abbreviate "does not exist" as DNE. Thus, we would write $\lim\limits_{x\to 0} \sin(1/x)$ DNE.) The graph of $f(x)=\sin(1/x)$ is shown in Figure 2.1.6 and it gives a clearer picture of the behavior of $\sin(1/x)$ as x approaches 0. You can see that $\sin(1/x)$ oscillates ever more wildly between -1 and 1 as x approaches 0.

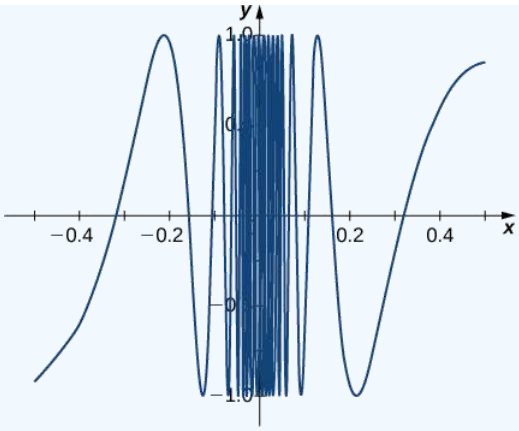


Figure 2.1.6: The graph of $f(x) = \sin(1/x)$ oscillates rapidly between -1 and 1 as x approaches 0.

? Exercise 2.1.3

Use a table of functional values to evaluate $\lim\limits_{x \to 2} \dfrac{|x^2 - 4|}{x - 2}$, if possible.

Hint

Use x-values 1.9, 1.99, 1.999, 1.9999, 1.99999 and 2.1, 2.01, 2.001, 2.0001, 2.00001 in your table.

Answer

$\lim\limits_{x \to 2} \dfrac{|x^2 - 4|}{x - 2}$ does not exist.

Previously, we evaluated limits by looking at graphs or by constructing a table of values. Now, we establish laws for calculating limits and learn how to apply these laws.

Evaluating Limits with the Limit Laws

The first two limit laws were stated previously and we repeat them here. These basic results, together with the other limit laws, allow us to evaluate limits of many algebraic functions.

Basic Limit Results

For any real number a and any constant c,

I. $\lim\limits_{x \to a} x = a$
II. $\lim\limits_{x \to a} c = c$

✓ Example 2.1.4: Evaluating a Basic Limit

Evaluate each of the following limits using "Basic Limit Results."

a. $\lim\limits_{x \to 2} x$
b. $\lim\limits_{x \to 2} 5$

Solution

a. The limit of x as x approaches a is a: $\lim\limits_{x \to 2} x = 2$.
b. The limit of a constant is that constant: $\lim\limits_{x \to 2} 5 = 5$.

We now take a look at the **limit laws**, the individual properties of limits. The proofs that these laws hold are omitted here.

Limit Laws

Let $f(x)$ and $g(x)$ be defined for all $x \neq a$ over some open interval containing a. Assume that L and M are real numbers such that $\lim\limits_{x \to a} f(x) = L$ and $\lim\limits_{x \to a} g(x) = M$. Let c be a constant. Then, each of the following statements holds:

- **Sum law for limits**:

$$\lim_{x \to a}(f(x) + g(x)) = \lim_{x \to a} f(x) + \lim_{x \to a} g(x) = L + M$$

- **Difference law for limits**:

$$\lim_{x \to a}(f(x) - g(x)) = \lim_{x \to a} f(x) - \lim_{x \to a} g(x) = L - M$$

- **Constant multiple law for limits**:

$$\lim_{x \to a} cf(x) = c \cdot \lim_{x \to a} f(x) = cL$$

- **Product law for limits**:

$$\lim_{x \to a}(f(x) \cdot g(x)) = \lim_{x \to a} f(x) \cdot \lim_{x \to a} g(x) = L \cdot M$$

- **Quotient law for limits**:

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \frac{\lim\limits_{x \to a} f(x)}{\lim\limits_{x \to a} g(x)} = \frac{L}{M}$$

for $M \neq 0$.

- **Power law for limits**:

$$\lim_{x \to a} \big(f(x)\big)^n = \big(\lim_{x \to a} f(x)\big)^n = L^n$$

for every positive integer n.

- **Root law for limits**:

$$\lim_{x \to a} \sqrt[n]{f(x)} = \sqrt[n]{\lim_{x \to a} f(x)} = \sqrt[n]{L}$$

for all L if n is odd and for $L \geq 0$ if n is even.

We now practice applying these limit laws to evaluate a limit.

Example 2.1.5A: Evaluating a Limit Using Limit Laws

Use the limit laws to evaluate

$$\lim_{x \to -3}(4x + 2).$$

Solution

Let's apply the limit laws one step at a time to be sure we understand how they work. We need to keep in mind the requirement that, at each application of a limit law, the new limits must exist for the limit law to be applied.

$$\begin{aligned} \lim_{x \to -3}(4x + 2) &= \lim_{x \to -3} 4x + \lim_{x \to -3} 2 && \text{Apply the sum law.} \\ &= 4 \cdot \lim_{x \to -3} x + \lim_{x \to -3} 2 && \text{Apply the constant multiple law.} \\ &= 4 \cdot (-3) + 2 = -10. && \text{Apply the basic limit results and simplify.} \end{aligned}$$

Example 2.1.5B: Using Limit Laws Repeatedly

Use the limit laws to evaluate

$$\lim_{x\to 2}\frac{2x^2-3x+1}{x^3+4}.$$

Solution

To find this limit, we need to apply the limit laws several times. Again, we need to keep in mind that as we rewrite the limit in terms of other limits, each new limit must exist for the limit law to be applied.

$$\lim_{x\to 2}\frac{2x^2-3x+1}{x^3+4}=\frac{\lim_{x\to 2}(2x^2-3x+1)}{\lim_{x\to 2}(x^3+4)}$$ Apply the quotient law, make sure that $(2)^3+4\neq 0$.

$$=\frac{2\cdot\lim_{x\to 2}x^2-3\cdot\lim_{x\to 2}x+\lim_{x\to 2}1}{\lim_{x\to 2}x^3+\lim_{x\to 2}4}$$ Apply the sum law and constant multiple law.

$$=\frac{2\cdot\left(\lim_{x\to 2}x\right)^2-3\cdot\lim_{x\to 2}x+\lim_{x\to 2}1}{\left(\lim_{x\to 2}x\right)^3+\lim_{x\to 2}4}$$ Apply the power law.

$$=\frac{2(4)-3(2)+1}{(2)^3+4}=\frac{1}{4}.$$ Apply the basic limit laws and simplify.

Exercise 2.1.5

Use the limit laws to evaluate $\lim_{x\to 6}(2x-1)\sqrt{x+4}$. In each step, indicate the limit law applied.

Hint

Begin by applying the product law.

Answer

$11\sqrt{10}$

Limits of Polynomial and Rational Functions

By now you have probably noticed that, in each of the previous examples, it has been the case that $\lim_{x\to a}f(x)=f(a)$. This is not always true, but it does hold for all polynomials for any choice of a and for all rational functions at all values of a for which the rational function is defined.

Limits of Polynomial and Rational Functions

Let $p(x)$ and $q(x)$ be polynomial functions. Let a be a real number. Then,

$$\lim_{x\to a}p(x)=p(a)$$

$$\lim_{x\to a}\frac{p(x)}{q(x)}=\frac{p(a)}{q(a)}$$

when $q(a)\neq 0$.

To see that this theorem holds, consider the polynomial

$$p(x)=c_nx^n+c_{n-1}x^{n-1}+\cdots+c_1x+c_0.$$

By applying the sum, constant multiple, and power laws, we end up with

$$\begin{aligned}\lim_{x\to a} p(x) &= \lim_{x\to a}(c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0)\\ &= c_n\left(\lim_{x\to a} x\right)^n + c_{n-1}\left(\lim_{x\to a} x\right)^{n-1} + \cdots + c_1\left(\lim_{x\to a} x\right) + \lim_{x\to a} c_0\\ &= c_n a^n + c_{n-1}a^{n-1} + \cdots + c_1 a + c_0\\ &= p(a)\end{aligned}$$

It now follows from the quotient law that if $p(x)$ and $q(x)$ are polynomials for which $q(a) \neq 0$,

then

$$\lim_{x\to a}\frac{p(x)}{q(x)} = \frac{p(a)}{q(a)}.$$

Example 2.1.6: Evaluating a Limit of a Rational Function

Evaluate the $\lim_{x\to 3}\dfrac{2x^2-3x+1}{5x+4}$.

Solution

Since 3 is in the domain of the rational function $f(x) = \dfrac{2x^2-3x+1}{5x+4}$, we can calculate the limit by substituting 3 for x into the function. Thus,

$$\lim_{x\to 3}\frac{2x^2-3x+1}{5x+4} = \frac{10}{19}.$$

Exercise 2.1.6

Evaluate $\lim_{x\to -2}(3x^3 - 2x + 7)$.

Hint

Use LIMITS OF POLYNOMIAL AND RATIONAL FUNCTIONS as reference

Answer

−13

Additional Limit Evaluation Techniques

As we have seen, we may evaluate easily the limits of polynomials and limits of some (but not all) rational functions by direct substitution. However, as we saw in the introductory section on limits, it is certainly possible for $\lim_{x\to a} f(x)$ to exist when $f(a)$ is undefined. The following observation allows us to evaluate many limits of this type:

If for all $x \neq a$, $f(x) = g(x)$ over some open interval containing a, then

$$\lim_{x\to a} f(x) = \lim_{x\to a} g(x).$$

To understand this idea better, consider the limit $\lim_{x\to 1}\dfrac{x^2-1}{x-1}$.

The function

$$f(x) = \frac{x^2-1}{x-1} = \frac{(x-1)(x+1)}{x-1}$$

and the function $g(x) = x + 1$ are identical for all values of $x \neq 1$. The graphs of these two functions are shown in Figure 2.1.1.

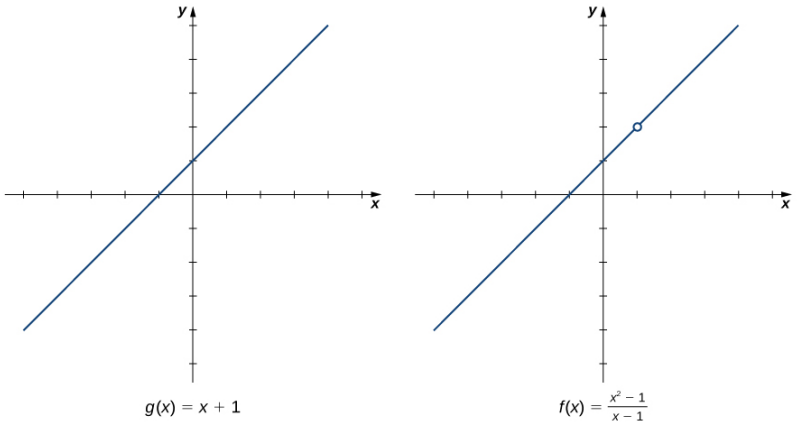


Figure 2.1.1: The graphs of $f(x)$ and $g(x)$ are identical for all $x \neq 1$. Their limits at 1 are equal.

We see that

$$\lim_{x\to 1}\frac{x^2-1}{x-1}=\lim_{x\to 1}\frac{(x-1)(x+1)}{x-1}=\lim_{x\to 1}(x+1)=2.$$

The limit has the form $\lim_{x\to a} f(x)/g(x)$, where $\lim_{x\to a} f(x)=0$ and $\lim_{x\to a} g(x)=0$. (In this case, we say that $f(x)/g(x)$ has the indeterminate form $0/0$.) The following Problem-Solving Strategy provides a general outline for evaluating limits of this type.

Problem-Solving Strategy: Calculating a Limit When $f(x)/g(x)$ has the Indeterminate Form $0/0$

1. First, we need to make sure that our function has the appropriate form and cannot be evaluated immediately using the limit laws.
2. We then need to find a function that is equal to $h(x)=f(x)/g(x)$ for all $x \neq a$ over some interval containing a. To do this, we may need to try one or more of the following steps:
   a. If $f(x)$ and $g(x)$ are polynomials, we should factor each function and cancel out any common factors.
   b. If the numerator or denominator contains a difference involving a square root, we should try multiplying the numerator and denominator by the conjugate of the expression involving the square root.
   c. If $f(x)/g(x)$ is a complex fraction, we begin by simplifying it.
3. Last, we apply the limit laws.

The next examples demonstrate the use of this Problem-Solving Strategy. Example 2.1.4 illustrates the factor-and-cancel technique; Example 2.1.5 shows multiplying by a conjugate. In Example 2.1.6, we look at simplifying a complex fraction.

Example 2.1.7: Evaluating a Limit by Factoring and Canceling

Evaluate $\lim_{x\to 3}\dfrac{x^2-3x}{2x^2-5x-3}$.

Solution

Step 1. The function $f(x)=\dfrac{x^2-3x}{2x^2-5x-3}$ is undefined for $x=3$. In fact, if we substitute 3 into the function we get $0/0$, which is undefined. Factoring and canceling is a good strategy:

$$\lim_{x\to 3}\frac{x^2-3x}{2x^2-5x-3}=\lim_{x\to 3}\frac{x(x-3)}{(x-3)(2x+1)}$$

Step 2. For all $x \neq 3$, $\dfrac{x^2-3x}{2x^2-5x-3}=\dfrac{x}{2x+1}$. Therefore,

$$\lim_{x\to 3}\frac{x(x-3)}{(x-3)(2x+1)}=\lim_{x\to 3}\frac{x}{2x+1}.$$

Step 3. Evaluate using the limit laws:

$$\lim_{x\to 3}\frac{x}{2x+1}=\frac{3}{7}.$$

Exercise 2.1.7

Evaluate $\lim_{x\to -3}\dfrac{x^2+4x+3}{x^2-9}$.

Hint

Follow the steps in the Problem-Solving Strategy

Answer

$\dfrac{1}{3}$

Example 2.1.8: Evaluating a Limit by Multiplying by a Conjugate

Evaluate $\lim_{x\to -1}\dfrac{\sqrt{x+2}-1}{x+1}$.

Solution

Step 1. $\dfrac{\sqrt{x+2}-1}{x+1}$ has the form $0/0$ at -1. Let's begin by multiplying by $\sqrt{x+2}+1$, the conjugate of $\sqrt{x+2}-1$, on the numerator and denominator:

$$\lim_{x\to -1}\frac{\sqrt{x+2}-1}{x+1}=\lim_{x\to -1}\frac{\sqrt{x+2}-1}{x+1}\cdot\frac{\sqrt{x+2}+1}{\sqrt{x+2}+1}.$$

Step 2. We then multiply out the numerator. We don't multiply out the denominator because we are hoping that the $(x+1)$ in the denominator cancels out in the end:

$$=\lim_{x\to -1}\frac{x+1}{(x+1)(\sqrt{x+2}+1)}.$$

Step 3. Then we cancel:

$$=\lim_{x\to -1}\frac{1}{\sqrt{x+2}+1}.$$

Step 4. Last, we apply the limit laws:

$$\lim_{x\to -1}\frac{1}{\sqrt{x+2}+1}=\frac{1}{2}.$$

Exercise 2.1.8

Evaluate $\lim_{x\to 5}\dfrac{\sqrt{x-1}-2}{x-5}$.

Hint

Follow the steps in the Problem-Solving Strategy

Answer

$\dfrac{1}{4}$

Example 2.1.9: Evaluating a Limit by Simplifying a Complex Fraction

Evaluate $\lim\limits_{x\to 1}\dfrac{\dfrac{1}{x+1}-\dfrac{1}{2}}{x-1}$.

Solution

Step 1. $\dfrac{\dfrac{1}{x+1}-\dfrac{1}{2}}{x-1}$ has the form $0/0$ at 1. We simplify the algebraic fraction by multiplying by $2(x+1)/2(x+1)$:

$$\lim_{x\to 1}\frac{\dfrac{1}{x+1}-\dfrac{1}{2}}{x-1}=\lim_{x\to 1}\frac{\dfrac{1}{x+1}-\dfrac{1}{2}}{x-1}\cdot\frac{2(x+1)}{2(x+1)}.$$

Step 2. Next, we multiply through the numerators. Do not multiply the denominators because we want to be able to cancel the factor $(x-1)$:

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

Step 3. Then, we simplify the numerator:

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

Step 4. Now we factor out -1 from the numerator:

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

Step 5. Then, we cancel the common factors of $(x-1)$:

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

Step 6. Last, we evaluate using the limit laws:

$$\lim_{x\to 1}\frac{-1}{2(x+1)}=-\frac{1}{4}.$$

Exercise 2.1.9

Evaluate $\lim\limits_{x\to -3}\dfrac{\dfrac{1}{x+2}+1}{x+3}$.

Hint

Follow the steps in the Problem-Solving Strategy

Answer

-1

Example 2.1.7 does not fall neatly into any of the patterns established in the previous examples. However, with a little creativity, we can still use these same techniques.

Example 2.1.10: Evaluating a Limit When the Limit Laws Do Not Apply

Evaluate $\lim\limits_{x\to 0}\left(\dfrac{1}{x}+\dfrac{5}{x(x-5)}\right)$.

Solution:

Both $1/x$ and $5/x(x-5)$ fail to have a limit at zero. Since neither of the two functions has a limit at zero, we cannot apply the sum law for limits; we must use a different strategy. In this case, we find the limit by performing addition and then applying one of our previous strategies. Observe that

$$\frac{1}{x}+\frac{5}{x(x-5)}=\frac{x-5+5}{x(x-5)}=\frac{x}{x(x-5)}.$$

Thus,

$$\lim_{x\to 0}\left(\frac{1}{x}+\frac{5}{x(x-5)}\right)=\lim_{x\to 0}\frac{x}{x(x-5)}=\lim_{x\to 0}\frac{1}{x-5}=-\frac{1}{5}.$$

Exercise 2.1.10

Evaluate $\lim\limits_{x\to 3}\left(\dfrac{1}{x-3}-\dfrac{4}{x^2-2x-3}\right)$.

Hint

Use the same technique as Example 2.1.7. Don't forget to factor x^2-2x-3 before getting a common denominator.

Answer

$\dfrac{1}{4}$

Let's now revisit one-sided limits. Simple modifications in the limit laws allow us to apply them to one-sided limits. For example, to apply the limit laws to a limit of the form $\lim\limits_{x\to a^-}h(x)$, we require the function $h(x)$ to be defined over an open interval of the form (b,a); for a limit of the form $\lim\limits_{x\to a^+}h(x)$, we require the function $h(x)$ to be defined over an open interval of the form (a,c). Example 2.1.8A illustrates this point.

Example 2.1.11A: Evaluating a One-Sided Limit Using the Limit Laws

Evaluate each of the following limits, if possible.

a. $\lim\limits_{x\to 3^-}\sqrt{x-3}$

b. $\lim\limits_{x\to 3^+}\sqrt{x-3}$

Solution

Figure 2.1.2 illustrates the function $f(x)=\sqrt{x-3}$ and aids in our understanding of these limits.

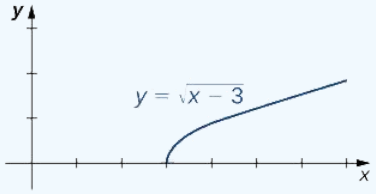


Figure 2.1.2: The graph shows the function $f(x)=\sqrt{x-3}$.

a. The function $f(x)=\sqrt{x-3}$ is defined over the interval $[3,+\infty)$. Since this function is not defined to the left of 3, we cannot apply the limit laws to compute $\lim\limits_{x\to 3^-}\sqrt{x-3}$. In fact, since $f(x)=\sqrt{x-3}$ is undefined to the left of 3, $\lim\limits_{x\to 3^-}\sqrt{x-3}$

does not exist.

b. Since $f(x) = \sqrt{x-3}$ is defined to the right of 3, the limit laws do apply to $\lim\limits_{x \to 3^+} \sqrt{x-3}$. By applying these limit laws we obtain $\lim\limits_{x \to 3^+} \sqrt{x-3} = 0$.

In Example 2.1.8B we look at one-sided limits of a piecewise-defined function and use these limits to draw a conclusion about a two-sided limit of the same function.

Example 2.1.11B: Evaluating a Two-Sided Limit Using the Limit Laws

For $f(x) = \begin{cases} 4x-3, & \text{if } x < 2 \\ (x-3)^2, & \text{if } x \geq 2 \end{cases}$, evaluate each of the following limits:

a. $\lim\limits_{x \to 2^-} f(x)$
b. $\lim\limits_{x \to 2^+} f(x)$
c. $\lim\limits_{x \to 2} f(x)$

Solution

Figure 2.1.3 illustrates the function $f(x)$ and aids in our understanding of these limits.

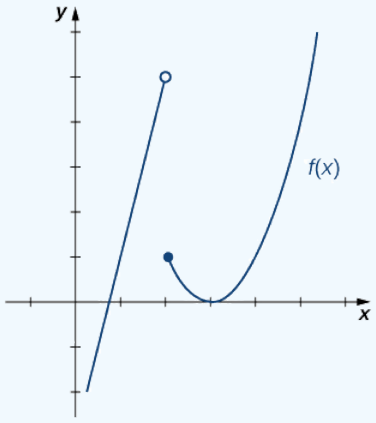


Figure 2.1.3: This graph shows a function $f(x)$.

a. Since $f(x) = 4x-3$ for all x in $(-\infty, 2)$, replace $f(x)$ in the limit with $4x-3$ and apply the limit laws:

$$\lim_{x \to 2^-} f(x) = \lim_{x \to 2^-} (4x-3) = 5$$

b. Since $f(x) = (x-3)^2$ for all x in $(2, +\infty)$, replace $f(x)$ in the limit with $(x-3)^2$ and apply the limit laws:

$$\lim_{x \to 2^+} f(x) = \lim_{x \to 2^-} (x-3)^2 = 1.$$

c. Since $\lim\limits_{x \to 2^-} f(x) = 5$ and $\lim\limits_{x \to 2^+} f(x) = 1$, we conclude that $\lim\limits_{x \to 2} f(x)$ does not exist.

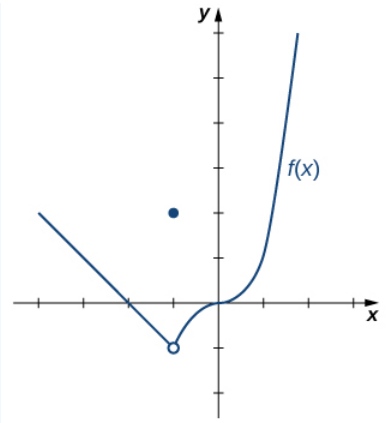
Exercise 2.1.11

Graph $f(x) = \begin{cases} -x-2, & \text{if } x < -1 \\ 2, & \text{if } x = -1 \\ x^3, & \text{if } x > -1 \end{cases}$ and evaluate $\lim\limits_{x \to -1^-} f(x)$.

Hint

Use the method in Example 2.1.8B to evaluate the limit.

Answer



$$\lim_{x \to -1^-} f(x) = -1$$

We now turn our attention to evaluating a limit of the form $\lim_{x \to a} \dfrac{f(x)}{g(x)}$, where $\lim_{x \to a} f(x) = K$, where $K \neq 0$ and $\lim_{x \to a} g(x) = 0$. That is, $f(x)/g(x)$ has the form $K/0, K \neq 0$ at a.

Example 2.1.12: Evaluating a Limit of the Form $K/0$, $K \neq 0$ Using the Limit Laws

Evaluate $\lim_{x \to 2^-} \dfrac{x-3}{x^2-2x}$.

Solution

Step 1. After substituting in $x = 2$, we see that this limit has the form $-1/0$. That is, as x approaches 2 from the left, the numerator approaches -1; and the denominator approaches 0. Consequently, the magnitude of $\dfrac{x-3}{x(x-2)}$ becomes infinite. To get a better idea of what the limit is, we need to factor the denominator:

$$\lim_{x \to 2^-} \frac{x-3}{x^2-2x} = \lim_{x \to 2^-} \frac{x-3}{x(x-2)}$$

Step 2. Since $x-2$ is the only part of the denominator that is zero when 2 is substituted, we then separate $1/(x-2)$ from the rest of the function:

$$= \lim_{x \to 2^-} \frac{x-3}{x} \cdot \frac{1}{x-2}$$

Step 3. Using the Limit Laws, we can write:

$$= \left(\lim_{x \to 2^-} \frac{x-3}{x} \right) \cdot \left(\lim_{x \to 2^-} \frac{1}{x-2} \right).$$

Step 4. $\lim_{x \to 2^-} \dfrac{x-3}{x} = -\dfrac{1}{2}$ and $\lim_{x \to 2^-} \dfrac{1}{x-2} = -\infty$. Therefore, the product of $(x-3)/x$ and $1/(x-2)$ has a limit of $+\infty$:

$$\lim_{x \to 2^-} \frac{x-3}{x^2-2x} = +\infty.$$

Exercise 2.1.12

Evaluate $\lim_{x \to 1} \dfrac{x+2}{(x-1)^2}$.

Solution

Use the methods from Example 2.1.9

Answer

$+\infty$

The Squeeze Theorem

The techniques we have developed thus far work very well for algebraic functions, but we are still unable to evaluate limits of very basic trigonometric functions. The next theorem, called the **squeeze theorem,** proves very useful for establishing basic trigonometric limits. This theorem allows us to calculate limits by "squeezing" a function, with a limit at a point a that is unknown, between two functions having a common known limit at a. Figure 2.1.4 illustrates this idea.

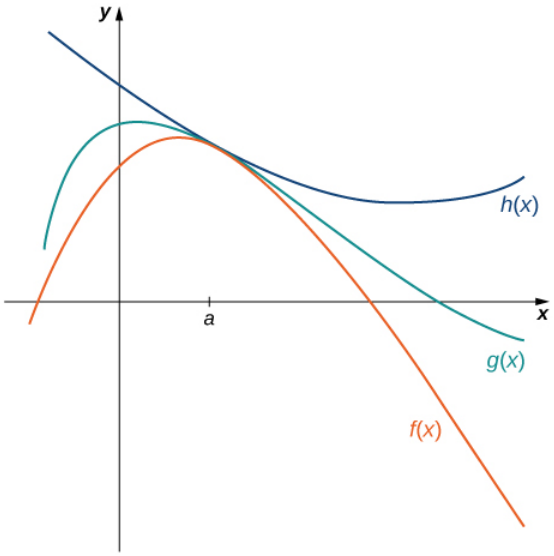


Figure 2.1.4: The Squeeze Theorem applies when $f(x) \le g(x) \le h(x)$ and $\lim\limits_{x \to a} f(x) = \lim\limits_{x \to a} h(x)$.

The Squeeze Theorem

Let $f(x)$, $g(x)$, and $h(x)$ be defined for all $x \neq a$ over an open interval containing a. If

$$f(x) \le g(x) \le h(x)$$

for all $x \neq a$ in an open interval containing a and

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

where L is a real number, then $\lim\limits_{x \to a} g(x) = L$.

Example 2.1.13: Applying the Squeeze Theorem

Apply the squeeze theorem to evaluate $\lim\limits_{x \to 0} x \cos x$.

Solution

Because $-1 \le \cos x \le 1$ for all x, we have $-x \le x \cos x \le x$ for $x \ge 0$ and $-x \ge x \cos x \ge x$ for $x \le 0$ (if x is negative the direction of the inequalities changes when we multiply). Since $\lim\limits_{x \to 0}(-x) = 0 = \lim\limits_{x \to 0} x$, from the squeeze theorem, we obtain $\lim\limits_{x \to 0} x \cos x = 0$. The graphs of $f(x) = -x$, $g(x) = x \cos x$, and $h(x) = x$ are shown in Figure 2.1.5.

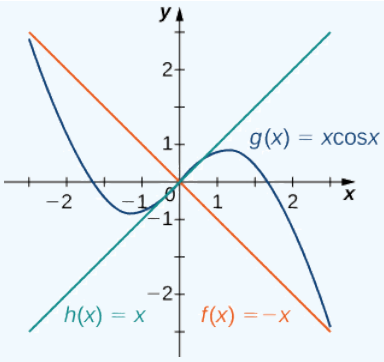


Figure 2.1.5: The graphs of $f(x)$, $g(x)$, and $h(x)$ are shown around the point $x = 0$.

Exercise 2.1.13

Use the squeeze theorem to evaluate $\lim_{x \to 0} x^2 \sin \frac{1}{x}$.

Hint

Use the fact that $-x^2 \leq x^2 \sin(1/x) \leq x^2$ to help you find two functions such that $x^2 \sin(1/x)$ is squeezed between them.

Answer

0

2.1: Limit of a Function and Limit Laws is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **2.4: Continuity** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **2.2: The Limit of a Function** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **2.3: The Limit Laws** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

2.2: The Precise Definition of a Limit

Learning Objectives

- Describe the epsilon-delta definition of a limit.
- Apply the epsilon-delta definition to find the limit of a function.
- Describe the epsilon-delta definitions of one-sided limits and infinite limits.
- Use the epsilon-delta definition to prove the limit laws.

By now you have progressed from the very informal definition of a limit in the introduction of this chapter to the intuitive understanding of a limit. At this point, you should have a very strong intuitive sense of what the limit of a function means and how you can find it. In this section, we convert this intuitive idea of a limit into a formal definition using precise mathematical language. The formal definition of a limit is quite possibly one of the most challenging definitions you will encounter early in your study of calculus; however, it is well worth any effort you make to reconcile it with your intuitive notion of a limit. Understanding this definition is the key that opens the door to a better understanding of calculus.

Quantifying Closeness

Before stating the formal definition of a limit, we must introduce a few preliminary ideas. Recall that the distance between two points a and b on a number line is given by $|a-b|$.

- The statement $|f(x)-L|<\varepsilon$ may be interpreted as: The distance between $f(x)$ and L is less than ε.
- The statement $0<|x-a|<\delta$ may be interpreted as: $x\neq a$ and the distance between x and a is less than δ.

It is also important to look at the following equivalences for absolute value:

- The statement $|f(x)-L|<\varepsilon$ is equivalent to the statement $L-\varepsilon<f(x)<L+\varepsilon$.
- The statement $0<|x-a|<\delta$ is equivalent to the statement $a-\delta<x<a+\delta$ and $x\neq a$.

With these clarifications, we can state the formal **epsilon-delta definition of the limit**.

Definition: Finite Limits (Formal)

Let $f(x)$ be defined for all $x\neq a$ over an open interval containing a. Let L be a real number. Then

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

if, for every $\varepsilon>0$, there exists a $\delta>0$, such that if $0<|x-a|<\delta$, then $|f(x)-L|<\varepsilon$.

This definition may seem rather complex from a mathematical point of view, but it becomes easier to understand if we break it down phrase by phrase. The statement itself involves something called a **universal quantifier** (for every $\varepsilon>0$), an **existential quantifier** (there exists a $\delta>0$), and, last, a **conditional statement** (if $0<|x-a|<\delta$, then $|f(x)-L|<\varepsilon$). Let's take a look at Table 2.2.1, which breaks down the definition and translates each part.

Table 2.2.1

Definition	Translation
1. For every $\varepsilon>0$,	1. For every positive distance ε from L,
2. there exists a $\delta>0$,	2. There is a positive distance δ from a,
3. such that	3. such that
4. if $0<\|x-a\|<\delta$, then $\|f(x)-L\|<\varepsilon$.	4. if x is closer than δ to a and $x\neq a$, then $f(x)$ is closer than ε to L.

We can get a better handle on this definition by looking at the definition geometrically. Figure 2.2.1 shows possible values of δ for various choices of $\varepsilon>0$ for a given function $f(x)$, a number a, and a limit L at a. Notice that as we choose smaller values of ε

(the distance between the function and the limit), we can always find a δ small enough so that if we have chosen an x value within δ of a, then the value of $f(x)$ is within ε of the limit L.

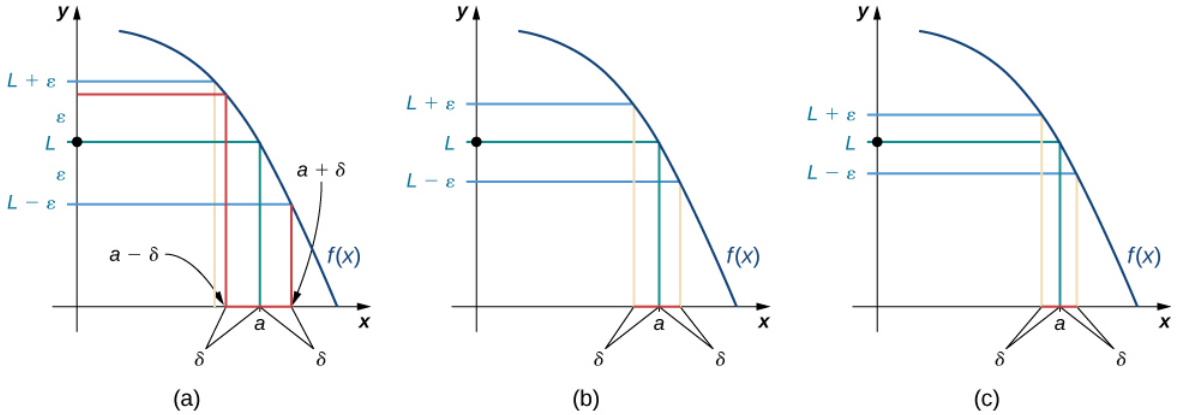


Figure 2.2.1: These graphs show possible values of δ, given successively smaller choices of ε.

Note

Visit the following applet to experiment with finding values of δ for selected values of ε:

- The epsilon-delta definition of limit

Example 2.2.1 shows how you can use this definition to prove a statement about the limit of a specific function at a specified value.

Example 2.2.1: Proving a Statement about the Limit of a Specific Function

Prove that $\lim\limits_{x\to 1}(2x+1)=3$.

Solution

Let $\varepsilon>0$.

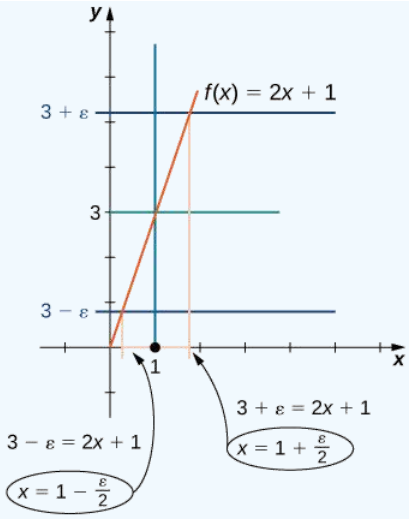
The first part of the definition begins "For every $\varepsilon>0$."This means we must prove that whatever follows is true no matter what positive value of ε is chosen. By stating "Let $\varepsilon>0$," we signal our intent to do so.

Choose $\delta=\frac{\varepsilon}{2}$.

The definition continues with "there exists a $\delta>0$." The phrase "there exists" in a mathematical statement is always a signal for a scavenger hunt. In other words, we must go and find δ. So, where exactly did $\delta=\varepsilon/2$ come from? There are two basic approaches to tracking down δ. One method is purely algebraic and the other is geometric.

We begin by tackling the problem from an algebraic point of view. Since ultimately we want $|(2x+1)-3|<\varepsilon$, we begin by manipulating this expression: $|(2x+1)-3|<\varepsilon$ is equivalent to $|2x-2|<\varepsilon$, which in turn is equivalent to $|2||x-1|<\varepsilon$. Last, this is equivalent to $|x-1|<\varepsilon/2$. Thus, it would seem that $\delta=\varepsilon/2$ is appropriate.

We may also find δ through geometric methods. Figure 2.2.2 demonstrates how this is done.



δ is the length of the smaller of the two distances marked in brown.

$\delta = \min\{1 + \frac{\varepsilon}{2} - 1, 1 - (1 - \frac{\varepsilon}{2})\}$

$= \min\{\frac{\varepsilon}{2}, \frac{\varepsilon}{2}\}$

$= \frac{\varepsilon}{2}$

Figure 2.2.2: This graph shows how we find δ geometrically.

Assume $0 < |x-1| < \delta$. When δ has been chosen, our goal is to show that if $0 < |x-1| < \delta$, then $|(2x+1)-3| < \varepsilon$. To prove any statement of the form "If this, then that," we begin by assuming "this" and trying to get "that."

Thus,

$|(2x+1)-3| = |2x-2|$ **property of absolute value**

$= |2(x-1)|$

$= |2||x-1|$ $|\mathbf{2}| = \mathbf{2}$

$= 2|x-1|$

$< 2 \cdot \delta$ **here's where we use the assumption that** $0 < |x-1| < \delta$

$= 2 \cdot \frac{\varepsilon}{2} = \varepsilon$ **here's where we use our choice of** $\delta = \varepsilon/2$

Analysis

In this part of the proof, we started with $|(2x+1)-3|$ and used our assumption $0 < |x-1| < \delta$ in a key part of the chain of inequalities to get $|(2x+1)-3|$ to be less than ε. We could just as easily have manipulated the assumed inequality $0 < |x-1| < \delta$ to arrive at $|(2x+1)-3| < \varepsilon$ as follows:

$0 < |x-1| < \delta \Rightarrow |x-1| < \delta$

$\Rightarrow -\delta < x-1 < \delta$

$\Rightarrow -\frac{\varepsilon}{2} < x-1 < \frac{\varepsilon}{2}$

$\Rightarrow -\varepsilon < 2x-2 < \varepsilon$

$\Rightarrow |2x-2| < \varepsilon$

$\Rightarrow |(2x+1)-3| < \varepsilon.$

Therefore, $\lim\limits_{x \to 1} (2x+1) = 3.$ (Having completed the proof, we state what we have accomplished.)

After removing all the remarks, here is a final version of the proof:

Let $\varepsilon > 0$.

Choose $\delta = \varepsilon/2$.

Assume $0 < |x-1| < \delta$.

Thus,

$$
\begin{aligned}
|(2x+1)-3| &= |2x-2| \\
&= |2(x-1)| \\
&= |2||x-1| \\
&= 2|x-1| \\
&< 2\cdot\delta \\
&= 2\cdot\frac{\varepsilon}{2} \\
&= \varepsilon.
\end{aligned}
$$

Therefore, $\lim\limits_{x\to 1}(2x+1)=3$.

The following Problem-Solving Strategy summarizes the type of proof we worked out in Example 2.2.1.

Problem-Solving Strategy: Proving That $\lim\limits_{x\to a} f(x)=L$ for a Specific Function $f(x)$

1. Let's begin the proof with the following statement: Let $\varepsilon>0$.
2. Next, we need to obtain a value for δ. After we have obtained this value, we make the following statement, filling in the blank with our choice of δ: Choose $\delta=$_____.
3. The next statement in the proof should be (at this point, we fill in our given value for a): Assume $0<|x-a|<\delta$.
4. Next, based on this assumption, we need to show that $|f(x)-L|<\varepsilon$, where $f(x)$ and L are our function $f(x)$ and our limit L. At some point, we need to use $0<|x-a|<\delta$.
5. We conclude our proof with the statement: Therefore, $\lim\limits_{x\to a} f(x)=L$.

Example 2.2.2: Proving a Statement about a Limit

Complete the proof that $\lim\limits_{x\to -1}(4x+1)=-3$ by filling in the blanks.

Let _____.

Choose $\delta=$_____.

Assume $0<|x-$_____$|<\delta$.

Thus, |_____−_____|=_____ε.

Solution

We begin by filling in the blanks where the choices are specified by the definition. Thus, we have

Let $\varepsilon>0$.

Choose $\delta=$_____.

Assume $0<|x-(-1)|<\delta$. (or equivalently, $0<|x+1|<\delta$.)

Thus, $|(4x+1)-(-3)|=|4x+4|=|4||x+1|<4\delta$ _____ε.

Focusing on the final line of the proof, we see that we should choose $\delta=\frac{\varepsilon}{4}$.

We now complete the final write-up of the proof:

Let $\varepsilon>0$.

Choose $\delta=\frac{\varepsilon}{4}$.

Assume $0<|x-(-1)|<\delta$ (or equivalently, $0<|x+1|<\delta$.)

Thus, $|(4x+1)-(-3)|=|4x+4|=|4||x+1|<4\delta=4(\varepsilon/4)=\varepsilon$.

Exercise 2.2.1

Complete the proof that $\lim_{x\to 2}(3x-2)=4$ by filling in the blanks.

Let _____.

Choose δ =_____.

Assume $0<|x-$____$|<$____.

Thus,

|_____−____|=_____ε.

Therefore, $\lim_{x\to 2}(3x-2)=4$.

Hint

Follow the outline in the Problem-Solving Strategy that we worked out in full in Example 2.2.2

Answer

Let $\varepsilon>0$; choose $\delta=\frac{\varepsilon}{3}$; assume $0<|x-2|<\delta$.

Thus, $|(3x-2)-4|=|3x-6|=|3|\cdot|x-2|<3\cdot\delta=3\cdot(\varepsilon/3)=\varepsilon$.

Therefore, $\lim_{x\to 2}(3x-2)=4$.

In Examples 2.2.1 and 2.2.2, the proofs were fairly straightforward, since the functions with which we were working were linear. In Example 2.2.3, we see how to modify the proof to accommodate a nonlinear function.

Example 2.2.3: Proving a Statement about the Limit of a Specific Function (Geometric Approach)

Prove that $\lim_{x\to 2}x^2=4$.

Solution

1. Let $\varepsilon>0$. The first part of the definition begins "For every $\varepsilon>0$," so we must prove that whatever follows is true no matter what positive value of ε is chosen. By stating "Let $\varepsilon>0$," we signal our intent to do so.

2. Without loss of generality, assume $\varepsilon\le 4$. Two questions present themselves: Why do we want $\varepsilon\le 4$ and why is it okay to make this assumption? In answer to the first question: Later on, in the process of solving for δ, we will discover that δ involves the quantity $\sqrt{4-\varepsilon}$. Consequently, we need $\varepsilon\le 4$. In answer to the second question: If we can find $\delta>0$ that "works" for $\varepsilon\le 4$, then it will "work" for any $\varepsilon>4$ as well. Keep in mind that, although it is always okay to put an upper bound on ε, it is never okay to put a lower bound (other than zero) on ε.

3. Choose $\delta=\min\{2-\sqrt{4-\varepsilon},\sqrt{4+\varepsilon}-2\}$. Figure 2.2.3 shows how we made this choice of δ.

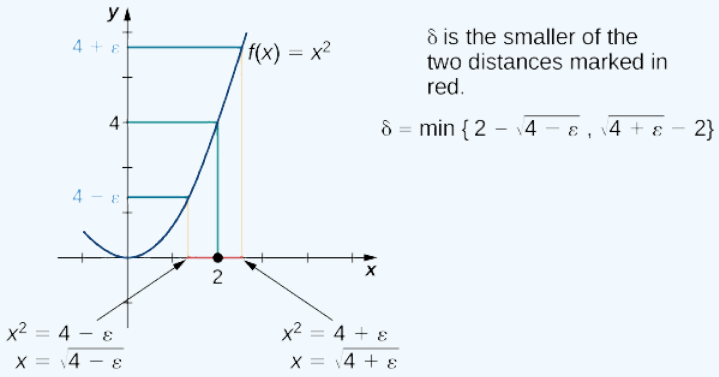


Figure 2.2.3: This graph shows how we find δ geometrically for a given ε for the proof in Example 2.2.3.

4. We must show: If $0<|x-2|<\delta$, then $|x^2-4|<\varepsilon$, so we must begin by assuming

$$0 < |x-2| < \delta.$$

We don't really need $0 < |x-2|$ (in other words, $x \neq 2$) for this proof. Since $0 < |x-2| < \delta \Rightarrow |x-2| < \delta$, it is okay to drop $0 < |x-2|$.

$$|x-2| < \delta.$$

Hence,

$$-\delta < x-2 < \delta.$$

Recall that $\delta = \min\{2-\sqrt{4-\varepsilon}, \sqrt{4+\varepsilon}-2\}$. Thus, $\delta \geq 2-\sqrt{4-\varepsilon}$ and consequently $-(2-\sqrt{4-\varepsilon}) \leq -\delta$. We also use $\delta \leq \sqrt{4+\varepsilon}-2$ here. We might ask at this point: Why did we substitute $2-\sqrt{4-\varepsilon}$ for δ on the left-hand side of the inequality and $\sqrt{4+\varepsilon}-2$ on the right-hand side of the inequality? If we look at Figure 2.2.3, we see that $2-\sqrt{4-\varepsilon}$ corresponds to the distance on the left of 2 on the x-axis and $\sqrt{4+\varepsilon}-2$ corresponds to the distance on the right. Thus,

$$-(2-\sqrt{4-\varepsilon}) \leq -\delta < x-2 < \delta \leq \sqrt{4+\varepsilon}-2.$$

We simplify the expression on the left:

$$-2+\sqrt{4-\varepsilon} < x-2 < \sqrt{4+\varepsilon}-2 \quad .$$

Then, we add 2 to all parts of the inequality:

$$\sqrt{4-\varepsilon} < x < \sqrt{4+\varepsilon}.$$

We square all parts of the inequality. It is okay to do so, since all parts of the inequality are positive:

$$4-\varepsilon < x^2 < 4+\varepsilon.$$

We subtract 4 from all parts of the inequality:

$$-\varepsilon < x^2-4 < \varepsilon.$$

Last,

$$|x^2-4| < \varepsilon.$$

5. Therefore,

$$\lim_{x\to 2} x^2 = 4.$$

? Exercise 2.2.2

Find δ corresponding to $\varepsilon > 0$ for a proof that $\lim_{x\to 9} \sqrt{x} = 3$.

Hint

Draw a graph similar to the one in Example 2.2.3

Answer

Choose $\delta = \min\{9-(3-\varepsilon)^2,\ (3+\varepsilon)^2-9\}$.

The geometric approach to proving that the limit of a function takes on a specific value works quite well for some functions. Also, the insight into the formal definition of the limit that this method provides is invaluable. However, we may also approach limit proofs from a purely algebraic point of view. In many cases, an algebraic approach may not only provide us with additional insight into the definition, it may prove to be simpler as well. Furthermore, an algebraic approach is the primary tool used in proofs of statements about limits. For Example 2.2.4, we take on a purely algebraic approach.

✓ Example 2.2.4:Proving a Statement about the Limit of a Specific Function (Algebraic Approach)

Prove that $\lim_{x\to -1} (x^2-2x+3) = 6$.

Solution

Let's use our outline from the Problem-Solving Strategy:

1. Let $\varepsilon > 0$.

2. Choose $\delta = \min\{1, \varepsilon/5\}$. This choice of δ may appear odd at first glance, but it was obtained by taking a look at our ultimate desired inequality: $|(x^2 - 2x + 3) - 6| < \varepsilon$. This inequality is equivalent to $|x+1| \cdot |x-3| < \varepsilon$. At this point, the temptation simply to choose $\delta = \frac{\varepsilon}{x-3}$ is very strong. Unfortunately, our choice of δ must depend on ε only and no other variable. If we can replace $|x-3|$ by a numerical value, our problem can be resolved. This is the place where assuming $\delta \leq 1$ comes into play. The choice of $\delta \leq 1$ here is arbitrary. We could have just as easily used any other positive number. In some proofs, greater care in this choice may be necessary. Now, since $\delta \leq 1$ and $|x+1| < \delta \leq 1$, we are able to show that $|x-3| < 5$. Consequently, $|x+1| \cdot |x-3| < |x+1| \cdot 5$. At this point we realize that we also need $\delta \leq \varepsilon/5$. Thus, we choose $\delta = \min\{1, \varepsilon/5\}$.

3. Assume $0 < |x+1| < \delta$. Thus,

$$|x+1| < 1 \text{ and } |x+1| < \frac{\varepsilon}{5}.$$

Since $|x+1| < 1$, we may conclude that $-1 < x+1 < 1$. Thus, by subtracting 4 from all parts of the inequality, we obtain $-5 < x-3 < -1$. Consequently, $|x-3| < 5$. This gives us

$$|(x^2 - 2x + 3) - 6| = |x+1| \cdot |x-3| < \frac{\varepsilon}{5} \cdot 5 = \varepsilon.$$

Therefore,

$$\lim_{x \to -1} (x^2 - 2x + 3) = 6.$$

Exercise 2.2.3

Complete the proof that $\lim_{x \to 1} x^2 = 1$.

Let $\varepsilon > 0$; choose $\delta = \min\{1, \varepsilon/3\}$; assume $0 < |x-1| < \delta$.

Since $|x-1| < 1$, we may conclude that $-1 < x-1 < 1$. Thus, $1 < x+1 < 3$. Hence, $|x+1| < 3$.

Hint

Use Example 2.2.4as a guide.

Answer

$|x^2 - 1| = |x-1| \cdot |x+1| < \varepsilon/3 \cdot 3 = \varepsilon$

You will find that, in general, the more complex a function, the more likely it is that the algebraic approach is the easiest to apply. The algebraic approach is also more useful in proving statements about limits.

Proving Limit Laws

We now demonstrate how to use the epsilon-delta definition of a limit to construct a rigorous proof of one of the limit laws. The **triangle inequality** is used at a key point of the proof, so we first review this key property of absolute value.

Definition: The Triangle Inequality

The **triangle inequality** states that if a and b are any real numbers, then $|a+b| \leq |a| + |b|$.

Proof

We prove the following limit law: If $\lim_{x \to a} f(x) = L$ and $\lim_{x \to a} g(x) = M$, then $\lim_{x \to a} (f(x) + g(x)) = L + M$.

Let $\varepsilon > 0$.

Choose $\delta_1 > 0$ so that if $0 < |x-a| < \delta_1$, then $|f(x)-L| < \varepsilon/2$.

Choose $\delta_2 > 0$ so that if $0 < |x-a| < \delta_2$, then $|g(x)-M| < \varepsilon/2$.

Choose $\delta = \min\{\delta_1, \delta_2\}$.

Assume $0 < |x-a| < \delta$.

Thus,

$$0 < |x-a| < \delta_1 \text{ and } 0 < |x-a| < \delta_2 .$$

Hence,

$$\begin{aligned} |(f(x)+g(x))-(L+M)| &= |(f(x)-L)+(g(x)-M)| \\ &\leq |f(x)-L|+|g(x)-M| \\ &< \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon \end{aligned} .$$

□

We now explore what it means for a limit not to exist. The limit $\lim\limits_{x\to a} f(x)$ does not exist if there is no real number L for which $\lim\limits_{x\to a} f(x) = L$. Thus, for all real numbers L, $\lim\limits_{x\to a} f(x) \neq L$. To understand what this means, we look at each part of the definition of $\lim\limits_{x\to a} f(x) = L$ together with its opposite. A translation of the definition is given in Table 2.2.2.

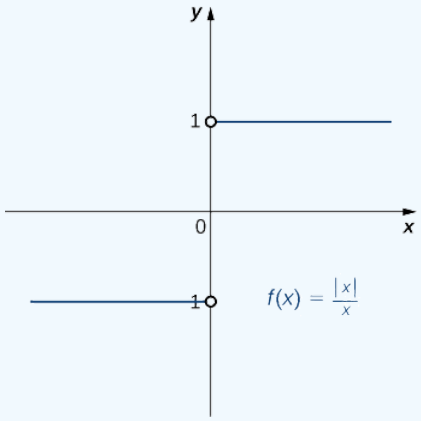
Table 2.2.2. Translation of the Definition of $\lim_{x\to a} f(x) = L$ and its Opposite

Definition	Opposite
1. For every $\varepsilon > 0$,	1. There exists $\varepsilon > 0$ so that
2. there exists a $\delta > 0$, so that	2. for every $\delta > 0$,
3. if $0 < \|x-a\| < \delta$, then $\|f(x)-L\| < \varepsilon$.	3. There is an x satisfying $0 < \|x-a\| < \delta$ so that $\|f(x)-L\| \geq \varepsilon$.

Finally, we may state what it means for a limit not to exist. The limit $\lim\limits_{x\to a} f(x)$ does not exist if for every real number L, there exists a real number $\varepsilon > 0$ so that for all $\delta > 0$, there is an x satisfying $0 < |x-a| < \delta$, so that $|f(x)-L| \geq \varepsilon$. Let's apply this in Example 2.2.5 to show that a limit does not exist.

Example 2.2.5: Showing That a Limit Does Not Exist

Show that $\lim\limits_{x\to 0} \dfrac{|x|}{x}$ does not exist. The graph of $f(x) = |x|/x$ is shown here:



Solution

Suppose that L is a candidate for a limit. Choose $\varepsilon = 1/2$.

Let $\delta > 0$. Either $L \geq 0$ or $L < 0$. If $L \geq 0$, then let $x = -\delta/2$.

Thus,

$$|x-0| = |-\tfrac{\delta}{2} - 0| = \tfrac{\delta}{2} < \delta$$

and

$$\left|\frac{\left|-\frac{\delta}{2}\right|}{-\frac{\delta}{2}} - L\right| = |-1-L| = L+1 \geq 1 > \tfrac{1}{2} = \varepsilon \quad .$$

On the other hand, if $L < 0$, then let $x = \delta/2$. Thus,

$$|x-0| = |\tfrac{\delta}{2} - 0| = \tfrac{\delta}{2} < \delta$$

and

$$\left|\frac{\left|\frac{\delta}{2}\right|}{\frac{\delta}{2}} - L\right| = |1-L| = |L|+1 \geq 1 > \tfrac{1}{2} = \varepsilon \quad .$$

Thus, for any value of L, $\lim\limits_{x\to 0} \dfrac{|x|}{x} \neq L$.

One-Sided Limits

Just as we first gained an intuitive understanding of limits and then moved on to a more rigorous definition of a limit, we now revisit one-sided limits. To do this, we modify the epsilon-delta definition of a limit to give formal epsilon-delta definitions for limits from the right and left at a point. These definitions only require slight modifications from the definition of the limit. In the definition of the limit from the right, the inequality $0 < x - a < \delta$ replaces $0 < |x-a| < \delta$, which ensures that we only consider values of x that are greater than (to the right of) a. Similarly, in the definition of the limit from the left, the inequality $-\delta < x - a < 0$ replaces $0 < |x-a| < \delta$, which ensures that we only consider values of x that are less than (to the left of) a.

Definition: One-Sided Limits (Formal)

Limit from the Right: Let $f(x)$ be defined over an open interval of the form (a, b) where $a < b$. Then

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

if for every $\varepsilon > 0$, there exists a $\delta > 0$, such that if $0 < x - a < \delta$, then $|f(x) - L| < \varepsilon$.

Limit from the Left: Let $f(x)$ be defined over an open interval of the form (b, c) where $b < c$. Then,

$$\lim_{x\to c^-} f(x) = L$$

if for every $\varepsilon > 0$,there exists a $\delta > 0$ such that if $-\delta < x - c < 0$, then $|f(x) - L| < \varepsilon$.

Example 2.2.6: Proving a Statement about a Limit From the Right

Prove that

$$\lim_{x\to 4^+} \sqrt{x-4} = 0.$$

Solution

Let $\varepsilon > 0$.

Choose $\delta = \varepsilon^2$. Since we ultimately want $|\sqrt{x-4} - 0| < \varepsilon$, we manipulate this inequality to get $\sqrt{x-4} < \varepsilon$ or, equivalently, $0 < x - 4 < \varepsilon^2$, making $\delta = \varepsilon^2$ a clear choice. We may also determine δ geometrically, as shown in Figure 2.2.4.

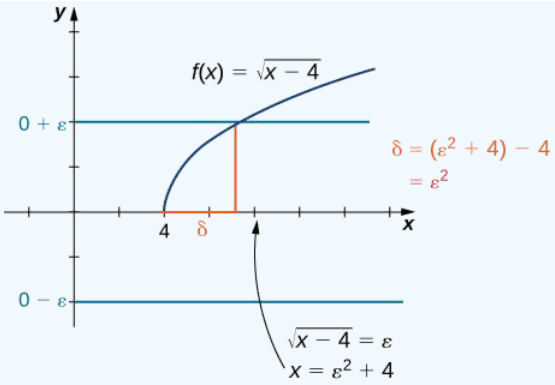


Figure 2.2.4: This graph shows how we find δ for the proof in Example 2.2.6.

Assume $0 < x - 4 < \delta$. Thus, $0 < x - 4 < \varepsilon^2$. Hence, $0 < \sqrt{x-4} < \varepsilon$. Finally, $|\sqrt{x-4} - 0| < \varepsilon$. Therefore, $\lim\limits_{x \to 4^+} \sqrt{x-4} = 0$.

? Exercise 2.2.4

Find δ corresponding to ε for a proof that $\lim\limits_{x \to 1^-} \sqrt{1-x} = 0$.

Hint

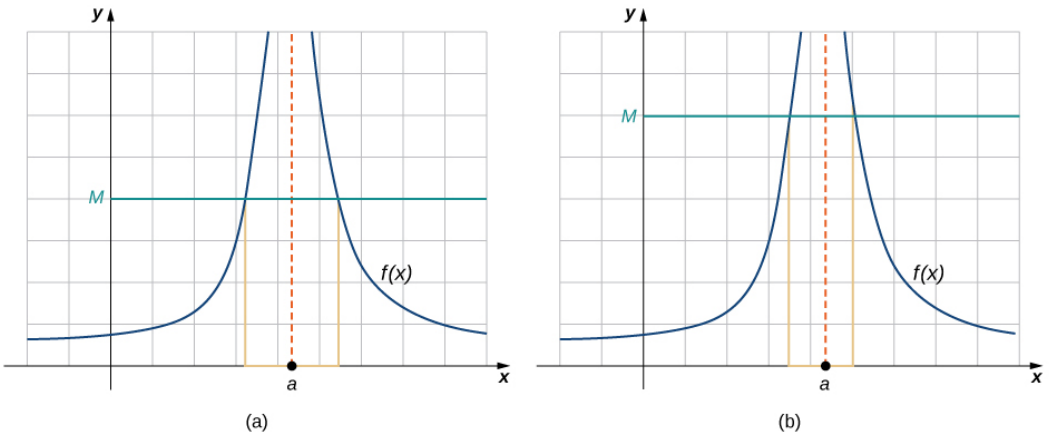
Sketch the graph and use Example 2.2.6as a solving guide.

Answer

$\delta = \varepsilon^2$

Infinite Limits

We conclude the process of converting our intuitive ideas of various types of limits to rigorous formal definitions by pursuing a formal definition of infinite limits. To have $\lim\limits_{x \to a} f(x) = +\infty$, we want the values of the function $f(x)$ to get larger and larger as x approaches a. Instead of the requirement that $|f(x) - L| < \varepsilon$ for arbitrarily small ε when $0 < |x - a| < \delta$ for small enough δ, we want $f(x) > M$ for arbitrarily large positive M when $0 < |x - a| < \delta$ for small enough δ. Figure 2.2.5 illustrates this idea by showing the value of δ for successively larger values of M.



In each graph, δ is the smaller of the lengths of the two brown intervals.

Figure 2.2.5: These graphs plot values of δ for M to show that $\lim\limits_{x \to a} f(x) = +\infty$.

Definition: Infinite Limits (Formal)

Let $f(x)$ be defined for all $x \neq a$ in an open interval containing a. Then, we have an infinite limit

$$\lim_{x \to a} f(x) = +\infty$$

if for every $M > 0$, there exists $\delta > 0$ such that if $0 < |x - a| < \delta$, then $f(x) > M$.

Let $f(x)$ be defined for all $x \neq a$ in an open interval containing a. Then, we have a negative infinite limit

$$\lim_{x \to a} f(x) = -\infty$$

if for every $M > 0$, there exists $\delta > 0$ such that if $0 < |x - a| < \delta$, then $f(x) < -M$.

Example 2.2.7: Proving a Statement about an Infinite Limit

Prove that $\lim_{x \to 3} \dfrac{1}{(x-3)^2} = \infty$.

Solution

We use a very similar approach to our previous Problem-Solving Strategy. We first find an appropriate $\delta > 0$. Then we write our proof.

Step 1: First we find an appropriate $\delta > 0$.

1. Let M be any real number such that $M > 0$.

2. Let $f(x) = \dfrac{1}{(x-3)^2} > M$. Then we solve for the expression $x - 3$.

Multiplying both sides of the inequality by the positive quantity $(x-3)^2$ and dividing both sides by the positive quantity M gives us:

$$\frac{1}{M} > (x-3)^2$$

Taking the square root of both sides, we have,

$$\sqrt{\frac{1}{M}} > |x-3|. \qquad \left(\text{Remember that } \sqrt{x^2} = |x|.\right)$$

Rewriting this statement gives us, $0 < |x-3| < \sqrt{\dfrac{1}{M}}$. From this we choose $\delta = \sqrt{\dfrac{1}{M}}$.

Step 2: Now we write a proof.

3. Let $\delta = \sqrt{\dfrac{1}{M}}$ and assume $0 < |x-3| < \delta = \sqrt{\dfrac{1}{M}}$.

Thus,

$$|x-3| < \sqrt{\frac{1}{M}}.$$

Squaring both sides gives us,

$$(x-3)^2 < \frac{1}{M}.$$

Taking the reciprocal of both sides (and remembering that this will reverse the direction of the inequality),

$$\frac{1}{(x-3)^2} > M.$$

Therefore, we have proven that

$$\lim_{x \to 3} \frac{1}{(x-3)^2} = \infty.$$

A very similar proof will be needed for a limit that is equal to $-\infty$.

Note that a one-sided limit approach will often need to be taken with this type of limit. For example, to prove: $\lim_{x \to 0^+} \frac{1}{x} = \infty$.

Key Concepts

- The intuitive notion of a limit may be converted into a rigorous mathematical definition known as the epsilon-delta definition of the limit.
- The epsilon-delta definition may be used to prove statements about limits.
- The epsilon-delta definition of a limit may be modified to define one-sided limits.
- A similar definition of an infinite limit can be used to prove statements about infinite limits.

Glossary

epsilon-delta definition of the limit

$\lim_{x \to a} f(x) = L$ if for every $\varepsilon > 0$, there exists a $\delta > 0$ such that if $0 < |x - a| < \delta$, then $|f(x) - L| < \varepsilon$

triangle inequality

If a and b are any real numbers, then $|a + b| \leq |a| + |b|$

formal definition of an infinite limit

$\lim_{x \to a} f(x) = \infty$ if for every $M > 0$, there exists a $\delta > 0$ such that if $0 < |x - a| < \delta$, then $f(x) > M$
$\lim_{x \to a} f(x) = -\infty$ if for every $M > 0$, there exists a $\delta > 0$ such that if $0 < |x - a| < \delta$, then $f(x) < -M$

- Gilbert Strang (MIT) and Edwin “Jed” Herman (Harvey Mudd) with many contributing authors. This content by OpenStax is licensed with a CC-BY-SA-NC 4.0 license. Download for free at http://cnx.org.
- Paul Seeburger (Monroe Community College), added Example $2.2.7$ and entries for infinite limits under Key Concepts and the Glossary.

This page titled 2.2: The Precise Definition of a Limit is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **2.5: The Precise Definition of a Limit** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

2.3: One-Sided Limits

Learning Objectives

- Define one-sided limits and provide examples.
- Explain the relationship between one-sided and two-sided limits.
- Define a vertical asymptote.

One-Sided Limits

Sometimes indicating that the limit of a function fails to exist at a point does not provide us with enough information about the behavior of the function at that particular point. To see this, we now revisit the function $g(x) = |x-2|/(x-2)$ introduced at the beginning of the section (see Figure 2.3.1(b)). As we pick values of x close to 2, $g(x)$ does not approach a single value, so the limit as x approaches 2 does not exist—that is, $\lim_{x\to 2} g(x)$ DNE. However, this statement alone does not give us a complete picture of the behavior of the function around the x-value 2. To provide a more accurate description, we introduce the idea of a **one-sided limit**. For all values to the left of 2 (or the negative side of 2), $g(x) = -1$. Thus, as x approaches 2 from the left, $g(x)$ approaches -1. Mathematically, we say that the limit as x approaches 2 from the left is -1. Symbolically, we express this idea as

$$\lim_{x\to 2^-} g(x) = -1.$$

Similarly, as x approaches 2 from the right (or from the positive side), $g(x)$ approaches 1. Symbolically, we express this idea as

$$\lim_{x\to 2^+} g(x) = 1.$$

We can now present an informal definition of one-sided limits.

Definition: One-sided Limits

We define two types of one-sided limits.

Limit from the left:

Let $f(x)$ be a function defined at all values in an open interval of the form (z, a), and let L be a real number. If the values of the function $f(x)$ approach the real number L as the values of x (where $x < a$) approach the number a, then we say that L is the limit of $f(x)$ as x approaches a from the left. Symbolically, we express this idea as

$$\lim_{x\to a^-} f(x) = L.$$

Limit from the right:

Let $f(x)$ be a function defined at all values in an open interval of the form (a, c), and let L be a real number. If the values of the function $f(x)$ approach the real number L as the values of x (where $x > a$) approach the number a, then we say that L is the limit of $f(x)$ as x approaches a from the right. Symbolically, we express this idea as

$$\lim_{x\to a^+} f(x) = L.$$

Example 2.3.1: Evaluating One-Sided Limits

For the function $f(x) = \begin{cases} x+1, & \text{if } x < 2 \\ x^2-4, & \text{if } x \geq 2 \end{cases}$, evaluate each of the following limits.

a. $\lim_{x\to 2^-} f(x)$

b. $\lim_{x\to 2^+} f(x)$

Solution

We can use tables of functional values again. Observe in Table 2.3.6 that for values of x less than 2, we use $f(x) = x + 1$ and for values of x greater than 2, we use $f(x) = x^2 - 4$.

Table 2.3.6

x	$f(x) = x + 1$	x	$f(x) = x^2 - 4$
1.9	2.9	2.1	0.41
1.99	2.99	2.01	0.0401
1.999	2.999	2.001	0.004001
1.9999	2.9999	2.0001	0.00040001
1.99999	2.99999	2.00001	0.0000400001

Based on this table, we can conclude that a. $\lim_{x \to 2^-} f(x) = 3$ and b. $\lim_{x \to 2^+} f(x) = 0$. Therefore, the (two-sided) limit of $f(x)$ does not exist at $x = 2$. Figure 2.3.7 shows a graph of $f(x)$ and reinforces our conclusion about these limits.

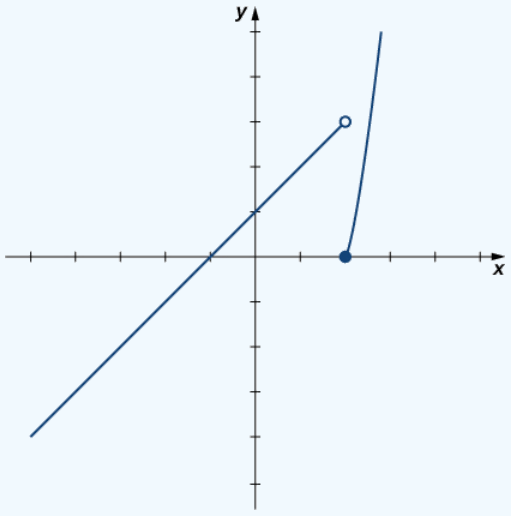


Figure 2.3.7: The graph of $f(x) = \begin{cases} x+1, & \text{if } x < 2 \\ x^2 - 4, & \text{if } x \geq 2 \end{cases}$ has a break at $x = 2$.

Exercise 2.3.1

Use a table of functional values to estimate the following limits, if possible.

a. $\lim_{x \to 2^-} \dfrac{|x^2 - 4|}{x - 2}$

b. $\lim_{x \to 2^+} \dfrac{|x^2 - 4|}{x - 2}$

Hint

Use x-values 1.9, 1.99, 1.999, 1.9999, 1.99999 to estimate $\lim_{x \to 2^-} \dfrac{|x^2 - 4|}{x - 2}$.

Use x-values 2.1, 2.01, 2.001, 2.0001, 2.00001 to estimate $\lim_{x \to 2^+} \dfrac{|x^2 - 4|}{x - 2}$.

(These tables are available from a previous Checkpoint problem.)

Solution a

a. $\lim_{x \to 2^-} \dfrac{|x^2 - 4|}{x - 2} = -4$

Solution b

$$\lim_{x\to 2^+} \frac{|x^2-4|}{x-2} = 4$$

Let us now consider the relationship between the limit of a function at a point and the limits from the right and left at that point. It seems clear that if the limit from the right and the limit from the left have a common value, then that common value is the limit of the function at that point. Similarly, if the limit from the left and the limit from the right take on different values, the limit of the function does not exist. These conclusions are summarized in Note.

Relating One-Sided and Two-Sided Limits

Let $f(x)$ be a function defined at all values in an open interval containing a, with the possible exception of a itself, and let L be a real number. Then,

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

if and only if $\lim_{x\to a^-} f(x) = L$ and $\lim_{x\to a^+} f(x) = L$.

Example 2.3.2 : Evaluating one sided limits

Let $f(x) = \begin{cases} x & 0 \le x \le 1 \\ 3-x & 1 < x < 2 \end{cases}$, as shown in Figure 1.21. Find each of the following:

1. $\lim_{x\to 1^-} f(x)$
2. $\lim_{x\to 1^+} f(x)$
3. $\lim_{x\to 1} f(x)$
4. $f(1)$
5. $\lim_{x\to 0^+} f(x)$
6. $f(0)$
7. $\lim_{x\to 2^-} f(x)$
8. $f(2)$

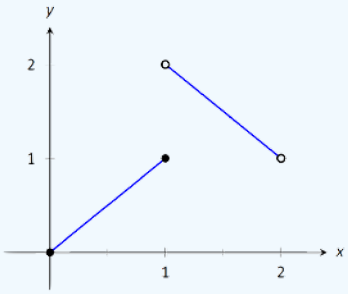


FIGURE 1.21: A graph of f in Example 17.

Solution

For these problems, the visual aid of the graph is likely more effective in evaluating the limits than using f itself. Therefore we will refer often to the graph.

1. As x goes to 1 *from the left*, we see that $f(x)$ is approaching the value of 1. Therefore $\lim_{x\to 1^-} f(x) = 1$.
2. As x goes to 1 *from the right*, we see that $f(x)$ is approaching the value of 2. Recall that it does not matter that there is an "open circle" there; we are evaluating a limit, not the value of the function. Therefore $\lim_{x\to 1^+} f(x) = 2$.
3. *The* limit of f as x approaches 1 does not exist, as discussed in the first section. The function does not approach one particular value, but two different values from the left and the right.

4. Using the definition and by looking at the graph we see that $f(1) = 1$.
5. As x goes to 0 from the right, we see that $f(x)$ is also approaching 0. Therefore $\lim\limits_{x \to 0^+} f(x) = 0$. Note we cannot consider a left-hand limit at 0 as f is not defined for values of $x < 0$.
6. Using the definition and the graph, $f(0) = 0$.
7. As x goes to 2 from the left, we see that $f(x)$ is approaching the value of 1. Therefore $\lim\limits_{x \to 2^-} f(x) = 1$.
8. The graph and the definition of the function show that $f(2)$ is not defined.

Note how the left and right-hand limits were different at $x = 1$. This, of course, causes *the* limit to not exist. The following theorem states what is fairly intuitive: *the* limit exists precisely when the left and right-hand limits are equal.

? Theorem: Limits and One Sided Limits

Let f be a function defined on an open interval I containing c. Then

$$\lim_{x \to c} f(x) = L \tag{2.3.1}$$

if, and only if,

$$\lim_{x \to c^-} f(x) = L \quad \text{and} \quad \lim_{x \to c^+} f(x) = L. \tag{2.3.2}$$

The phrase "if, and only if" means the two statements are *equivalent*: they are either both true or both false. If the limit equals L, then the left and right hand limits both equal L. If the limit is not equal to L, then at least one of the left and right-hand limits is not equal to L (it may not even exist).

One thing to consider in Examples 17 - 20 is that the value of the function may/may not be equal to the value(s) of its left/right-hand limits, even when these limits agree.

✓ Example ✓ 2.3.3 : Evaluating limits of a piecewise-defined function

Let $f(x) = \begin{cases} 2 - x & 0 < x < 1 \\ (x-2)^2 & 1 < x < 2 \end{cases}$, as shown in Figure 1.22. Evaluate the following.

1. $\lim\limits_{x \to 1^-} f(x)$
2. $\lim\limits_{x \to 1^+} f(x)$
3. $\lim\limits_{x \to 1} f(x)$
4. $f(1)$
5. $\lim\limits_{x \to 0^+} f(x)$
6. $f(0)$
7. $\lim\limits_{x \to 2^-} f(x)$
8. $f(2)$

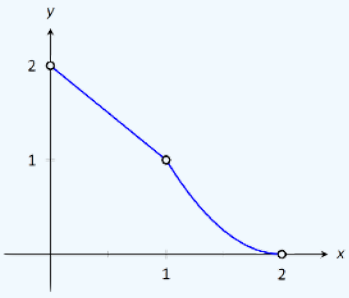


FIGURE 1.22: A graph of f from Example 18.

Solution

Again we will evaluate each using both the definition of f and its graph.

1. As x approaches 1 from the left, we see that $f(x)$ approaches 1. Therefore $\lim_{x\to 1^-} f(x) = 1$.
2. As x approaches 1 from the right, we see that again $f(x)$ approaches 1. Therefore $\lim_{x\to 1+} f(x) = 1$.
3. *The* limit of f as x approaches 1 exists and is 1, as f approaches 1 from both the right and left. Therefore $\lim_{x\to 1} f(x) = 1$.
4. $f(1)$ is not defined. Note that 1 is not in the domain of f as defined by the problem, which is indicated on the graph by an open circle when $x = 1$.
5. As x goes to 0 from the right, $f(x)$ approaches 2. So $\lim_{x\to 0^+} f(x) = 2$.
6. $f(0)$ is not defined as 0 is not in the domain of f.
7. As x goes to 2 from the left, $f(x)$ approaches 0. So $\lim_{x\to 2^-} f(x) = 0$.
8. $f(2)$ is not defined as 2 is not in the domain of f.

✓ Example ✓ 2.3.4 : Evaluating limits of a piecewise-defined function

Let $f(x) = \begin{cases} (x-1)^2 & 0 \le x \le 2, x \ne 1 \\ 1 & x = 1 \end{cases}$, as shown in Figure 1.23. Evaluate the following.

1. $\lim_{x\to 1^-} f(x)$
2. $\lim_{x\to 1^+} f(x)$
3. $\lim_{x\to 1} f(x)$
4. $f(1)$

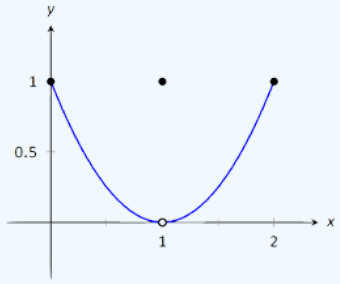


FIGURE 1.23: Graphing f in Example 19.

It is clear by looking at the graph that both the left and right-hand limits of f, as x approaches 1, is 0. Thus it is also clear that *the* limit is 0; i.e., $\lim_{x\to 1} f(x) = 0$. It is also clearly stated that $f(1) = 1$.

✓ Example ✓ 2.3.5 : Evaluating limits of a piecewise-defined function

Let $f(x) = \begin{cases} x^2 & 0 \le x \le 1 \\ 2-x & 1 < x \le 2 \end{cases}$, as shown in Figure 1.24. Evaluate the following.

1. $\lim_{x\to 1^-} f(x)$
2. $\lim_{x\to 1^+} f(x)$
3. $\lim_{x\to 1} f(x)$
4. $f(1)$

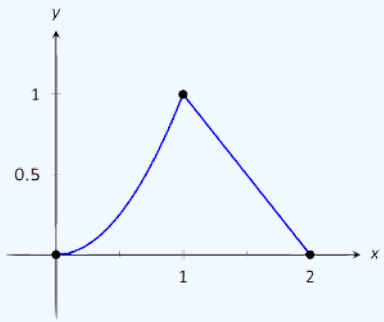


FIGURE 1.24: Graphing f in Example 20.

Solution

It is clear from the definition of the function and its graph that all of the following are equal:

$$\lim_{x\to 1^-} f(x) = \lim_{x\to 1^+} f(x) = \lim_{x\to 1} f(x) = f(1) = 1. \tag{2.3.3}$$

In Examples 17 - 20 we were asked to find both $\lim_{x\to 1} f(x)$ and $f(1)$. Consider the following table:

	$\lim_{x\to 1} f(x)$	$f(1)$
Example 17	does not exist	1
Example 18	1	not defined
Example 19	0	1
Example 20	1	1

(2.3.4)

Only in Example 20 do both the function and the limit exist and agree. This seems "nice;" in fact, it seems "normal." This is in fact an important situation which we explore in the next section, entitled "Continuity." In short, a *continuous function* is one in which when a function approaches a value as $x \to c$ (i.e., when $\lim_{x\to c} f(x) = L$), it actually *attains* that value at c. Such functions behave nicely as they are very predictable.

Example 2.3.6 : Using the Squeeze Theorem

Use the Squeeze Theorem to show that

$$\lim_{x\to 0} \frac{\sin x}{x} = 1.$$

Solution

We begin by considering the unit circle. Each point on the unit circle has coordinates $(\cos\theta, \sin\theta)$ for some angle θ as shown in Figure 2.3.1. Using similar triangles, we can extend the line from the origin through the point to the point $(1, \tan\theta)$, as shown. (Here we are assuming that $0 \le \theta \le \pi/2$ .Later we will show that we can also consider $\theta \le 0$.)

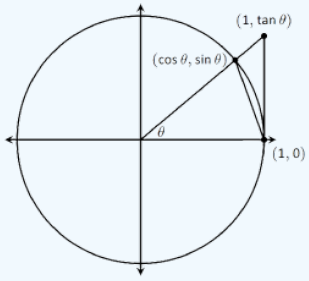
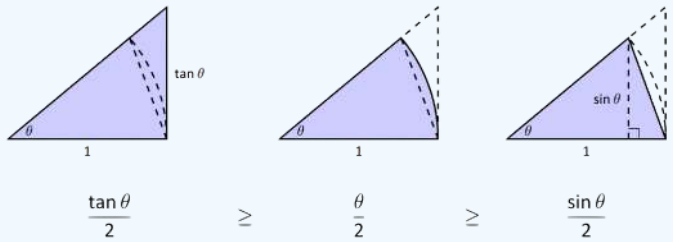


Figure 2.3.1: The unit circle and related triangles.

Figure 1.19 shows three regions have been constructed in the first quadrant, two triangles and a sector of a circle, which are also drawn below. The area of the large triangle is $\frac{1}{2}\tan\theta$; the area of the sector is $\theta/2$; the area of the triangle contained inside

the sector is $\frac{1}{2}\sin\theta$. It is then clear from the diagram that



$$\frac{\tan\theta}{2} \quad \geq \quad \frac{\theta}{2} \quad \geq \quad \frac{\sin\theta}{2}$$

Multiply all terms by $\frac{2}{\sin\theta}$, giving

$$\frac{1}{\cos\theta} \geq \frac{\theta}{\sin\theta} \geq 1. \tag{2.3.5}$$

Taking reciprocals reverses the inequalities, giving

$$\cos\theta \leq \frac{\sin\theta}{\theta} \leq 1. \tag{2.3.6}$$

(These inequalities hold for all values of θ near 0, even negative values, since $\cos(-\theta) = \cos\theta$ and $\sin(-\theta) = -\sin\theta$.)

Now take limits.

$$\lim_{\theta\to 0}\cos\theta \leq \lim_{\theta\to 0}\frac{\sin\theta}{\theta} \leq \lim_{\theta\to 0} 1 \tag{2.3.7}$$

$$\cos 0 \leq \lim_{\theta\to 0}\frac{\sin\theta}{\theta} \leq 1 \tag{2.3.8}$$

$$1 \leq \lim_{\theta\to 0}\frac{\sin\theta}{\theta} \leq 1 \tag{2.3.9}$$

Clearly this means that $\lim_{\theta\to 0}\frac{\sin\theta}{\theta} = 1$

Two notes about the previous example are worth mentioning. First, one might be discouraged by this application, thinking "I would *never* have come up with that on my own. This is too hard!" Don't be discouraged; within this text we will guide you in your use of the Squeeze Theorem. As one gains mathematical maturity, clever proofs like this are easier and easier to create.

Second, this limit tells us more than just that as x approaches 0, $\sin(x)/x$ approaches 1. Both x and $\sin x$ are approaching 0, but the *ratio* of x and $\sin x$ approaches 1, meaning that they are approaching 0 in essentially the same way. Another way of viewing this is: for small x, the functions $y = x$ and $y = \sin x$ are essentially indistinguishable.

We include this special limit, along with three others, in the following theorem.

? Theorem: Special Limits

1. $\lim_{x\to 0}\frac{\sin x}{x} = 1$
2. $\lim_{x\to 0}\frac{\cos x - 1}{x} = 0$

A short word on how to interpret the latter three limits. We know that as x goes to 0, $\cos x$ goes to 1. So, in the second limit, both the numerator and denominator are approaching 0. However, since the limit is 0, we can interpret this as saying that "$\cos x$ is approaching 1 faster than x is approaching 0."

✓ Example ✓ 2.3.7 : Using Special Limits

Evaluate the following limit:

$$\lim_{x\to 0}\frac{\sin 2x}{3x}.$$

Solution

The answer here is not going to be 1 since the numbers in front of each x do not match. We need the bottom to be 3x. We can do this by multiplying the numerator and denominator by the fraction $\frac{1}{2x}$.

$$\lim_{x\to 0}\frac{\sin 2x}{3x}=\lim_{x\to 0}\frac{(\frac{1}{2x})\sin 2x}{(\frac{1}{2x})3x}=\lim_{x\to 0}\frac{\frac{\sin 2x}{2x}}{\frac{3}{2}}=\frac{2}{3}\lim_{x\to 0}\frac{\sin 2x}{2x}=\frac{2}{3}(1)=\frac{2}{3}.$$

2.3: One-Sided Limits is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **2.4: Continuity** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **2.2: The Limit of a Function** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **1.4: One Sided Limits** by Gregory Hartman et al. is licensed CC BY-NC 3.0. Original source: http://www.apexcalculus.com/.
- **1.3: Finding Limits Analytically** by Gregory Hartman et al. is licensed CC BY-NC 3.0. Original source: http://www.apexcalculus.com/.

2.4: Continuity

Learning Objectives

- Explain the three conditions for continuity at a point.
- Describe three kinds of discontinuities.
- Define continuity on an interval.
- State the theorem for limits of composite functions.
- Provide an example of the intermediate value theorem.

Many functions have the property that their graphs can be traced with a pencil without lifting the pencil from the page. Such functions are called *continuous*. Other functions have points at which a break in the graph occurs, but satisfy this property over intervals contained in their domains. They are continuous on these intervals and are said to have a *discontinuity* at a point where a break occurs.

We begin our investigation of continuity by exploring what it means for a function to have *continuity at a point*. Intuitively, a function is continuous at a particular point if there is no break in its graph at that point.

Continuity at a Point

Before we look at a formal definition of what it means for a function to be continuous at a point, let's consider various functions that fail to meet our intuitive notion of what it means to be continuous at a point. We then create a list of conditions that prevent such failures.

Our first function of interest is shown in Figure 2.4.1. We see that the graph of $f(x)$ has a hole at a. In fact, $f(a)$ is undefined. At the very least, for $f(x)$ to be continuous at a, we need the following condition:

i. $f(a)$ is defined

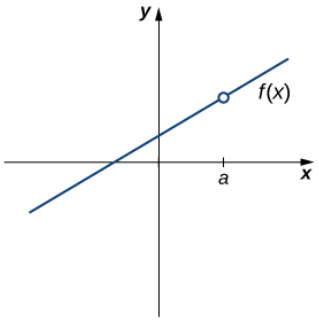


Figure 2.4.1: The function $f(x)$ is not continuous at a because $f(a)$ is undefined.

However, as we see in Figure 2.4.2, this condition alone is insufficient to guarantee continuity at the point a. Although $f(a)$ is defined, the function has a gap at a. In this example, the gap exists because $\lim_{x \to a} f(x)$ does not exist. We must add another condition for continuity at a—namely,

ii. $\lim_{x \to a} f(x)$ exists

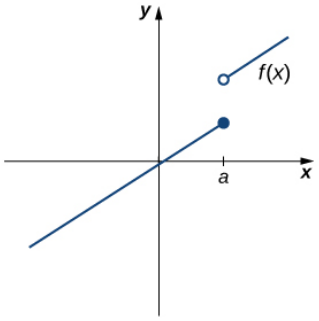


Figure 2.4.2: The function $f(x)$ is not continuous at a because $\lim_{x \to a} f(x)$ does not exist.

However, as we see in Figure 2.4.3, these two conditions by themselves do not guarantee continuity at a point. The function in this figure satisfies both of our first two conditions, but is still not continuous at a. We must add a third condition to our list:

iii. $\lim_{x \to a} f(x) = f(a)$

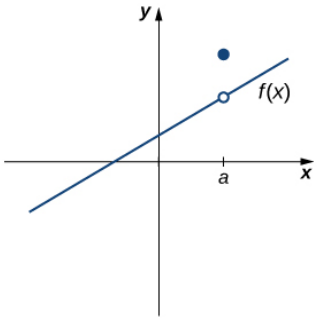


Figure 2.4.3: The function $f(x)$ is not continuous at a because $\lim_{x \to a} f(x) \neq f(a)$.

Now we put our list of conditions together and form a definition of continuity at a point.

Definition: Continuous at a Point

A function $f(x)$ is **continuous at a point a** if and only if the following three conditions are satisfied:

i. $f(a)$ is defined
ii. $\lim_{x \to a} f(x)$ exists
iii. $\lim_{x \to a} f(x) = f(a)$

A function is **discontinuous at a point a** if it fails to be continuous at a.

The following procedure can be used to analyze the continuity of a function at a point using this definition.

Problem-Solving Strategy: Determining Continuity at a Point

1. Check to see if $f(a)$ is defined. If $f(a)$ is undefined, we need go no further. The function is not continuous at a. If $f(a)$ is defined, continue to step 2.
2. Compute $\lim_{x \to a} f(x)$. In some cases, we may need to do this by first computing $\lim_{x \to a^-} f(x)$ and $\lim_{x \to a^+} f(x)$. If $\lim_{x \to a} f(x)$ does not exist (that is, it is not a real number), then the function is not continuous at a and the problem is solved. If $\lim_{x \to a} f(x)$ exists, then continue to step 3.
3. Compare $f(a)$ and $\lim_{x \to a} f(x)$. If $\lim_{x \to a} f(x) \neq f(a)$, then the function is not continuous at a. If $\lim_{x \to a} f(x) = f(a)$, then the function is continuous at a.

The next three examples demonstrate how to apply this definition to determine whether a function is continuous at a given point. These examples illustrate situations in which each of the conditions for continuity in the definition succeed or fail.

Example 2.4.1A: Determining Continuity at a Point, Condition 1

Using the definition, determine whether the function $f(x) = \dfrac{x^2 - 4}{x - 2}$ is continuous at $x = 2$. Justify the conclusion.

Solution

Let's begin by trying to calculate $f(2)$. We can see that $f(2) = 0/0$, which is undefined. Therefore, $f(x) = \dfrac{x^2 - 4}{x - 2}$ is discontinuous at 2 because $f(2)$ is undefined. The graph of $f(x)$ is shown in Figure 2.4.4.

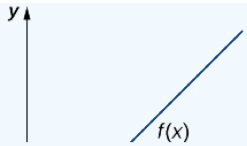


Figure 2.4.4: The function $f(x)$ is discontinuous at 2 because $f(2)$ is undefined.

Example 2.4.1B: Determining Continuity at a Point, Condition 2

Using the definition, determine whether the function $f(x) = \begin{cases} -x^2+4, & \text{if } x \leq 3 \\ 4x-8, & \text{if } x > 3 \end{cases}$ is continuous at $x = 3$. Justify the conclusion.

Solution

Let's begin by trying to calculate $f(3)$.

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

Thus, $f(3)$ is defined. Next, we calculate $\lim\limits_{x\to 3} f(x)$. To do this, we must compute $\lim\limits_{x\to 3^-} f(x)$ and $\lim\limits_{x\to 3^+} f(x)$:

$$\lim_{x\to 3^-} f(x) = -(3^2)+4 = -5$$

and

$$\lim_{x\to 3^+} f(x) = 4(3)-8 = 4 \, .$$

Therefore, $\lim\limits_{x\to 3} f(x)$ does not exist. Thus, $f(x)$ is not continuous at 3. The graph of $f(x)$ is shown in Figure 2.4.5.

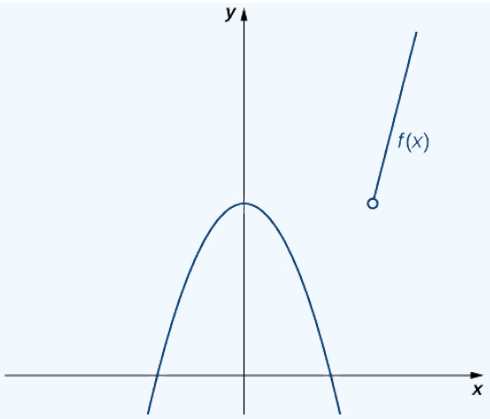


Figure 2.4.5: The function $f(x)$ is not continuous at 3 because $\lim_{x\to 3} f(x)$ does not exist.

Example 2.4.1C: Determining Continuity at a Point, Condition 3

Using the definition, determine whether the function $f(x)=\begin{cases}\frac{\sin x}{x}, & \text{if } x\neq 0\\ 1, & \text{if } x=0\end{cases}$ is continuous at $x=0$.

Solution

First, observe that

$$f(0)=1$$

Next,

$$\lim_{x\to 0} f(x)=\lim_{x\to 0}\frac{\sin x}{x}=1\,.$$

Last, compare $f(0)$ and $\lim_{x\to 0} f(x)$. We see that

$$f(0)=1=\lim_{x\to 0} f(x).$$

Since all three of the conditions in the definition of continuity are satisfied, $f(x)$ is continuous at $x=0$.

Exercise 2.4.1

Using the definition, determine whether the function $f(x)=\begin{cases}2x+1, & \text{if } x<1\\ 2, & \text{if } x=1\\ -x+4, & \text{if } x>1\end{cases}$ is continuous at $x=1$. If the function is not continuous at 1, indicate the condition for continuity at a point that fails to hold.

Hint

Check each condition of the definition.

Answer

f is not continuous at 1 because $f(1)=2\neq 3=\lim_{x\to 1} f(x)$.

By applying the definition of continuity and previously established theorems concerning the evaluation of limits, we can state the following theorem.

Theorem 2.4.1: Continuity of Polynomials and Rational Functions

Polynomials and rational functions are continuous at every point in their domains.

Proof

Previously, we showed that if $p(x)$ and $q(x)$ are polynomials, $\lim_{x \to a} p(x) = p(a)$ for every polynomial $p(x)$ and $\lim_{x \to a} \frac{p(x)}{q(x)} = \frac{p(a)}{q(a)}$ as long as $q(a) \neq 0$. Therefore, polynomials and rational functions are continuous on their domains.

□

We now apply Theorem 2.4.1 to determine the points at which a given rational function is continuous.

Example 2.4.2:Continuity of a Rational Function

For what values of x is $f(x) = \frac{x+1}{x-5}$ continuous?

Solution

The rational function $f(x) = \frac{x+1}{x-5}$ is continuous for every value of x except $x = 5$.

Exercise 2.4.2

For what values of x is $f(x) = 3x^4 - 4x^2$ continuous?

Hint

Use the Continuity of Polynomials and Rational Functions stated above.

Answer

$f(x)$ is continuous at every real number.

Types of Discontinuities

As we have seen in Example 2.4.1A and Example 2.4.1B, discontinuities take on several different appearances. We classify the types of discontinuities we have seen thus far as removable discontinuities, infinite discontinuities, or jump discontinuities. Intuitively, a **removable discontinuity** is a discontinuity for which there is a hole in the graph, a **jump discontinuity** is a noninfinite discontinuity for which the sections of the function do not meet up, and an **infinite discontinuity** is a discontinuity located at a vertical asymptote. Figure 2.4.6 illustrates the differences in these types of discontinuities. Although these terms provide a handy way of describing three common types of discontinuities, keep in mind that not all discontinuities fit neatly into these categories.

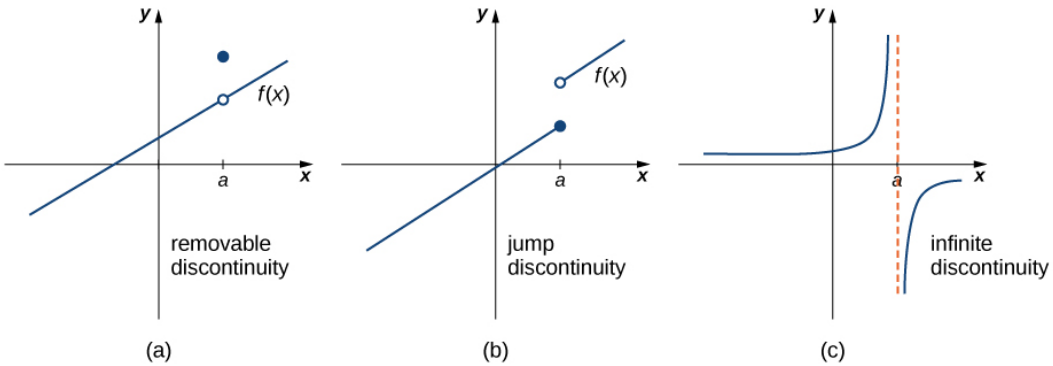


Figure 2.4.6: Discontinuities are classified as (a) removable, (b) jump, or (c) infinite.

These three discontinuities are formally defined as follows:

Definition

If $f(x)$ is discontinuous at a, then

1. f has a **removable discontinuity** at a if $\lim\limits_{x\to a} f(x)$ exists. (Note: When we state that $\lim\limits_{x\to a} f(x)$ exists, we mean that $\lim\limits_{x\to a} f(x) = L$, where L is a real number.)
2. f has a **jump discontinuity** at a if $\lim\limits_{x\to a^-} f(x)$ and $\lim\limits_{x\to a^+} f(x)$ both exist, but $\lim\limits_{x\to a^-} f(x) \neq lim_{x\to a^+} f(x)$. (Note: When we state that $\lim\limits_{x\to a^-} f(x)$ and $\lim\limits_{x\to a^+} f(x)$ both exist, we mean that both are real-valued and that neither take on the values $\pm\infty$.)
3. f has an **infinite discontinuity** at a if $\lim\limits_{x\to a^-} f(x) = \pm\infty$ or $\lim\limits_{x\to a^+} f(x) = \pm\infty$.

Example 2.4.3: Classifying a Discontinuity

In Example 2.4.1A, we showed that $f(x) = \dfrac{x^2-4}{x-2}$ is discontinuous at $x = 2$. Classify this discontinuity as removable, jump, or infinite.

Solution

To classify the discontinuity at 2 we must evaluate $\lim\limits_{x\to 2} f(x)$:

$$\begin{aligned}\lim_{x\to 2} f(x) &= \lim_{x\to 2} \frac{x^2-4}{x-2} \\ &= \lim_{x\to 2} \frac{(x-2)(x+2)}{x-2} \\ &= \lim_{x\to 2} (x+2) \\ &= 4.\end{aligned}$$

Since f is discontinuous at 2 and $\lim\limits_{x\to 2} f(x)$ exists, f has a removable discontinuity at $x = 2$.

Example 2.4.4: Classifying a Discontinuity

In Example 2.4.1B, we showed that $f(x) = \begin{cases} -x^2+4, & \text{if } x \leq 3 \\ 4x-8, & \text{if } x > 3 \end{cases}$ is discontinuous at $x = 3$. Classify this discontinuity as removable, jump, or infinite.

Solution

Earlier, we showed that f is discontinuous at 3 because $\lim\limits_{x\to 3} f(x)$ does not exist. However, since $\lim\limits_{x\to 3^-} f(x) = -5$ and $\lim\limits_{x\to 3^+} f(x) = 4$ both exist, we conclude that the function has a jump discontinuity at 3.

Example 2.4.5: Classifying a Discontinuity

Determine whether $f(x) = \dfrac{x+2}{x+1}$ is continuous at -1. If the function is discontinuous at -1, classify the discontinuity as removable, jump, or infinite.

Solution

The function value $f(-1)$ is undefined. Therefore, the function is not continuous at -1. To determine the type of discontinuity, we must determine the limit at -1. We see that $\lim\limits_{x\to -1^-} \dfrac{x+2}{x+1} = -\infty$ and $\lim\limits_{x\to -1^+} \dfrac{x+2}{x+1} = +\infty$. Therefore, the function has an infinite discontinuity at -1.

Exercise 2.4.3

For $f(x)=\begin{cases} x^2, & \text{if } x \neq 1 \\ 3, & \text{if } x=1 \end{cases}$, decide whether f is continuous at 1. If f is not continuous at 1, classify the discontinuity as removable, jump, or infinite.

Hint

Consider the definitions of the various kinds of discontinuity stated above. If the function is discontinuous at 1, look at $\lim_{x\to 1} f(x)$

Answer

Discontinuous at 1; removable

Continuity over an Interval

Now that we have explored the concept of continuity at a point, we extend that idea to **continuity over an interval**. As we develop this idea for different types of intervals, it may be useful to keep in mind the intuitive idea that a function is continuous over an interval if we can use a pencil to trace the function between any two points in the interval without lifting the pencil from the paper. In preparation for defining continuity on an interval, we begin by looking at the definition of what it means for a function to be continuous from the right at a point and continuous from the left at a point.

Definition: Continuity from the Right and from the Left

A function $f(x)$ is said to be **continuous from the right** at a if $\lim_{x\to a^+} f(x)=f(a)$.

A function $f(x)$ is said to be **continuous from the left** at a if $\lim_{x\to a^-} f(x)=f(a)$

A function is continuous over an open interval if it is continuous at every point in the interval. A function $f(x)$ is continuous over a closed interval of the form $[a,b]$ if it is continuous at every point in (a,b) and is continuous from the right at a and is continuous from the left at b. Analogously, a function $f(x)$ is continuous over an interval of the form $(a,b]$ if it is continuous over (a,b) and is continuous from the left at b. Continuity over other types of intervals are defined in a similar fashion.

Requiring that $\lim_{x\to a^+} f(x)=f(a)$ and $\lim_{x\to b^-} f(x)=f(b)$ ensures that we can trace the graph of the function from the point $(a,f(a))$ to the point $(b,f(b))$ without lifting the pencil. If, for example, $\lim_{x\to a^+} f(x)\neq f(a)$, we would need to lift our pencil to jump from $f(a)$ to the graph of the rest of the function over $(a,b]$.

Example 2.4.6: Continuity on an Interval

State the interval(s) over which the function $f(x)=\dfrac{x-1}{x^2+2x}$ is continuous.

Solution

Since $f(x)=\dfrac{x-1}{x^2+2x}$ is a rational function, it is continuous at every point in its domain. The domain of $f(x)$ is the set $(-\infty,-2)\cup(-2,0)\cup(0,+\infty)$. Thus, $f(x)$ is continuous over each of the intervals $(-\infty,-2)$, $(-2,0)$, and $(0,+\infty)$.

Example 2.4.7: Continuity over an Interval

State the interval(s) over which the function $f(x)=\sqrt{4-x^2}$ is continuous.

Solution

From the limit laws, we know that $\lim_{x\to a}\sqrt{4-x^2}=\sqrt{4-a^2}$ for all values of a in $(-2,2)$. We also know that $\lim_{x\to -2^+}\sqrt{4-x^2}=0$ exists and $\lim_{x\to 2^-}\sqrt{4-x^2}=0$ exists. Therefore, $f(x)$ is continuous over the interval $[-2,2]$.

Exercise 2.4.4

State the interval(s) over which the function $f(x)=\sqrt{x+3}$ is continuous.

Hint

Use Example 2.4.7as a guide.

Answer

$[-3,+\infty)$

Theorem 2.4.2 allows us to expand our ability to compute limits. In particular, this theorem ultimately allows us to demonstrate that trigonometric functions are continuous over their domains.

Theorem 2.4.2: Composite Function Theorem

If $f(x)$ is continuous at L and $\lim\limits_{x\to a} g(x)=L$, then

$$\lim_{x\to a} f\big(g(x)\big)=f\big(\lim_{x\to a} g(x)\big)=f(L).$$

Before we move on to Example 2.4.8, recall that earlier, in the section on limit laws, we showed $\lim\limits_{x\to 0}\cos x=1=\cos(0)$. Consequently, we know that $f(x)=\cos x$ is continuous at 0. In Example 2.4.8, we see how to combine this result with the composite function theorem.

Example 2.4.8: Limit of a Composite Cosine Function

Evaluate $\lim\limits_{x\to \pi/2}\cos\left(x-\dfrac{\pi}{2}\right)$.

Solution

The given function is a composite of $\cos x$ and $x-\frac{\pi}{2}$. Since $\lim\limits_{x\to \pi/2}\left(x-\dfrac{\pi}{2}\right)=0$ and $\cos x$ is continuous at 0, we may apply the composite function theorem. Thus,

$$\lim_{x\to \pi/2}\cos\left(x-\frac{\pi}{2}\right)=\cos\left(\lim_{x\to \pi/2}\left(x-\frac{\pi}{2}\right)\right)=\cos(0)=1.$$

Exercise 2.4.4:

Evaluate $\lim\limits_{x\to \pi}\sin(x-\pi)$.

Hint

$f(x)=\sin x$ is continuous at 0. Use Example 2.4.8as a guide.

Answer

0

The proof of the next theorem uses the composite function theorem as well as the continuity of $f(x)=\sin x$ and $g(x)=\cos x$ at the point 0 to show that trigonometric functions are continuous over their entire domains.

Theorem 2.4.3: Continuity of Trigonometric Functions

Trigonometric functions are continuous over their entire domains.

Proof

We begin by demonstrating that $\cos x$ is continuous at every real number. To do this, we must show that $\lim\limits_{x\to a}\cos x=\cos a$ for all values of a.

$$\begin{aligned}
\lim_{x\to a}\cos x &= \lim_{x\to a}\cos((x-a)+a) && \text{Rewrite } x = x-a+a.\\
&= \lim_{x\to a}(\cos(x-a)\cos a - \sin(x-a)\sin a) && \text{Apply the identity for the cosine of the sum of two angles.}\\
&= \cos(\lim_{x\to a}(x-a))\cos a - \sin(\lim_{x\to a}(x-a))\sin a && \text{Since } \lim_{x\to a}(x-a)=0, \text{ and } \sin x \text{ and } \cos x \text{ are continuous at } 0.\\
&= \cos(0)\cos a - \sin(0)\sin a && \text{Evaluate } \cos(0) \text{ and } \sin(0) \text{ and simplify.}\\
&= 1\cdot\cos a - 0\cdot\sin a = \cos a.
\end{aligned}$$

The proof that $\sin x$ is continuous at every real number is analogous. Because the remaining trigonometric functions may be expressed in terms of $\sin x$ and $\cos x$, their continuity follows from the quotient limit law.

□

As you can see, the composite function theorem is invaluable in demonstrating the continuity of trigonometric functions. As we continue our study of calculus, we revisit this theorem many times.

The Intermediate Value Theorem

Functions that are continuous over intervals of the form $[a, b]$, where a and b are real numbers, exhibit many useful properties. Throughout our study of calculus, we will encounter many powerful theorems concerning such functions. The first of these theorems is the **Intermediate Value Theorem.**

The Intermediate Value Theorem

Let f be continuous over a closed, bounded interval $[a, b]$. If z is any real number between $f(a)$ and $f(b)$, then there is a number c in $[a, b]$ satisfying $f(c) = z$ in Figure 2.4.7.

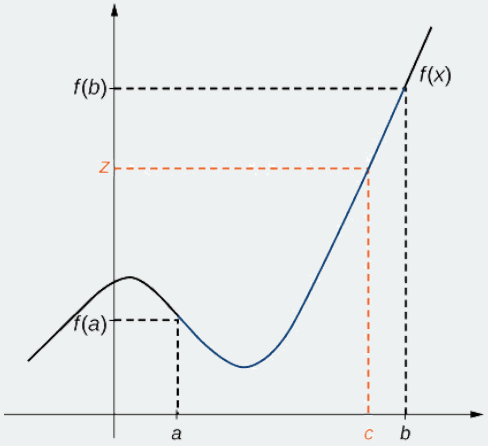


Figure 2.4.7: There is a number $c \in [a, b]$ that satisfies $f(c) = z$.

Example 2.4.9: Application of the Intermediate Value Theorem

Show that $f(x) = x - \cos x$ has at least one zero.

Solution

Since $f(x) = x - \cos x$ is continuous over $(-\infty, +\infty)$, it is continuous over any closed interval of the form $[a, b]$. If you can find an interval $[a, b]$ such that $f(a)$ and $f(b)$ have opposite signs, you can use the Intermediate Value Theorem to conclude there must be a real number c in (a, b) that satisfies $f(c) = 0$. Note that

$$f(0) = 0 - \cos(0) = -1 < 0$$

and

$$f(\tfrac{\pi}{2}) = \tfrac{\pi}{2} - \cos\tfrac{\pi}{2} = \tfrac{\pi}{2} > 0\ .$$

Using the Intermediate Value Theorem, we can see that there must be a real number c in $[0, \pi/2]$ that satisfies $f(c) = 0$. Therefore, $f(x) = x - \cos x$ has at least one zero.

Example 2.4.10: When Can You Apply the Intermediate Value Theorem?

If $f(x)$ is continuous over $[0,2]$, $f(0)>0$ and $f(2)>0$, can we use the Intermediate Value Theorem to conclude that $f(x)$ has no zeros in the interval $[0,2]$? Explain.

Solution

No. The Intermediate Value Theorem only allows us to conclude that we can find a value between $f(0)$ and $f(2)$; it doesn't allow us to conclude that we can't find other values. To see this more clearly, consider the function $f(x)=(x-1)^2$. It satisfies $f(0)=1>0, f(2)=1>0$, and $f(1)=0$.

Example 2.4.11: When Can You Apply the Intermediate Value Theorem?

For $f(x)=1/x, f(-1)=-1<0$ and $f(1)=1>0$. Can we conclude that $f(x)$ has a zero in the interval $[-1,1]$?

Solution

No. The function is not continuous over $[-1,1]$. The Intermediate Value Theorem does not apply here.

Exercise 2.4.5

Show that $f(x)=x^3-x^2-3x+1$ has a zero over the interval $[0,1]$.

Hint

Find $f(0)$ and $f(1)$. Apply the Intermediate Value Theorem.

Answer

$f(0)=1>0,\ f(1)=-2<0;\ f(x)$ is continuous over $[0,1]$. It must have a zero on this interval.

Key Concepts

- For a function to be continuous at a point, it must be defined at that point, its limit must exist at the point, and the value of the function at that point must equal the value of the limit at that point.
- Discontinuities may be classified as removable, jump, or infinite.
- A function is continuous over an open interval if it is continuous at every point in the interval. It is continuous over a closed interval if it is continuous at every point in its interior and is continuous at its endpoints.
- The composite function theorem states: If $f(x)$ is continuous at L and $\lim\limits_{x\to a} g(x)=L$, then $\lim\limits_{x\to a} f\big(g(x)\big)=f\big(\lim\limits_{x\to a} g(x)\big)=f(L)$.
- The Intermediate Value Theorem guarantees that if a function is continuous over a closed interval, then the function takes on every value between the values at its endpoints.

Glossary

continuity at a point

A function $f(x)$ is continuous at a point a if and only if the following three conditions are satisfied: (1) $f(a)$ is defined, (2) $\lim\limits_{x\to a} f(x)$ exists, and (3) $\lim x\to af(x)=f(a)$

continuity from the left

A function is continuous from the left at b if $\lim\limits_{x\to b^-} f(x)=f(b)$

continuity from the right

A function is continuous from the right at a if $\lim\limits_{x\to a^+} f(x)=f(a)$

continuity over an interval

a function that can be traced with a pencil without lifting the pencil; a function is continuous over an open interval if it is continuous at every point in the interval; a function $f(x)$ is continuous over a closed interval of the form $[a,b]$ if it is continuous at every point in (a,b), and it is continuous from the right at a and from the left at b

discontinuity at a point

A function is discontinuous at a point or has a discontinuity at a point if it is not continuous at the point

infinite discontinuity

An infinite discontinuity occurs at a point a if $\lim\limits_{x \to a^-} f(x) = \pm\infty$ or $\lim\limits_{x \to a^+} f(x) = \pm\infty$

Intermediate Value Theorem

Let f be continuous over a closed bounded interval $[a, b]$ if z is any real number between $f(a)$ and $f(b)$, then there is a number c in $[a, b]$ satisfying $f(c) = z$

jump discontinuity

A jump discontinuity occurs at a point a if $\lim\limits_{x \to a^-} f(x)$ and $\lim\limits_{x \to a^+} f(x)$ both exist, but $\lim\limits_{x \to a^-} f(x) \neq \lim\limits_{x \to a^+} f(x)$

removable discontinuity

A removable discontinuity occurs at a point a if $f(x)$ is discontinuous at a, but $\lim\limits_{x \to a} f(x)$ exists

---

2.4: Continuity is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **2.4: Continuity** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

2.5: Limits Involving Infinity; Asymptotes of Graphs

Learning Objectives

- Calculate the limit of a function as x increases or decreases without bound.
- Using correct notation, describe an infinite limit.
- Define a vertical asymptote.
- Recognize a horizontal asymptote on the graph of a function.
- Estimate the end behavior of a function as x increases or decreases without bound.
- Recognize an oblique asymptote on the graph of a function.

Infinite Limits

Evaluating the limit of a function at a point or evaluating the limit of a function from the right and left at a point helps us to characterize the behavior of a function around a given value. As we shall see, we can also describe the behavior of functions that do not have finite limits.

We now turn our attention to $h(x) = 1/(x-2)^2$, the third and final function introduced at the beginning of this section (see Figure 2.5.1(c)). From its graph we see that as the values of x approach 2, the values of $h(x) = 1/(x-2)^2$ become larger and larger and, in fact, become infinite. Mathematically, we say that the limit of $h(x)$ as x approaches 2 is positive infinity. Symbolically, we express this idea as

$$\lim_{x \to 2} h(x) = +\infty.$$

More generally, we define *infinite limits* as follows:

Definitions: Infinite Limits

We define three types of *infinite limits*.

Infinite limits from the left: Let $f(x)$ be a function defined at all values in an open interval of the form (b, a).

i. If the values of $f(x)$ increase without bound as the values of x (where $x < a$) approach the number a, then we say that the limit as x approaches a from the left is positive infinity and we write

$$\lim_{x \to a^-} f(x) = +\infty.$$

ii. If the values of $f(x)$ decrease without bound as the values of x (where $x < a$) approach the number a, then we say that the limit as x approaches a from the left is negative infinity and we write

$$\lim_{x \to a^-} f(x) = -\infty.$$

Infinite limits from the right: Let $f(x)$ be a function defined at all values in an open interval of the form (a, c).

i. If the values of $f(x)$ increase without bound as the values of x (where $x > a$) approach the number a, then we say that the limit as x approaches a from the right is positive infinity and we write

$$\lim_{x \to a^+} f(x) = +\infty.$$

ii. If the values of $f(x)$ decrease without bound as the values of x (where $x > a$) approach the number a, then we say that the limit as x approaches a from the right is negative infinity and we write

$$\lim_{x \to a^+} f(x) = -\infty.$$

Two-sided infinite limit: Let $f(x)$ be defined for all $x \neq a$ in an open interval containing a

i. If the values of $f(x)$ increase without bound as the values of x (where $x \neq a$) approach the number a, then we say that the limit as x approaches a is positive infinity and we write

$$\lim_{x \to a} f(x) = +\infty.$$

ii. If the values of $f(x)$ decrease without bound as the values of x (where $x \neq a$) approach the number a, then we say that the limit as x approaches a is negative infinity and we write

$$\lim_{x \to a} f(x) = -\infty.$$

It is important to understand that when we write statements such as $\lim_{x \to a} f(x) = +\infty$ or $\lim_{x \to a} f(x) = -\infty$ we are describing the behavior of the function, as we have just defined it. We are not asserting that a limit exists. For the limit of a function $f(x)$ to exist at a, it must approach a real number L as x approaches a. That said, if, for example, $\lim_{x \to a} f(x) = +\infty$, we always write $\lim_{x \to a} f(x) = +\infty$ rather than $\lim_{x \to a} f(x)$ DNE.

Example 2.5.1: Recognizing an Infinite Limit

Evaluate each of the following limits, if possible. Use a table of functional values and graph $f(x) = 1/x$ to confirm your conclusion.

a. $\lim_{x \to 0^-} \dfrac{1}{x}$
b. $\lim_{x \to 0^+} \dfrac{1}{x}$
c. $\lim_{x \to 0} \dfrac{1}{x}$

Solution

Begin by constructing a table of functional values.

Table 2.5.7

x	$\frac{1}{x}$	x	$\frac{1}{x}$
-0.1	-10	0.1	10
-0.01	-100	0.01	100
-0.001	-1000	0.001	1000
-0.0001	-10,000	0.0001	10,000
-0.00001	-100,000	0.00001	100,000
-0.000001	-1,000,000	0.000001	1,000,000

a. The values of $1/x$ decrease without bound as x approaches 0 from the left. We conclude that

$$\lim_{x \to 0^-} \frac{1}{x} = -\infty.$$

b. The values of $1/x$ increase without bound as x approaches 0 from the right. We conclude that

$$\lim_{x \to 0^+} \frac{1}{x} = +\infty.$$

c. Since $\lim_{x \to 0^-} \dfrac{1}{x} = -\infty$ and $\lim_{x \to 0^+} \dfrac{1}{x} = +\infty$ have different values, we conclude that

$$\lim_{x \to 0} \frac{1}{x} \quad \text{DNE.}$$

The graph of $f(x) = 1/x$ in Figure 2.5.8 confirms these conclusions.

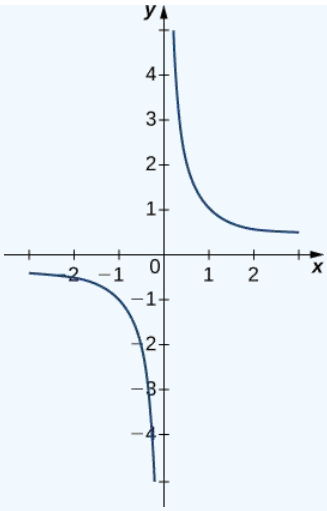


Figure 2.5.8: The graph of $f(x) = 1/x$ confirms that the limit as x approaches 0 does not exist.

? Exercise 2.5.1:

Evaluate each of the following limits, if possible. Use a table of functional values and graph $f(x) = 1/x^2$ to confirm your conclusion.

a. $\lim\limits_{x \to 0^-} \dfrac{1}{x^2}$

b. $\lim\limits_{x \to 0^+} \dfrac{1}{x^2}$

c. $\lim\limits_{x \to 0} \dfrac{1}{x^2}$

Hint

Follow the procedures from Example 2.5.5

Answer

a. $\lim\limits_{x \to 0^-} \dfrac{1}{x^2} = +\infty$;

b. $\lim\limits_{x \to 0^+} \dfrac{1}{x^2} = +\infty$;

c. $\lim\limits_{x \to 0} \dfrac{1}{x^2} = +\infty$

It is useful to point out that functions of the form $f(x) = 1/(x-a)^n$, where n is a positive integer, have infinite limits as x approaches a from either the left or right (Figure 2.5.9). These limits are summarized in the above definitions.

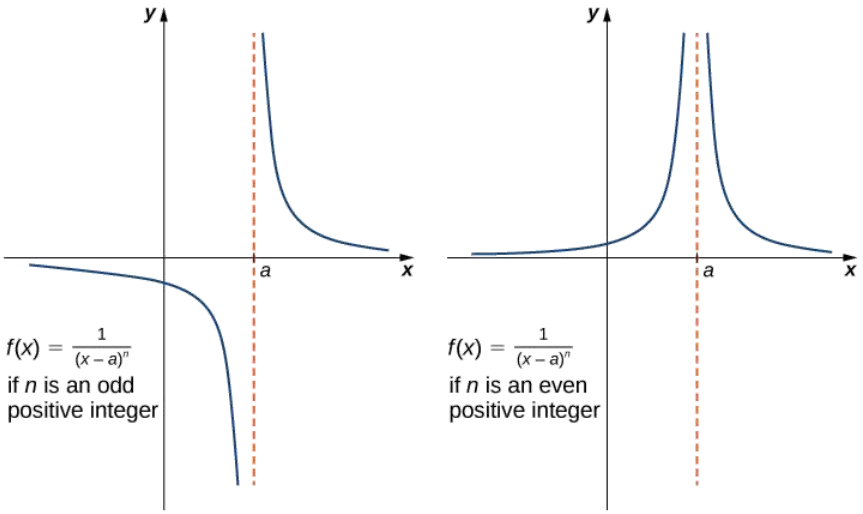


Figure 2.5.9: The function $f(x) = 1/(x-a)^n$ has infinite limits at a.

Infinite Limits from Positive Integers

If n is a positive even integer, then

$$\lim_{x \to a} \frac{1}{(x-a)^n} = +\infty. \tag{2.5.1}$$

If n is a positive odd integer, then

$$\lim_{x \to a^+} \frac{1}{(x-a)^n} = +\infty \tag{2.5.2}$$

and

$$\lim_{x \to a^-} \frac{1}{(x-a)^n} = -\infty. \tag{2.5.3}$$

We should also point out that in the graphs of $f(x) = 1/(x-a)^n$, points on the graph having x-coordinates very near to a are very close to the vertical line $x = a$. That is, as x approaches a, the points on the graph of $f(x)$ are closer to the line $x = a$. The line $x = a$ is called a vertical asymptote of the graph. We formally define a *vertical asymptote* as follows:

Definition: Vertical Asymptotes

Let $f(x)$ be a function. If any of the following conditions hold, then the line $x = a$ is a vertical asymptote of $f(x)$.

$$\lim_{x \to a^-} f(x) = +\infty$$

$$\lim_{x \to a^-} f(x) = -\infty$$

$$\lim_{x \to a^+} f(x) = +\infty$$

$$\lim_{x \to a^+} f(x) = -\infty$$

$$\lim_{x \to a} f(x) = +\infty$$

$$\lim_{x \to a} f(x) = -\infty$$

Example 2.5.2: Finding a Vertical Asymptote

Evaluate each of the following limits using Equations 2.5.1, 2.5.2, and 2.5.3 above. Identify any vertical asymptotes of the function $f(x) = 1/(x+3)^4$.

a. $\lim\limits_{x \to -3^-} \dfrac{1}{(x+3)^4}$
b. $\lim\limits_{x \to -3^+} \dfrac{1}{(x+3)^4}$
c. $\lim\limits_{x \to -3} \dfrac{1}{(x+3)^4}$

Solution

We can use the above equations directly.

a. $\lim\limits_{x \to -3^-} \dfrac{1}{(x+3)^4} = +\infty$
b. $\lim\limits_{x \to -3^+} \dfrac{1}{(x+3)^4} = +\infty$
c. $\lim\limits_{x \to -3} \dfrac{1}{(x+3)^4} = +\infty$

The function $f(x) = 1/(x+3)^4$ has a vertical asymptote of $x = -3$.

Exercise 2.5.2:

Evaluate each of the following limits. Identify any vertical asymptotes of the function $f(x) = \dfrac{1}{(x-2)^3}$.

a. $\lim\limits_{x \to 2^-} \dfrac{1}{(x-2)^3}$
b. $\lim\limits_{x \to 2^+} \dfrac{1}{(x-2)^3}$
c. $\lim\limits_{x \to 2} \dfrac{1}{(x-2)^3}$

Answer a

$\lim\limits_{x \to 2^-} \dfrac{1}{(x-2)^3} = -\infty$

Answer b

$\lim\limits_{x \to 2^+} \dfrac{1}{(x-2)^3} = +\infty$

Answer c

$\lim\limits_{x \to 2} \dfrac{1}{(x-2)^3}$ DNE. The line $x = 2$ is the vertical asymptote of $f(x) = 1/(x-2)^3$.

In the next example we put our knowledge of various types of limits to use to analyze the behavior of a function at several different points.

Example 2.5.3: Behavior of a Function at Different Points

Use the graph of $f(x)$ in Figure 2.5.10 to determine each of the following values:

a. $\lim\limits_{x \to -4^-} f(x)$; $\lim\limits_{x \to -4^+} f(x)$; $\lim\limits_{x \to -4} f(x)$; $f(-4)$
b. $\lim\limits_{x \to -2^-} f(x)$; $\lim\limits_{x \to -2^+} f(x)$; $\lim\limits_{x \to -2} f(x)$; $f(-2)$

c. $\lim\limits_{x\to 1^-} f(x)$; $\lim\limits_{x\to 1^+} f(x)$; $\lim\limits_{x\to 1} f(x)$; $f(1)$

d. $\lim\limits_{x\to 3^-} f(x)$; $\lim\limits_{x\to 3^+} f(x)$; $\lim\limits_{x\to 3} f(x)$; $f(3)$

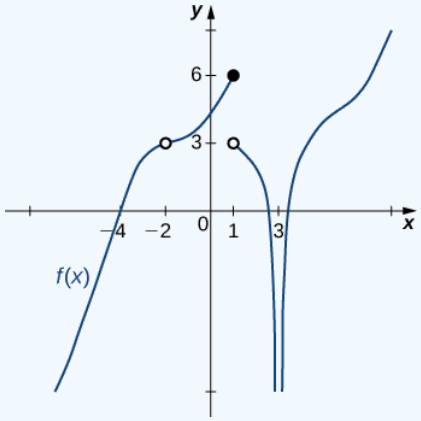


Figure 2.5.10: The graph shows $f(x)$.

Solution

Using the definitions above and the graph for reference, we arrive at the following values:

a. $\lim\limits_{x\to -4^-} f(x) = 0$; $\lim\limits_{x\to -4^+} f(x) = 0$; $\lim\limits_{x\to -4} f(x) = 0$; $f(-4) = 0$

b. $\lim\limits_{x\to -2^-} f(x) = 3$; $\lim\limits_{x\to -2^+} f(x) = 3$; $\lim\limits_{x\to -2} f(x) = 3$; $f(-2)$ is undefined

c. $\lim\limits_{x\to 1^-} f(x) = 6$; $\lim\limits_{x\to 1^+} f(x) = 3$; $\lim\limits_{x\to 1} f(x)$ DNE; $f(1) = 6$

d. $\lim\limits_{x\to 3^-} f(x) = -\infty$; $\lim\limits_{x\to 3^+} f(x) = -\infty$; $\lim\limits_{x\to 3} f(x) = -\infty$; $f(3)$ is undefined

? Exercise 2.5.3:

Evaluate $\lim\limits_{x\to 1} f(x)$ for $f(x)$ shown here:

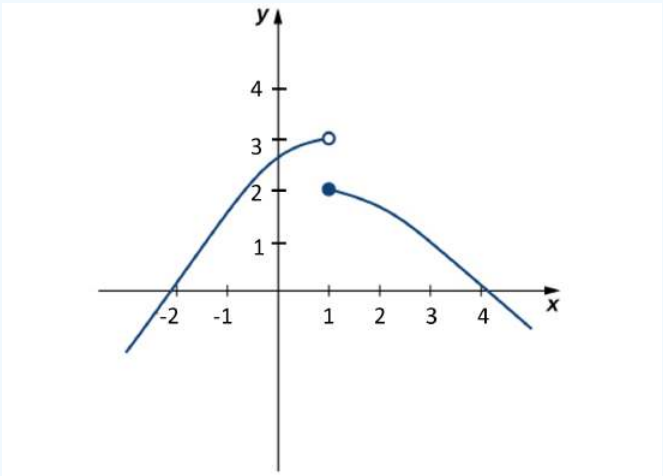


Figure 2.5.11. The graph of a piecewise function f.

Hint

Compare the limit from the right with the limit from the left.

Answer

$\lim\limits_{x\to 1} f(x)$ does not exist

Vertical Asymptotes

If the limit of $f(x)$ as x approaches c from either the left or right (or both) is ∞ or $-\infty$, we say the function has a **vertical asymptote** at c.

✔ Example ✔ 2.5.4: Finding vertical asymptotes

Find the vertical asymptotes of $f(x) = \dfrac{3x}{x^2-4}$.

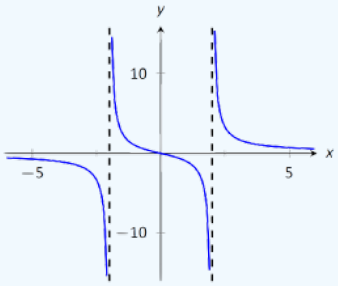


FIGURE 1.33: Graphing $f(x) = \frac{3x}{x^2-4}$.

Vertical asymptotes occur where the function grows without bound; this can occur at values of c where the denominator is 0. When x is near c, the denominator is small, which in turn can make the function take on large values. In the case of the given function, the denominator is 0 at $x = \pm 2$. Substituting in values of x close to 2 and -2 seems to indicate that the function tends toward ∞ or $-\infty$ at those points. We can graphically confirm this by looking at Figure 1.33. Thus the vertical asymptotes are at $x = \pm 2$.

When a rational function has a vertical asymptote at $x = c$, we can conclude that the denominator is 0 at $x = c$. However, just because the denominator is 0 at a certain point does not mean there is a vertical asymptote there. For instance, $f(x) = (x^2-1)/(x-1)$ does not have a vertical asymptote at $x = 1$, as shown in Figure 1.34. While the denominator does get small near $x = 1$, the numerator gets small too, matching the denominator step for step. In fact, factoring the numerator, we get

$$f(x) = \frac{(x-1)(x+1)}{x-1}. \tag{2.5.4}$$

Canceling the common term, we get that $f(x) = x+1$ for $x \neq 1$. So there is clearly no asymptote, rather a hole exists in the graph at $x = 1$.

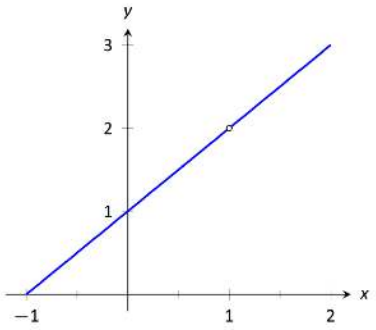


FIGURE 1.34: Graphically showing that $f(x) = \frac{x^2-1}{x-1}$ does not have an asymptote at $x = 1$.

The above example may seem a little contrived. Another example demonstrating this important concept is $f(x) = (\sin x)/x$. We have considered this function several times in the previous sections. We found that $\lim\limits_{x\to 0} \frac{\sin x}{x} = 1$; i.e., there is no vertical asymptote. No simple algebraic cancellation makes this fact obvious; we used the Squeeze Theorem in Section 1.3 to prove this.

If the denominator is 0 at a certain point but the numerator is not, then there will usually be a vertical asymptote at that point. On the other hand, if the numerator and denominator are both zero at that point, then there may or may not be a vertical asymptote at

that point. This case where the numerator and denominator are both zero returns us to an important topic.

Indeterminate Forms

We have seen how the limits

$$\lim_{x\to 0}\frac{\sin x}{x} \quad \text{and} \quad \lim_{x\to 1}\frac{x^2-1}{x-1} \tag{2.5.5}$$

each return the indeterminate form "$0/0$" when we blindly plug in $x=0$ and $x=1$, respectively. However, $0/0$ is not a valid arithmetical expression. It gives no indication that the respective limits are 1 and 2.

With a little cleverness, one can come up $0/0$ expressions which have a limit of ∞, 0, or any other real number. That is why this expression is called *indeterminate*.

A key concept to understand is that such limits do not really return $0/0$. Rather, keep in mind that we are taking *limits*. What is really happening is that the numerator is shrinking to 0 while the denominator is also shrinking to 0. The respective rates at which they do this are very important and determine the actual value of the limit.

An indeterminate form indicates that one needs to do more work in order to compute the limit. That work may be algebraic (such as factoring and canceling) or it may require a tool such as the Squeeze Theorem. In a later section we will learn a technique called l'Hospital's Rule that provides another way to handle indeterminate forms.

Some other common indeterminate forms are $\infty-\infty$, $\infty\cdot 0$, ∞/∞, 0^0, ∞^0 and 1^∞. Again, keep in mind that these are the "blind" results of evaluating a limit, and each, in and of itself, has no meaning. The expression $\infty-\infty$ does not really mean "subtract infinity from infinity." Rather, it means "One quantity is subtracted from the other, but both are growing without bound." What is the result? It is possible to get every value between $-\infty$ and ∞

Note that $1/0$ and $\infty/0$ are not indeterminate forms, though they are not exactly valid mathematical expressions, either. In each, the function is growing without bound, indicating that the limit will be ∞, $-\infty$, or simply not exist if the left- and right-hand limits do not match.

Limits at Infinity and Horizontal Asymptotes

At the beginning of this section we briefly considered what happens to $f(x)=1/x^2$ as x grew very large. Graphically, it concerns the behavior of the function to the "far right" of the graph. We make this notion more explicit in the following definition.

Definition: Limits at Infinity and Horizontal Asymptote

1. We say $\lim\limits_{x\to\infty} f(x)=L$ if for every $\epsilon>0$ there exists $M>0$ such that if $x\geq M$, then $|f(x)-L|<\epsilon$.
2. We say $\lim\limits_{x\to-\infty} f(x)=L$ if for every $\epsilon>0$ there exists $M<0$ such that if $x\leq M$, then $|f(x)-L|<\epsilon$.
3. If $\lim\limits_{x\to\infty} f(x)=L$ or $\lim\limits_{x\to-\infty} f(x)=L$, we say that $y=L$ is a **horizontal asymptote** of f.

We can also define limits such as $\lim\limits_{x\to\infty} f(x)=\infty$ by combining this definition with Definition 5.

Example 2.5.5: Approximating horizontal asymptotes

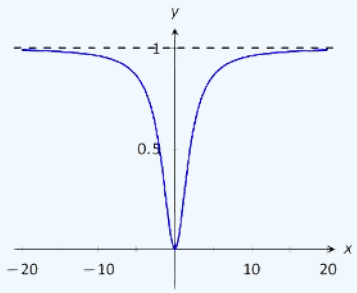
Approximate the horizontal asymptote(s) of $f(x)=\frac{x^2}{x^2+4}$.

Solution

We will approximate the horizontal asymptotes by approximating the limits

$$\lim_{x\to-\infty}\frac{x^2}{x^2+4} \quad \text{and} \quad \lim_{x\to\infty}\frac{x^2}{x^2+4}. \tag{2.5.6}$$

Figure 1.35(a) shows a sketch of f, and part (b) gives values of $f(x)$ for large magnitude values of x. It seems reasonable to conclude from both of these sources that f has a horizontal asymptote at $y=1$.



(a)

x	$f(x)$
10	0.9615
100	0.9996
10000	0.999996
-10	0.9615
-100	0.9996
-10000	0.999996

(b)

FIGURE 1.35: Using a graph and a table to approximate a horizontal asymptote in Example 29.

Later, we will show how to determine this analytically.

Horizontal asymptotes can take on a variety of forms. Figure 1.36(a) shows that $f(x) = x/(x^2+1)$ has a horizontal asymptote of $y = 0$, where 0 is approached from both above and below.

Figure 1.36(b) shows that $f(x) = x/\sqrt{x^2+1}$ has two horizontal asymptotes; one at $y = 1$ and the other at $y = -1$.

Figure 1.36(c) shows that $f(x) = (\sin x)/x$ has even more interesting behavior than at just $x = 0$; as x approaches $\pm\infty$, $f(x)$ approaches 0, but oscillates as it does this.

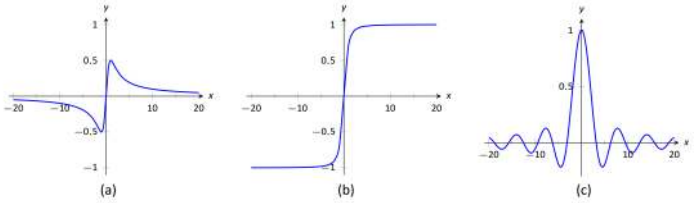


FIGURE 1.36: Considering different types of horizontal asymptotes.

We can analytically evaluate limits at infinity for rational functions once we understand $\lim_{x\to\infty} 1/x$. As x gets larger and larger, the $1/x$ gets smaller and smaller, approaching 0. We can, in fact, make $1/x$ as small as we want by choosing a large enough value of x. Given ϵ, we can make $1/x < \epsilon$ by choosing $x > 1/\epsilon$. Thus we have $\lim_{x\to\infty} 1/x = 0$.

It is now not much of a jump to conclude the following:

$$\lim_{x\to\infty} \frac{1}{x^n} = 0 \quad \text{and} \quad \lim_{x\to-\infty} \frac{1}{x^n} = 0 \tag{2.5.7}$$

Now suppose we need to compute the following limit:

$$\lim_{x\to\infty} \frac{x^3+2x+1}{4x^3-2x^2+9}. \tag{2.5.8}$$

A good way of approaching this is to divide through the numerator and denominator by x^3 (hence dividing by 1), which is the largest power of x to appear in the function. Doing this, we get

$$\begin{aligned}
\lim_{x\to\infty} \frac{x^3+2x+1}{4x^3-2x^2+9} &= \lim_{x\to\infty} \frac{1/x^3}{1/x^3} \cdot \frac{x^3+2x+1}{4x^3-2x^2+9} \\
&= \lim_{x\to\infty} \frac{x^3/x^3+2x/x^3+1/x^3}{4x^3/x^3-2x^2/x^3+9/x^3} \\
&= \lim_{x\to\infty} \frac{1+2/x^2+1/x^3}{4-2/x+9/x^3}.
\end{aligned}$$

Then using the rules for limits (which also hold for limits at infinity), as well as the fact about limits of $1/x^n$, we see that the limit becomes

$$\frac{1+0+0}{4-0+0} = \frac{1}{4}. \tag{2.5.9}$$

This procedure works for any rational function. In fact, it gives us the following theorem.

? Theorem: Limits of Rational Functions at Infinity

Let $f(x)$ be a rational function of the following form:

$$f(x) = \frac{a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0}{b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0}, \tag{2.5.10}$$

where any of the coefficients may be 0 except for a_n and b_m.

1. If $n = m$, then $\lim\limits_{x\to\infty} f(x) = \lim\limits_{x\to-\infty} f(x) = \frac{a_n}{b_m}$.
2. If $n < m$, then $\lim\limits_{x\to\infty} f(x) = \lim\limits_{x\to-\infty} f(x) = 0$.
3. If $n > m$, then $\lim\limits_{x\to\infty} f(x)$ and $\lim\limits_{x\to-\infty} f(x)$ are both infinite.

We can see why this is true. If the highest power of x is the same in both the numerator and denominator (i.e. $n = m$), we will be in a situation like the example above, where we will divide by x^n and in the limit all the terms will approach 0 except for $a_n x^n/x^n$ and $b_m x^m/x^n$. Since $n = m$, this will leave us with the limit a_n/b_m. If $n < m$, then after dividing through by x^m, all the terms in the numerator will approach 0 in the limit, leaving us with $0/b_m$ or 0. If $n > m$, and we try dividing through by x^n, we end up with all the terms in the denominator tending toward 0, while the x^n term in the numerator does not approach 0. This is indicative of some sort of infinite limit.

Intuitively, as x gets very large, all the terms in the numerator are small in comparison to $a_n x^n$, and likewise all the terms in the denominator are small compared to $b_n x^m$. If $n = m$, looking only at these two important terms, we have $(a_n x^n)/(b_n x^m)$. This reduces to a_n/b_m. If $n < m$, the function behaves like $a_n/(b_m x^{m-n})$, which tends toward 0. If $n > m$, the function behaves like $a_n x^{n-m}/b_m$, which will tend to either ∞ or $-\infty$ depending on the values of n, m, a_n, b_m and whether you are looking for $\lim\limits_{x\to\infty} f(x)$ or $\lim\limits_{x\to-\infty} f(x)$.

With care, we can quickly evaluate limits at infinity for a large number of functions by considering the largest powers of x. For instance, consider again $\lim\limits_{x\to\pm\infty} \frac{x}{\sqrt{x^2+1}}$, graphed in Figure ??? (b). When x is very large, $x^2 + 1 \approx x^2$. Thus

$$\sqrt{x^2+1} \approx \sqrt{x^2} = |x|, \quad \text{and} \quad \frac{x}{\sqrt{x^2+1}} \approx \frac{x}{|x|}. \tag{2.5.11}$$

This expression is 1 when x is positive and -1 when x is negative. Hence we get asymptotes of $y = 1$ and $y = -1$, respectively.

✓ Example ✓ 2.5.6: Finding a limit of a rational function

Confirm analytically that $y = 1$ is the horizontal asymptote of $f(x) = \frac{x^2}{x^2+4}$, as approximated in Example 29.

Solution

Before using Theorem 11, let's use the technique of evaluating limits at infinity of rational functions that led to that theorem. The largest power of x in f is 2, so divide the numerator and denominator of f by x^2, then take limits.

$$\begin{aligned}\lim_{x\to\infty} \frac{x^2}{x^2+4} &= \lim_{x\to\infty} \frac{x^2/x^2}{x^2/x^2 + 4/x^2} \\ &= \lim_{x\to\infty} \frac{1}{1+4/x^2} \\ &= \frac{1}{1+0} \\ &= 1.\end{aligned}$$

We can also use Theorem 11 directly; in this case $n = m$ so the limit is the ratio of the leading coefficients of the numerator and denominator, i.e., 1/1 = 1.

✔ Example ✔ 2.5.6: Finding limits of rational functions

Use Theorem 11 to evaluate each of the following limits.

$$1.\ \lim_{x\to-\infty}\frac{x^2+2x-1}{x^3+1} \qquad 3.\ \lim_{x\to\infty}\frac{x^2-1}{3-x} \tag{2.5.12}$$

$$2.\ \lim_{x\to\infty}\frac{x^2+2x-1}{1-x-3x^2} \tag{2.5.13}$$

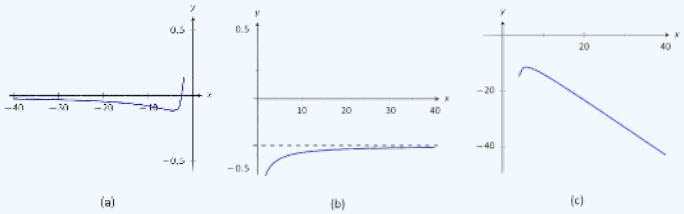


FIGURE 1.37: Visualizing the functions in Example 31.

Solution

1. The highest power of x is in the denominator. Therefore, the limit is 0; see Figure 1.37(a).
2. The highest power of x is x^2, which occurs in both the numerator and denominator. The limit is therefore the ratio of the coefficients of x^2, which is $-1/3$. See Figure 1.37(b).
3. The highest power of x is in the numerator so the limit will be ∞ or $-\infty$. To see which, consider only the dominant terms from the numerator and denominator, which are x^2 and $-x$. The expression in the limit will behave like $x^2/(-x) = -x$ for large values of x. Therefore, the limit is $-\infty$. See Figure 1.37(c).

2.5: Limits Involving Infinity; Asymptotes of Graphs is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **2.4: Continuity** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **2.2: The Limit of a Function** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **1.6: Limits Involving Infinity** by Gregory Hartman et al. is licensed CC BY-NC 3.0. Original source: http://www.apexcalculus.com/.

CHAPTER OVERVIEW

3: Differentiation

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

3.1: Defining the Derivative
3.2: The Derivative as a Function
3.3: Differentiation Rules
3.4: The Derivative as a Rate of Change
3.5: Derivatives of Trigonometric Functions
3.6: The Chain Rule
3.7: Implicit Differentiation
3.8: Derivatives of Inverse and Logarithmic Functions
3.9: Inverse Trigonometric Functions
3.10: Related Rates
3.11: Linearization and Differentials

3: Differentiation is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

3.1: Defining the Derivative

Learning Objectives

- Recognize the meaning of the tangent to a curve at a point.
- Calculate the slope of a tangent line.
- Identify the derivative as the limit of a difference quotient.
- Calculate the derivative of a given function at a point.
- Describe the velocity as a rate of change.
- Explain the difference between average velocity and instantaneous velocity.
- Estimate the derivative from a table of values.

Now that we have both a conceptual understanding of a limit and the practical ability to compute limits, we have established the foundation for our study of calculus, the branch of mathematics in which we compute derivatives and integrals. Most mathematicians and historians agree that calculus was developed independently by the Englishman Isaac **Newton (1643–1727)** and the German Gottfried **Leibniz (1646–1716)**, whose images appear in Figure 3.1.1. When we credit Newton and Leibniz with developing calculus, we are really referring to the fact that Newton and Leibniz were the first to understand the relationship between the derivative and the integral. Both mathematicians benefited from the work of predecessors, such as Barrow, Fermat, and Cavalieri. The initial relationship between the two mathematicians appears to have been amicable; however, in later years a bitter controversy erupted over whose work took precedence. Although it seems likely that Newton did, indeed, arrive at the ideas behind calculus first, we are indebted to Leibniz for the notation that we commonly use today.



Figure 3.1.1: Newton and Leibniz are credited with developing calculus independently.

Tangent Lines

We begin our study of calculus by revisiting the notion of secant lines and tangent lines. Recall that we used the slope of a secant line to a function at a point $(a, f(a))$ to estimate the rate of change, or the rate at which one variable changes in relation to another variable. We can obtain the slope of the secant by choosing a value of x near a and drawing a line through the points $(a, f(a))$ and $(x, f(x))$, as shown in Figure 3.1.2. The slope of this line is given by an equation in the form of a difference quotient:

$$m_{sec} = \frac{f(x) - f(a)}{x - a}$$

We can also calculate the slope of a secant line to a function at a value a by using this equation and replacing x with $a + h$, where h is a value close to a. We can then calculate the slope of the line through the points $(a, f(a))$ and $(a + h, f(a + h))$. In this case, we find the secant line has a slope given by the following difference quotient with increment h:

$$m_{sec} = \frac{f(a+h) - f(a)}{a + h - a} = \frac{f(a+h) - f(a)}{h}$$

Definition: Difference Quotient

Let f be a function defined on an interval I containing a. If $x \neq a$ is in I, then

$$Q = \frac{f(x) - f(a)}{x - a}$$

is a *difference quotient.*

Also, if $h \neq 0$ is chosen so that $a + h$ is in I, then

$$Q = \frac{f(a+h) - f(a)}{h}$$

is a difference quotient with increment h.

These two expressions for calculating the slope of a secant line are illustrated in Figure 3.1.2. We will see that each of these two methods for finding the slope of a secant line is of value. Depending on the setting, we can choose one or the other. The primary consideration in our choice usually depends on ease of calculation.

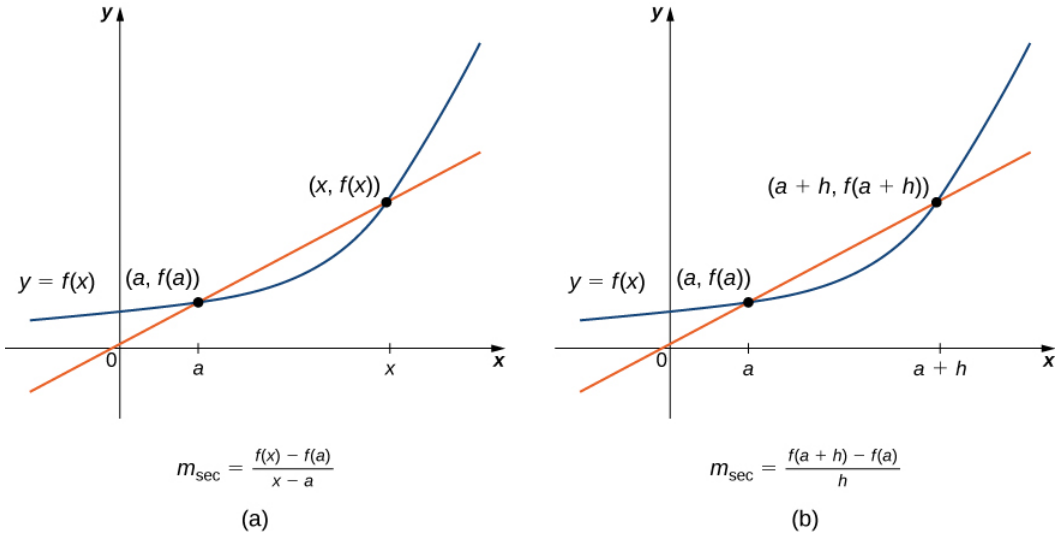


Figure 3.1.2: We can calculate the slope of a secant line in either of two ways.

In Figure 3.1.3*a* we see that, as the values of x approach a, the slopes of the secant lines provide better estimates of the rate of change of the function at a. Furthermore, the secant lines themselves approach the tangent line to the function at a, which represents the limit of the secant lines. Similarly, Figure 3.1.3*b* shows that as the values of h get closer to 0, the secant lines also approach the tangent line. The slope of the tangent line at a is the rate of change of the function at a, as shown in Figure 3.1.3*c*.

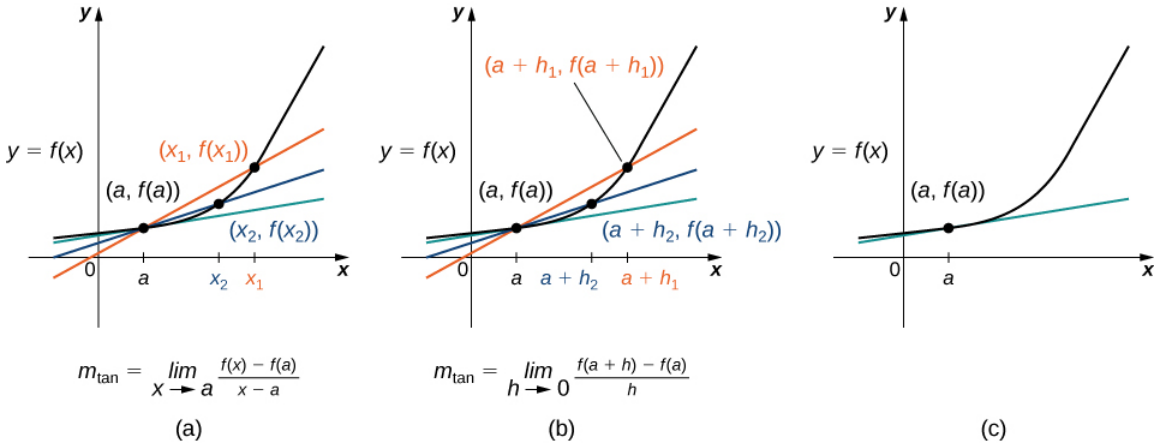


Figure 3.1.3: The secant lines approach the tangent line (shown in green) as the second point approaches the first.

In Figure 3.1.4 we show the graph of $f(x) = \sqrt{x}$ and its tangent line at $(1, 1)$ in a series of tighter intervals about $x = 1$. As the intervals become narrower, the graph of the function and its tangent line appear to coincide, making the values on the tangent line a good approximation to the values of the function for choices of x close to 1. In fact, the graph of $f(x)$ itself appears to be **locally linear** in the immediate vicinity of $x = 1$.

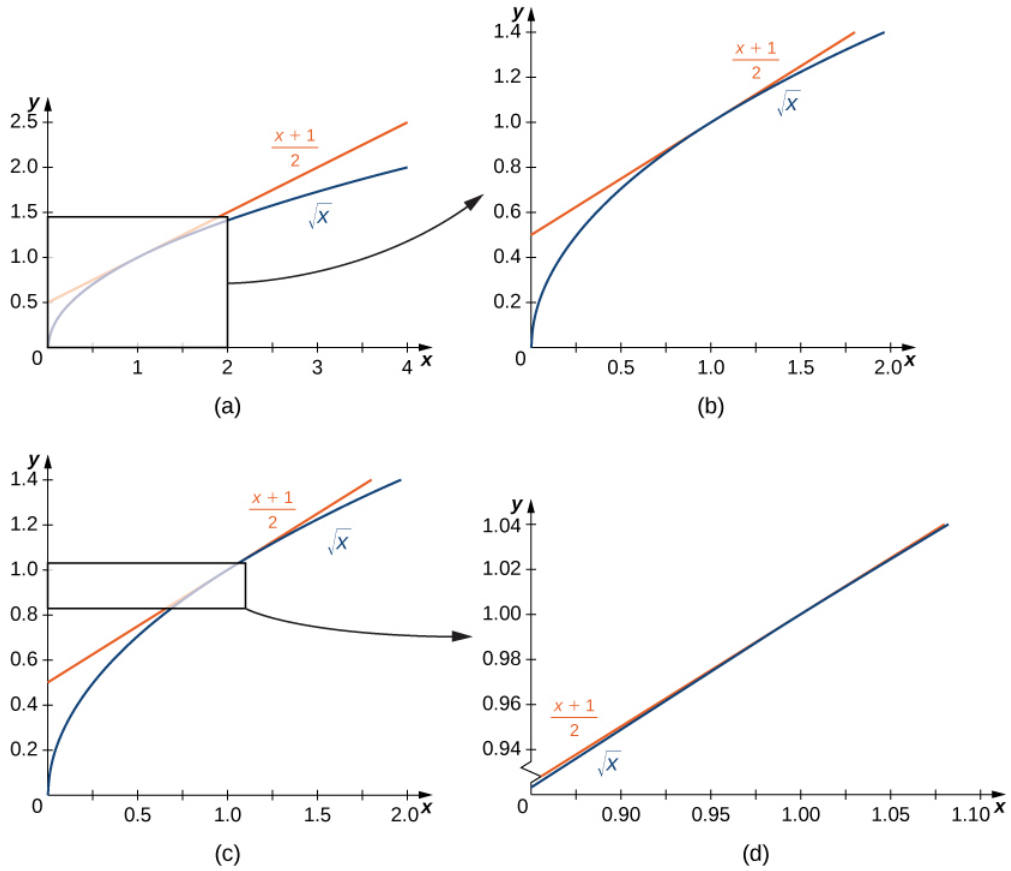


Figure 3.1.4: For values of x close to 1, the graph of $f(x) = \sqrt{x}$ and its tangent line appear to coincide.

Formally we may define the tangent line to the graph of a function as follows.

Definition: Tangent Line

Let $f(x)$ be a function defined in an open interval containing a. The tangent line to $f(x)$ at a is the line passing through the point $(a, f(a))$ having slope

$$m_{tan} = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \tag{3.1.1}$$

provided this limit exists.

Equivalently, we may define the *tangent line* to $f(x)$ at a to be the line passing through the point $(a, f(a))$ having slope

$$m_{tan} = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} \tag{3.1.2}$$

provided this limit exists.

Just as we have used two different expressions to define the slope of a secant line, we use two different forms to define the slope of the tangent line. In this text we use both forms of the definition. As before, the choice of definition will depend on the setting. Now that we have formally defined a tangent line to a function at a point, we can use this definition to find equations of tangent lines.

Example 3.1.1: Finding a Tangent Line

Find the equation of the line tangent to the graph of $f(x) = x^2$ at $x = 3$.

Solution

First find the slope of the tangent line. In this example, use Equation 3.1.1.

$$m_{tan} = \lim_{x \to 3} \frac{f(x) - f(3)}{x - 3} \qquad \text{Apply the definition.}$$

$$= \lim_{x \to 3} \frac{x^2 - 9}{x - 3} \qquad \text{Substitute } f(x) = x^2 \text{ and } f(3) = 9$$

$$= \lim_{x \to 3} \frac{(x-3)(x+3)}{x-3} = \lim_{x \to 3}(x+3) = 6 \qquad \text{Factor the numerator to evaluate the limit.}$$

Next, find a point on the tangent line. Since the line is tangent to the graph of $f(x)$ at $x = 3$, it passes through the point $(3, f(3))$. We have $f(3) = 9$, so the tangent line passes through the point $(3, 9)$.

Using the point-slope equation of the line with the slope $m = 6$ and the point $(3, 9)$, we obtain the line $y - 9 = 6(x - 3)$. Simplifying, we have $y = 6x - 9$. The graph of $f(x) = x^2$ and its tangent line at 3 are shown in Figure 3.1.5.

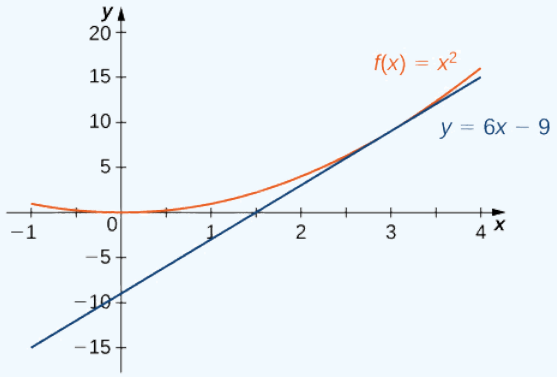


Figure 3.1.5: The tangent line to $f(x)$ at $x = 3$.

Example 3.1.2: The Slope of a Tangent Line Revisited

Use Equation 3.1.2 to find the slope of the line tangent to the graph of $f(x) = x^2$ at $x = 3$.

Solution

The steps are very similar to Example 3.1.1. See Equation 3.1.2 for the definition.

$$m_{tan} = \lim_{h \to 0} \frac{f(3+h) - f(3)}{h} \qquad \text{Apply the definition.}$$

$$= \lim_{h \to 0} \frac{(3+h)^2 - 9}{h} \qquad \text{Substitute } f(3+h) = (3+h)^2 \text{ and } f(3) = 9$$

$$= \lim_{h \to 0} \frac{9 + 6h + h^2 - 9}{h} \qquad \text{Expand and simplify to evaluate the limit.}$$

$$= \lim_{h \to 0} \frac{h(6+h)}{h} = \lim_{h \to 0}(6+h) = 6$$

We obtained the same value for the slope of the tangent line by using the other definition, demonstrating that the formulas can be interchanged.

Example 3.1.3: Finding the Equation of a Tangent Line

Find the equation of the line tangent to the graph of $f(x) = 1/x$ at $x = 2$.

Solution

We can use Equation 3.1.1, but as we have seen, the results are the same if we use Equation 3.1.2.

$$
\begin{aligned}
m_{tan} &= \lim_{x\to 2} \frac{f(x) - f(2)}{x-2} && \text{Apply the definition.}\\
&= \lim_{x\to 2} \frac{\frac{1}{x} - \frac{1}{2}}{x-2} && \text{Substitute } f(x) = \frac{1}{x} \text{ and } f(2) = \frac{1}{2}\\
&= \lim_{x\to 2} \frac{\frac{1}{x} - \frac{1}{2}}{x-2} \cdot \frac{2x}{2x} && \text{Multiply numerator and denominator by } 2x \text{ to simplify fractions.}\\
&= \lim_{x\to 2} \frac{(2-x)}{(x-2)(2x)} && \text{Simplify.}\\
&= \lim_{x\to 2} \frac{-1}{2x} && \text{Simplify using } \frac{2-x}{x-2} = -1, \text{ for } x \neq 2.\\
&= -\frac{1}{4} && \text{Evaluate the limit.}
\end{aligned}
$$

We now know that the slope of the tangent line is $-\frac{1}{4}$. To find the equation of the tangent line, we also need a point on the line. We know that $f(2) = \frac{1}{2}$. Since the tangent line passes through the point $(2, \frac{1}{2})$ we can use the point-slope equation of a line to find the equation of the tangent line. Thus the tangent line has the equation $y = -\frac{1}{4}x + 1$. The graphs of $f(x) = \frac{1}{x}$ and $y = -\frac{1}{4}x + 1$ are shown in Figure 3.1.6.

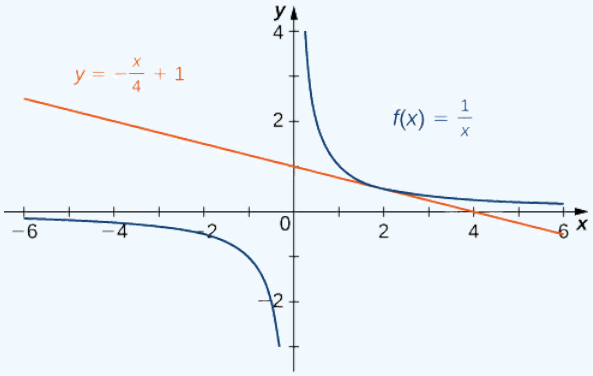


Figure 3.1.6:The line is tangent to $f(x)$ at $x = 2$.

Exercise 3.1.1

Find the slope of the line tangent to the graph of $f(x) = \sqrt{x}$ at $x = 4$.

Hint

Use either Equation 3.1.1or Equation 3.1.2 Multiply the numerator and the denominator by a conjugate.

Answer

$\frac{1}{4}$

The Derivative of a Function at a Point

The type of limit we compute in order to find the slope of the line tangent to a function at a point occurs in many applications across many disciplines. These applications include velocity and acceleration in physics, marginal profit functions in business, and

growth rates in biology. This limit occurs so frequently that we give this value a special name: the **derivative.** The process of finding a derivative is called **differentiation**.

Definition: Derivative

Let $f(x)$ be a function defined in an open interval containing a. The **derivative** of the function $f(x)$ at a, denoted by $f\,'(a)$, is defined by

$$f\,'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \tag{3.1.3}$$

provided this limit exists.

Alternatively, we may also define the derivative of $f(x)$ at a as

$$f\,'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}. \tag{3.1.4}$$

Example 3.1.4: Estimating a Derivative

For $f(x) = x^2$, use a table to estimate $f\,'(3)$ using Equation 3.1.3.

Solution

Create a table using values of x just below 3 and just above 3.

x	$\frac{x^2 - 9}{x - 3}$
2.9	5.9
2.99	5.99
2.999	5.999
3.001	6.001
3.01	6.01
3.1	6.1

After examining the table, we see that a good estimate is $f\,'(3) = 6$.

Exercise 3.1.2

For $f(x) = x^2$, use a table to estimate $f\,'(3)$ using Equation 3.1.4.

Hint

Evaluate $\frac{(x+h)^2 - x^2}{h}$ at $h = -0.1,\ -0.01,\ -0.001,\ 0.001,\ 0.01,\ 0.1$

Answer

6

Example 3.1.6: Finding a Derivative

For $f(x) = 3x^2 - 4x + 1$, find $f\,'(2)$ by using Equation 3.1.3.

Solution

Substitute the given function and value directly into the equation.

$$f'(x) = \lim_{x\to 2} \frac{f(x)-f(2)}{x-2}$$ Apply the definition.

$$= \lim_{x\to 2} \frac{(3x^2-4x+1)-5}{x-2}$$ Substitute $f(x) = 3x^2-4x+1$ and $f(2) = 5$.

$$= \lim_{x\to 2} \frac{(x-2)(3x+2)}{x-2}$$ Simplify and factor the numerator.

$$= \lim_{x\to 2} (3x+2)$$ Cancel the common factor.

$$= 8$$ Evaluate the limit.

Example 3.1.7: Revisiting the Derivative

For $f(x) = 3x^2-4x+1$, find $f'(2)$ by using Equation 3.1.4.

Solution

Using this equation, we can substitute two values of the function into the equation, and we should get the same value as in Example 3.1.6.

$$f'(2) = \lim_{h\to 0} \frac{f(2+h)-f(2)}{h}$$ Apply the definition.

$$= \lim_{h\to 0} \frac{(3(2+h)^2-4(2+h)+1)-5}{h}$$ Substitute $f(2) = 5$ and $f(2+h) = 3(2+h)^2-4(2+h)+1$.

$$= \lim_{h\to 0} \frac{3(4+4h+h^2)-8-4h+1-5}{h}$$ Expand the numerator.

$$= \lim_{h\to 0} \frac{12+12h+3h^2-12-4h}{h}$$ Distribute and begin simplifying the numerator.

$$= \lim_{h\to 0} \frac{3h^2+8h}{h}$$ Finish simplifying the numerator.

$$= \lim_{h\to 0} \frac{h(3h+8)}{h}$$ Factor the numerator.

$$= \lim_{h\to 0} (3h+8)$$ Cancel the common factor.

$$= 8$$ Evaluate the limit.

The results are the same whether we use Equation 3.1.3 or Equation 3.1.4.

Exercise 3.1.4

For $f(x) = x^2+3x+2$, find $f'(1)$.

Hint

Use either Equation 3.1.3, Equation 3.1.4, or try both.

Answer

$f'(1) = 5$

Velocities and Rates of Change

Now that we can evaluate a derivative, we can use it in velocity applications. Recall that if $s(t)$ is the position of an object moving along a coordinate axis, the **average velocity** of the object over a time interval $[a,t]$ if $t > a$ or $[t,a]$ if $t < a$ is given by the difference quotient

$$v_{ave} = \frac{s(t) - s(a)}{t - a}. \tag{3.1.5}$$

As the values of t approach a, the values of v_{ave} approach the value we call the **instantaneous velocity** at a. That is, instantaneous velocity at a, denoted $v(a)$, is given by

$$v(a) = s'(a) = \lim_{t \to a} \frac{s(t) - s(a)}{t - a}. \tag{3.1.6}$$

To better understand the relationship between average velocity and instantaneous velocity, see Figure 3.1.7. In this figure, the slope of the tangent line (shown in red) is the instantaneous velocity of the object at time $t = a$ whose position at time t is given by the function $s(t)$. The slope of the secant line (shown in green) is the average velocity of the object over the time interval $[a, t]$.

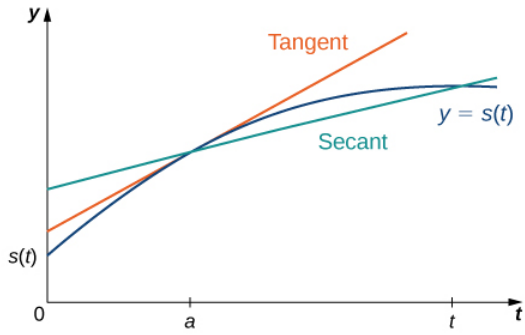


Figure 3.1.7: The slope of the secant line is the average velocity over the interval $[a, t]$. The slope of the tangent line is the instantaneous velocity.

We can use Equation 3.1.6 to calculate the instantaneous velocity, or we can estimate the velocity of a moving object by using a table of values. We can then confirm the estimate by using Equation 3.1.5.

Example 3.1.8: Estimating Velocity

A lead weight on a spring is oscillating up and down. Its position at time t with respect to a fixed horizontal line is given by $s(t) = \sin t$ (Figure 3.1.8). Use a table of values to estimate $v(0)$. Check the estimate by using Equation 3.1.6.

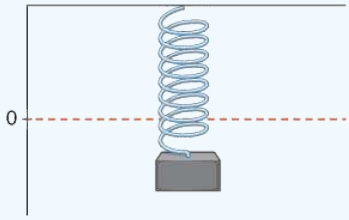


Figure 3.1.8: A lead weight suspended from a spring in vertical oscillatory motion.

Solution

We can estimate the instantaneous velocity at $t = 0$ by computing a table of average velocities using values of t approaching 0, as shown in Table 3.1.2.

Table 3.1.2: Average velocities using values of t approaching 0

t	$\frac{\sin t - \sin 0}{t - 0} = \frac{\sin t}{t}$
−0.1	0.998334166
−0.01	0.9999833333
−0.001	0.999999833
0.001	0.999999833
0.01	0.9999833333
0.1	0.998334166

From the table we see that the average velocity over the time interval $[-0.1, 0]$ is 0.998334166, the average velocity over the time interval $[-0.01, 0]$ is 0.9999833333, and so forth. Using this table of values, it appears that a good estimate is $v(0) = 1$.

By using Equation 3.1.6, we can see that

$$v(0) = s'(0) = \lim_{t \to 0} \frac{\sin t - \sin 0}{t - 0} = \lim_{t \to 0} \frac{\sin t}{t} = 1.$$

Thus, in fact, $v(0) = 1$.

Exercise 3.1.5

A rock is dropped from a height of 64 feet. Its height above ground at time t seconds later is given by $s(t) = -16t^2 + 64,\ 0 \le t \le 2$. Find its instantaneous velocity 1 second after it is dropped, using Equation 3.1.6.

Hint

$v(t) = s'(t)$. Follow the earlier examples of the derivative using Equation 3.1.6

Answer

−32 ft/s

As we have seen throughout this section, the slope of a tangent line to a function and instantaneous velocity are related concepts. Each is calculated by computing a derivative and each measures the instantaneous rate of change of a function, or the rate of change of a function at any point along the function.

Definition: Instantaneous Rate of Change

The **instantaneous rate of change** of a function $f(x)$ at a value a is its derivative $f'(a)$.

Example 3.1.9: Chapter Opener: Estimating Rate of Change of Velocity

Reaching a top speed of 270.49 mph, the Hennessey Venom GT is one of the fastest cars in the world. In tests it went from 0 to 60 mph in 3.05 seconds, from 0 to 100 mph in 5.88 seconds, from 0 to 200 mph in 14.51 seconds, and from 0 to 229.9 mph in 19.96 seconds. Use this data to draw a conclusion about the rate of change of velocity (that is, its **acceleration**) as it approaches 229.9 mph. Does the rate at which the car is accelerating appear to be increasing, decreasing, or constant?



Figure 3.1.9: (credit: modification of work by Codex41, Flickr)

Solution: First observe that 60 mph = 88 ft/s, 100 mph ≈146.67 ft/s, 200 mph ≈293.33 ft/s, and 229.9 mph ≈337.19 ft/s. We can summarize the information in a table.

Table 3.1.3: $v(t)$ at different values of t

t	$v(t)$
0	0
3.05	88
5.88	147.67

t	$v(t)$
14.51	293.33
19.96	337.19

Now compute the average acceleration of the car in feet per second on intervals of the form $[t, 19.96]$ as t approaches 19.96, as shown in the following table.

Average acceleration

t	$\frac{v(t)-v(19.96)}{t-19.96}=\frac{v(t)-337.19}{t-19.96}$
0.0	16.89
3.05	14.74
5.88	13.46
14.51	8.05

The rate at which the car is accelerating is decreasing as its velocity approaches $229.9\,\text{mph}$ ($337.19\,\text{ft/s}$).

Example 3.1.10: Rate of Change of Temperature

A homeowner sets the thermostat so that the temperature in the house begins to drop from $70^\circ F$ at 9 p.m., reaches a low of 60° during the night, and rises back to 70° by 7 a.m. the next morning. Suppose that the temperature in the house is given by $T(t)=0.4t^2-4t+70$ for $0\le t\le 10$, where t is the number of hours past 9 p.m. Find the instantaneous rate of change of the temperature at midnight.

Solution

Since midnight is 3 hours past 9 p.m., we want to compute $T'(3)$. Refer to Equation 3.1.3.

$$
\begin{aligned}
T'(3) &= \lim_{t\to 3}\frac{T(t)-T(3)}{t-3} && \text{Apply the definition.}\\
&= \lim_{t\to 3}\frac{0.4t^2-4t+70-61.6}{t-3} && \text{Substitute } T(t)=0.4t^2-4t+70 \text{ and } T(3)=61.6.\\
&= \lim_{t\to 3}\frac{0.4t^2-4t+8.4}{t-3} && \text{Simplify.}\\
&= \lim_{t\to 3}\frac{0.4(t-3)(t-7)}{t-3} &&\\
&= \lim_{t\to 3}0.4(t-7) && \text{Cancel.}\\
&= -1.6 && \text{Evaluate the limit.}
\end{aligned}
$$

The instantaneous rate of change of the temperature at midnight is $-1.6^\circ F$ per hour.

Example 3.1.11: Rate of Change of Profit

A toy company can sell x electronic gaming systems at a price of $p=-0.01x+400$ dollars per gaming system. The cost of manufacturing x systems is given by $C(x)=100x+10,000$ dollars. Find the rate of change of profit when $10,000$ games are produced. Should the toy company increase or decrease production?

Solution

The profit $P(x)$ earned by producing x gaming systems is $R(x)-C(x)$, where $R(x)$ is the revenue obtained from the sale of x games. Since the company can sell x games at $p=-0.01x+400$ per game,

$$R(x) = xp = x(-0.01x + 400) = -0.01x^2 + 400x.$$

Consequently,

$$P(x) = -0.01x^2 + 300x - 10,000.$$

Therefore, evaluating the rate of change of profit gives

$$\begin{aligned} P'(10000) &= \lim_{x \to 10000} \frac{P(x) - P(10000)}{x - 10000} \\ &= \lim_{x \to 10000} \frac{-0.01x^2 + 300x - 10000 - 1990000}{x - 10000} \\ &= \lim_{x \to 10000} \frac{-0.01x^2 + 300x - 2000000}{x - 10000} \\ &= 100 \end{aligned}.$$

Since the rate of change of profit $P'(10,000) > 0$ and $P(10,000) > 0$, the company should increase production.

? Exercise 3.1.6

A coffee shop determines that the daily profit on scones obtained by charging s dollars per scone is $P(s) = -20s^2 + 150s - 10$. The coffee shop currently charges $\$3.25$ per scone. Find $P'(3.25)$, the rate of change of profit when the price is $\$3.25$ and decide whether or not the coffee shop should consider raising or lowering its prices on scones.

Hint

Use Example 3.1.11for a guide.

Answer

$P'(3.25) = 20 > 0$; raise prices

Key Concepts

- The slope of the tangent line to a curve measures the instantaneous rate of change of a curve. We can calculate it by finding the limit of the difference quotient or the difference quotient with increment h.
- The derivative of a function $f(x)$ at a value a is found using either of the definitions for the slope of the tangent line.
- Velocity is the rate of change of position. As such, the velocity $v(t)$ at time t is the derivative of the position $s(t)$ at time t. Average velocity is given by

  $$v_{ave} = \frac{s(t) - s(a)}{t - a}.$$

  Instantaneous velocity is given by

  $$v(a) = s'(a) = \lim_{t \to a} \frac{s(t) - s(a)}{t - a}.$$

- We may estimate a derivative by using a table of values.

Key Equations

- **Difference quotient**

  $$Q = \frac{f(x) - f(a)}{x - a}$$

- **Difference quotient with increment h**

  $$Q = \frac{f(a+h) - f(a)}{a + h - a} = \frac{f(a+h) - f(a)}{h}$$

- **Slope of tangent line**

$$m_{tan} = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

$$m_{tan} = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

- **Derivative of f(x) at a**

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

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

- **Average velocity**

$$v_{ave} = \frac{s(t) - s(a)}{t - a}$$

- **Instantaneous velocity**

$$v(a) = s'(a) = \lim_{t \to a} \frac{s(t) - s(a)}{t - a}$$

Glossary

derivative

the slope of the tangent line to a function at a point, calculated by taking the limit of the difference quotient, is the derivative

difference quotient

of a function $f(x)$ at a is given by

$$\frac{f(a+h) - f(a)}{h} \text{ or } \frac{f(x) - f(a)}{x - a}$$

differentiation

the process of taking a derivative

instantaneous rate of change

the rate of change of a function at any point along the function a, also called $f'(a)$, or the derivative of the function at a

---

This page titled 3.1: Defining the Derivative is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **3.1: Defining the Derivative** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

3.2: The Derivative as a Function

Learning Objectives

- Define the derivative function of a given function.
- Graph a derivative function from the graph of a given function.
- State the connection between derivatives and continuity.
- Describe three conditions for when a function does not have a derivative.
- Explain the meaning of a higher-order derivative.

As we have seen, the derivative of a function at a given point gives us the rate of change or slope of the tangent line to the function at that point. If we differentiate a position function at a given time, we obtain the velocity at that time. It seems reasonable to conclude that knowing the derivative of the function at every point would produce valuable information about the behavior of the function. However, the process of finding the derivative at even a handful of values using the techniques of the preceding section would quickly become quite tedious. In this section we define the derivative function and learn a process for finding it.

Derivative Functions

The derivative function gives the derivative of a function at each point in the domain of the original function for which the derivative is defined. We can formally define a derivative function as follows.

Definition: Derivative Function

Let f be a function. The **derivative function**, denoted by f', is the function whose domain consists of those values of x such that the following limit exists:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}. \tag{3.2.1}$$

A function $f(x)$ is said to be **differentiable at** a if $f'(a)$ exists. More generally, a function is said to be **differentiable on** S if it is differentiable at every point in an open set S, and a differentiable function is one in which $f'(x)$ exists on its domain.

In the next few examples we use Equation 3.2.1 to find the derivative of a function.

Example 3.2.1: Finding the Derivative of a Square-Root Function

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

Solution

Start directly with the definition of the derivative function.

Substitute $f(x+h) = \sqrt{x+h}$ and $f(x) = \sqrt{x}$ into $f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$.

$f'(x) = \lim_{h \to 0} \frac{\sqrt{x+h} - \sqrt{x}}{h}$	
$= \lim_{h \to 0} \frac{\sqrt{x+h} - \sqrt{x}}{h} \cdot \frac{\sqrt{x+h} + \sqrt{x}}{\sqrt{x+h} + \sqrt{x}}$	Multiply numerator and denominator by $\sqrt{x+h} + \sqrt{x}$ without distributing in the denominator.
$= \lim_{h \to 0} \frac{h}{h\left(\sqrt{x+h} + \sqrt{x}\right)}$	Multiply the numerators and simplify.
$= \lim_{h \to 0} \frac{1}{\left(\sqrt{x+h} + \sqrt{x}\right)}$	Cancel the h.
$= \frac{1}{2\sqrt{x}}$	Evaluate the limit

Example 3.2.2: Finding the Derivative of a Quadratic Function

Find the derivative of the function $f(x) = x^2 - 2x$.

Solution

Follow the same procedure here, but without having to multiply by the conjugate.

Substitute $f(x+h) = (x+h)^2 - 2(x+h)$ and $f(x) = x^2 - 2x$ into $f'(x) = \lim_{h\to 0} \dfrac{f(x+h) - f(x)}{h}$.

$f'(x) = \lim_{h\to 0} \dfrac{((x+h)^2 - 2(x+h)) - (x^2 - 2x)}{h}$	
$= \lim_{h\to 0} \dfrac{x^2 + 2xh + h^2 - 2x - 2h - x^2 + 2x}{h}$	Expand $(x+h)^2 - 2(x+h)$.
$= \lim_{h\to 0} \dfrac{2xh - 2h + h^2}{h}$	Simplify
$= \lim_{h\to 0} \dfrac{h(2x - 2 + h)}{h}$	Factor out h from the numerator
$= \lim_{h\to 0} (2x - 2 + h)$	Cancel the common factor of h
$= 2x - 2$	Evaluate the limit

Exercise 3.2.1

Find the derivative of $f(x) = x^2$.

Hint

Use Equation 3.2.1 and follow the example.

Answer

$f'(x) = 2x$

We use a variety of different notations to express the derivative of a function. In Example 3.2.2 we showed that if $f(x) = x^2 - 2x$, then $f'(x) = 2x - 2$. If we had expressed this function in the form $y = x^2 - 2x$, we could have expressed the derivative as $y' = 2x - 2$ or $\dfrac{dy}{dx} = 2x - 2$. We could have conveyed the same information by writing $\dfrac{d}{dx}\left(x^2 - 2x\right) = 2x - 2$. Thus, for the function $y = f(x)$, each of the following notations represents the derivative of $f(x)$:

$$f'(x), \quad \frac{dy}{dx}, \quad y', \quad \frac{d}{dx}\left(f(x)\right).$$

In place of $f'(a)$ we may also use $\left.\dfrac{dy}{dx}\right|_{x=a}$. Use of the $\dfrac{dy}{dx}$ notation (called Leibniz notation) is quite common in engineering and physics. To understand this notation better, recall that the derivative of a function at a point is the limit of the slopes of secant lines as the secant lines approach the tangent line. The slopes of these secant lines are often expressed in the form $\dfrac{\Delta y}{\Delta x}$ where Δy is the difference in the y values corresponding to the difference in the x values, which are expressed as Δx (Figure 3.2.1). Thus the derivative, which can be thought of as the instantaneous rate of change of y with respect to x, is expressed as

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}.$$

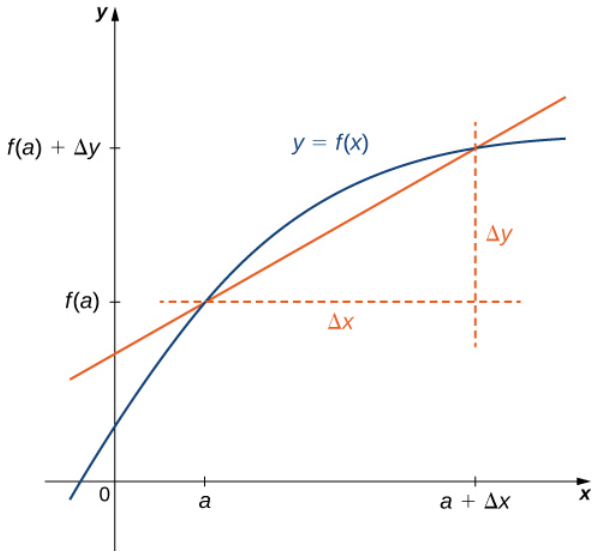


Figure 3.2.1: The derivative is expressed as $\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$.

Graphing a Derivative

We have already discussed how to graph a function, so given the equation of a function or the equation of a derivative function, we could graph it. Given both, we would expect to see a correspondence between the graphs of these two functions, since $f'(x)$ gives the rate of change of a function $f(x)$ (or slope of the tangent line to $f(x)$).

In Example 3.2.1, we found that for $f(x) = \sqrt{x}$, $f'(x) = \frac{1}{2\sqrt{x}}$. If we graph these functions on the same axes, as in Figure 3.2.2, we can use the graphs to understand the relationship between these two functions. First, we notice that $f(x)$ is increasing over its entire domain, which means that the slopes of its tangent lines at all points are positive. Consequently, we expect $f'(x) > 0$ for all values of x in its domain. Furthermore, as x increases, the slopes of the tangent lines to $f(x)$ are decreasing and we expect to see a corresponding decrease in $f'(x)$. We also observe that $f(0)$ is undefined and that $\lim_{x \to 0^+} f'(x) = +\infty$, corresponding to a vertical tangent to $f(x)$ at 0.

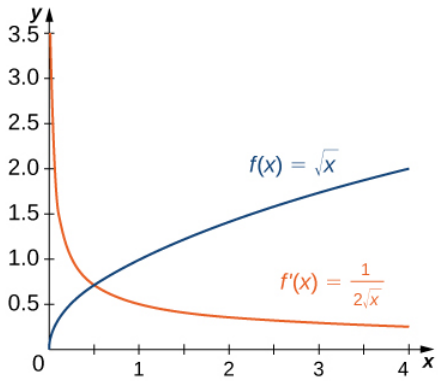


Figure 3.2.2: The derivative $f'(x)$ is positive everywhere because the function $f(x)$ is increasing.

In Example 3.2.2, we found that for $f(x) = x^2 - 2x$, $f'(x) = 2x - 2$. The graphs of these functions are shown in Figure 3.2.3. Observe that $f(x)$ is decreasing for $x < 1$. For these same values of x, $f'(x) < 0$. For values of $x > 1$, $f(x)$ is increasing and $f'(x) > 0$. Also, $f(x)$ has a horizontal tangent at $x = 1$ and $f'(1) = 0$.

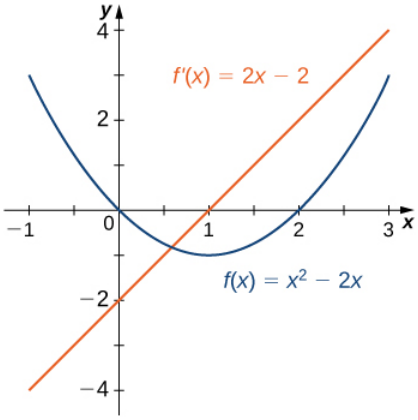
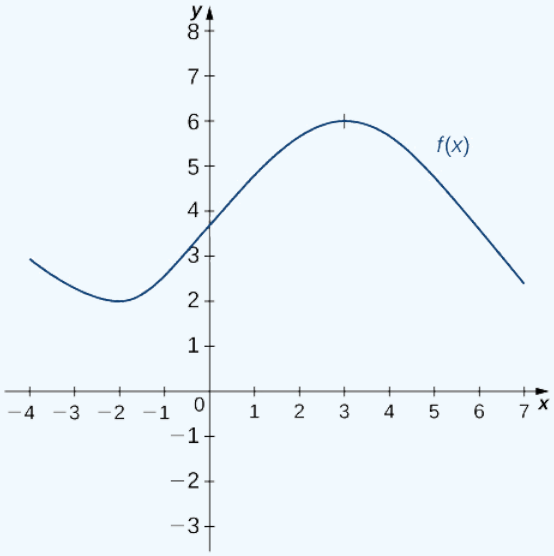


Figure 3.2.3: The derivative $f'(x) < 0$ where the function $f(x)$ is decreasing and $f'(x) > 0$ where $f(x)$ is increasing. The derivative is zero where the function has a horizontal tangent

Example 3.2.3: Sketching a Derivative Using a Function

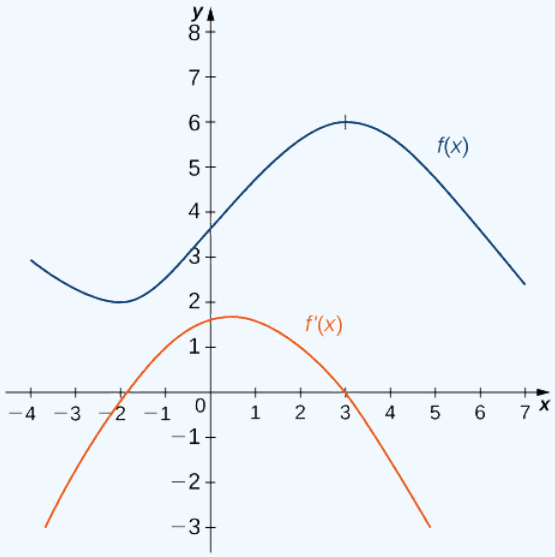
Use the following graph of $f(x)$ to sketch a graph of $f'(x)$.



Solution

The solution is shown in the following graph. Observe that $f(x)$ is increasing and $f'(x) > 0$ on $(-2, 3)$. Also, $f(x)$ is decreasing and $f'(x) < 0$ on $(-\infty, -2)$ and on $(3, +\infty)$. Also note that $f(x)$ has horizontal tangents at -2 and 3, and $f'(-2) = 0$ and $f'(3) = 0$.

https://math.libretexts.org/@go/page/138551



Exercise 3.2.2

Sketch the graph of $f(x) = x^2 - 4$. On what interval is the graph of $f'(x)$ above the x-axis?

Hint

The graph of $f'(x)$ is positive where $f(x)$ is increasing.

Answer

$(0, +\infty)$

Derivatives and Continuity

Now that we can graph a derivative, let's examine the behavior of the graphs. First, we consider the relationship between differentiability and continuity. We will see that if a function is differentiable at a point, it must be continuous there; however, a function that is continuous at a point need not be differentiable at that point. In fact, a function may be continuous at a point and fail to be differentiable at the point for one of several reasons.

Differentiability Implies Continuity

Let $f(x)$ be a function and a be in its domain. If $f(x)$ is differentiable at a, then f is continuous at a.

Proof

If $f(x)$ is differentiable at a, then $f'(a)$ exists and, if we let $h = x - a$, we have $x = a + h$, and as $h = x - a \to 0$, we can see that $x \to a$.

Then

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

can be rewritten as

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}.$$

We want to show that $f(x)$ is continuous at a by showing that $\lim_{x \to a} f(x) = f(a)$. Thus,

$$\begin{aligned}\lim_{x\to a} f(x) &= \lim_{x\to a}\ \big(f(x)-f(a)+f(a)\big)\\ &= \lim_{x\to a}\left(\frac{f(x)-f(a)}{x-a}\cdot(x-a)+f(a)\right) && \text{Multiply and divide } (f(x)-f(a)) \text{ by } x-a.\\ &= \left(\lim_{x\to a}\frac{f(x)-f(a)}{x-a}\right)\cdot\Big(\lim_{x\to a}(x-a)\Big)+\lim_{x\to a} f(a)\\ &= f'(a)\cdot 0+f(a)\\ &= f(a).\end{aligned}$$

Therefore, since $f(a)$ is defined and $\lim_{x\to a} f(x)=f(a)$, we conclude that f is continuous at a.

□

We have just proven that differentiability implies continuity, but now we consider whether continuity implies differentiability. To determine an answer to this question, we examine the function $f(x)=|x|$. This function is continuous everywhere; however, $f'(0)$ is undefined. This observation leads us to believe that continuity does not imply differentiability. Let's explore further. For $f(x)=|x|$,

$$f'(0)=\lim_{x\to 0}\frac{f(x)-f(0)}{x-0}=\lim_{x\to 0}\frac{|x|-|0|}{x-0}=\lim_{x\to 0}\frac{|x|}{x}\ .$$

This limit does not exist because

$$\lim_{x\to 0^-}\frac{|x|}{x}=-1 \text{ and } \lim_{x\to 0^+}\frac{|x|}{x}=1.$$

See Figure 3.2.4.

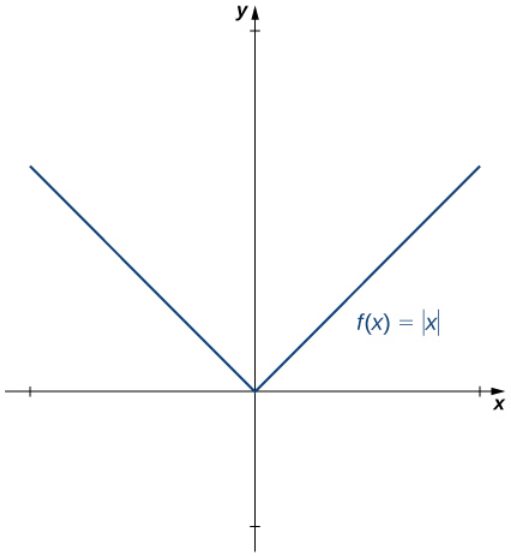


Figure 3.2.4: The function $f(x)=|x|$ is continuous at 0 but is not differentiable at 0.

Let's consider some additional situations in which a continuous function fails to be differentiable. Consider the function $f(x)=\sqrt[3]{x}$:

$$f'(0)=\lim_{x\to 0}\frac{\sqrt[3]{x}-0}{x-0}=\lim_{x\to 0}\frac{1}{\sqrt[3]{x^2}}=+\infty\ .$$

Thus $f'(0)$ does not exist. A quick look at the graph of $f(x)=\sqrt[3]{x}$ clarifies the situation. The function has a vertical tangent line at 0 (Figure 3.2.5).

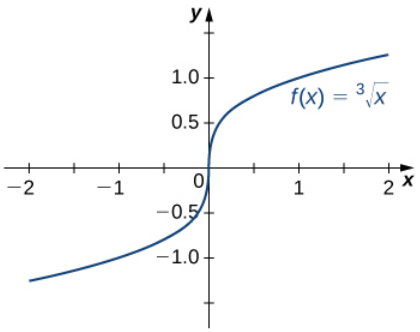


Figure 3.2.5: The function $f(x) = \sqrt[3]{x}$ has a vertical tangent at $x = 0$. It is continuous at 0 but is not differentiable at 0.

The function $f(x) = \begin{cases} x\sin\left(\frac{1}{x}\right), & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{cases}$ also has a derivative that exhibits interesting behavior at 0.

We see that

$$f'(0) = \lim_{x \to 0} \frac{x\sin(1/x) - 0}{x - 0} = \lim_{x \to 0} \sin\left(\frac{1}{x}\right).$$

This limit does not exist, essentially because the slopes of the secant lines continuously change direction as they approach zero (Figure 3.2.6).

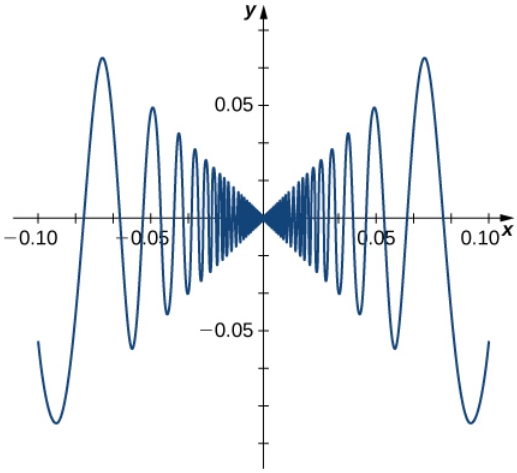


Figure 3.2.6: The function $f(x) = \begin{cases} x\sin\left(\frac{1}{x}\right), & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{cases}$ is not differentiable at 0.

In summary:

1. We observe that if a function is not continuous, it cannot be differentiable, since every differentiable function must be continuous. However, if a function is continuous, it may still fail to be differentiable.
2. We saw that $f(x) = |x|$ failed to be differentiable at 0 because the limit of the slopes of the tangent lines on the left and right were not the same. Visually, this resulted in a sharp corner on the graph of the function at 0. From this we conclude that in order to be differentiable at a point, a function must be "smooth" at that point.
3. As we saw in the example of $f(x) = \sqrt[3]{x}$, a function fails to be differentiable at a point where there is a vertical tangent line.
4. As we saw with $f(x) = \begin{cases} x\sin\left(\frac{1}{x}\right), & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{cases}$ a function may fail to be differentiable at a point in more complicated ways as well.

Example 3.2.4: A Piecewise Function that is Continuous and Differentiable

A toy company wants to design a track for a toy car that starts out along a parabolic curve and then converts to a straight line (Figure 3.2.7). The function that describes the track is to have the form $f(x) = \begin{cases} \frac{1}{10}x^2 + bx + c, & \text{if } x < -10 \\ -\frac{1}{4}x + \frac{5}{2}, & \text{if } x \geq -10 \end{cases}$ where x

and $f(x)$ are in inches. For the car to move smoothly along the track, the function $f(x)$ must be both continuous and differentiable at -10. Find values of b and c that make $f(x)$ both continuous and differentiable.

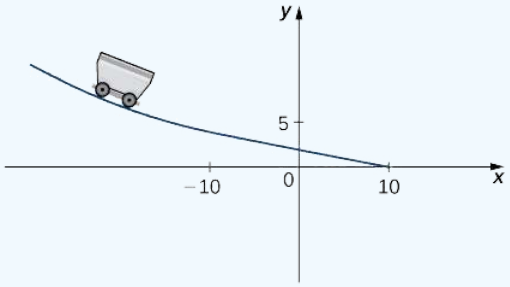


Figure 3.2.7: For the car to move smoothly along the track, the function must be both continuous and differentiable.

Solution

For the function to be continuous at $x=-10$, $\lim\limits_{x\to 10^-} f(x)=f(-10)$. Thus, since

$$\lim_{x\to -10^-} f(x)=\frac{1}{10}(-10)^2-10b+c=10-10b+c$$

and $f(-10)=5$, we must have $10-10b+c=5$. Equivalently, we have $c=10b-5$.

For the function to be differentiable at -10,

$$f'(10)=\lim_{x\to -10}\frac{f(x)-f(-10)}{x+10}$$

must exist. Since $f(x)$ is defined using different rules on the right and the left, we must evaluate this limit from the right and the left and then set them equal to each other:

$$\begin{aligned}
\lim_{x\to -10^-}\frac{f(x)-f(-10)}{x+10} &= \lim_{x\to -10^-}\frac{\frac{1}{10}x^2+bx+c-5}{x+10}\\
&= \lim_{x\to -10^-}\frac{\frac{1}{10}x^2+bx+(10b-5)-5}{x+10} \qquad \text{Substitute } c=10b-5.\\
&= \lim_{x\to -10^-}\frac{x^2-100+10bx+100b}{10(x+10)}\\
&= \lim_{x\to -10^-}\frac{(x+10)(x-10+10b)}{10(x+10)} \qquad \text{Factor by grouping}\\
&= b-2
\end{aligned}$$

We also have

$$\begin{aligned}
\lim_{x\to -10^+}\frac{f(x)-f(-10)}{x+10} &= \lim_{x\to -10^+}\frac{-\frac{1}{4}x+\frac{5}{2}-5}{x+10}\\
&= \lim_{x\to -10^+}\frac{-(x+10)}{4(x+10)}\\
&= -\frac{1}{4}
\end{aligned}$$

This gives us $b-2=-\frac{1}{4}$. Thus $b=\frac{7}{4}$ and $c=10(\frac{7}{4})-5=\frac{25}{2}$.

? Exercise 3.2.3

Find values of a and b that make $f(x)=\begin{cases} ax+b, & \text{if } x<3\\ x^2, & \text{if } x\geq 3\end{cases}$ both continuous and differentiable at 3.

Hint

Use Example 3.2.4as a guide.

Answer

$a = 6$ and $b = -9$

Higher-Order Derivatives

The derivative of a function is itself a function, so we can find the derivative of a derivative. For example, the derivative of a position function is the rate of change of position, or velocity. The derivative of velocity is the rate of change of velocity, which is acceleration. The new function obtained by differentiating the derivative is called the second derivative. Furthermore, we can continue to take derivatives to obtain the third derivative, fourth derivative, and so on. Collectively, these are referred to as **higher-order derivatives**. The notation for the higher-order derivatives of $y = f(x)$ can be expressed in any of the following forms:

$$f''(x),\ f'''(x),\ f^{(4)}(x),\ \ldots\ ,\ f^{(n)}(x)$$
$$y''(x),\ y'''(x),\ y^{(4)}(x),\ \ldots\ ,\ y^{(n)}(x)$$
$$\frac{d^2y}{dx^2},\ \frac{d^3y}{dy^3},\ \frac{d^4y}{dy^4},\ \ldots\ ,\ \frac{d^ny}{dy^n}.$$

It is interesting to note that the notation for $\frac{d^2y}{dx^2}$ may be viewed as an attempt to express $\frac{d}{dx}\left(\frac{dy}{dx}\right)$ more compactly.

Analogously, $\frac{d}{dx}\left(\frac{d}{dx}\left(\frac{dy}{dx}\right)\right) = \frac{d}{dx}\left(\frac{d^2y}{dx^2}\right) = \frac{d^3y}{dx^3}$.

Example 3.2.5: Finding a Second Derivative

For $f(x) = 2x^2 - 3x + 1$, find $f''(x)$.

Solution

First find $f'(x)$.

Substitute $f(x) = 2x^2 - 3x + 1$ and $f(x+h) = 2(x+h)^2 - 3(x+h) + 1$ into $f'(x) = \lim_{h\to 0}\frac{f(x+h) - f(x)}{h}$.

$f'(x) = \lim_{h\to 0}\frac{(2(x+h)^2 - 3(x+h) + 1) - (2x^2 - 3x + 1)}{h}$	
$= \lim_{h\to 0}\frac{4xh + 2h^2 - 3h}{h}$	Simplify the numerator.
$= \lim_{h\to 0}(4x + 2h - 3)$	Factor out the h in the numerator and cancel with the h in the denominator.
$= 4x - 3$	Take the limit.

Next, find $f''(x)$ by taking the derivative of $f'(x) = 4x - 3$.

$f''(x) = \lim_{h\to 0}\frac{f'(x+h) - f'(x)}{h}$	Use $f'(x) = \lim_{h\to 0}\frac{f(x+h) - f(x)}{h}$ with $f'(x)$ in place of $f(x)$.
$= \lim_{h\to 0}\frac{(4(x+h) - 3) - (4x - 3)}{h}$	Substitute $f'(x+h) = 4(x+h) - 3$ and $f'(x) = 4x - 3$.
$= \lim_{h\to 0} 4$	Simplify.
$= 4$	Take the limit.

Exercise 3.2.4

Find $f''(x)$ for $f(x) = x^2$.

Hint

We found $f'(x) = 2x$ in a previous checkpoint. Use Equation 3.2.1to find the derivative of $f'(x)$

Answer

$f''(x) = 2$

Example 3.2.6: Finding Acceleration

The position of a particle along a coordinate axis at time t (in seconds) is given by $s(t) = 3t^2 - 4t + 1$ (in meters). Find the function that describes its acceleration at time t.

Solution

Since $v(t) = s'(t)$ and $a(t) = v'(t) = s''(t)$, we begin by finding the derivative of $s(t)$:

$$\begin{aligned} s'(t) &= \lim_{h \to 0} \frac{s(t+h) - s(t)}{h} \\ &= \lim_{h \to 0} \frac{3(t+h)^2 - 4(t+h) + 1 - (3t^2 - 4t + 1)}{h} \\ &= 6t - 4. \end{aligned}$$

Next,

$$\begin{aligned} s''(t) &= \lim_{h \to 0} \frac{s'(t+h) - s'(t)}{h} \\ &= \lim_{h \to 0} \frac{6(t+h) - 4 - (6t - 4)}{h} \\ &= 6. \end{aligned}$$

Thus, $a = 6 \text{ m/s}^2$.

Exercise 3.2.5

For $s(t) = t^3$, find $a(t)$.

Hint

Use Example 3.2.6as a guide.

Answer

$a(t) = 6t$

Key Concepts

- The derivative of a function $f(x)$ is the function whose value at x is $f'(x)$.
- The graph of a derivative of a function $f(x)$ is related to the graph of $f(x)$. Where $f(x)$ has a tangent line with positive slope, $f'(x) > 0$. Where $f(x)$ has a tangent line with negative slope, $f'(x) < 0$. Where $f(x)$ has a horizontal tangent line, $f'(x) = 0$.
- If a function is differentiable at a point, then it is continuous at that point. A function is not differentiable at a point if it is not continuous at the point, if it has a vertical tangent line at the point, or if the graph has a sharp corner or cusp.
- Higher-order derivatives are derivatives of derivatives, from the second derivative to the nth derivative.

Key Equations

- **The derivative function**

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

Glossary

derivative function

gives the derivative of a function at each point in the domain of the original function for which the derivative is defined

differentiable at a

a function for which $f'(a)$ exists is differentiable at a

differentiable on S

a function for which $f'(x)$ exists for each x in the open set S is differentiable on S

differentiable function

a function for which $f'(x)$ exists is a differentiable function

higher-order derivative

a derivative of a derivative, from the second derivative to the nth derivative, is called a higher-order derivative

- Gilbert Strang (MIT) and Edwin “Jed” Herman (Harvey Mudd) with many contributing authors. This content by OpenStax is licensed with a CC-BY-SA-NC 4.0 license. Download for free at http://cnx.org.
- Paul Seeburger (Monroe Community College) added explanation of the alternative definition of the derivative used in the proof of that differentiability implies continuity.

This page titled 3.2: The Derivative as a Function is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **3.2: The Derivative as a Function** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

3.3: Differentiation Rules

Learning Objectives

- State the constant, constant multiple, and power rules.
- Apply the sum and difference rules to combine derivatives.
- Use the product rule for finding the derivative of a product of functions.
- Use the quotient rule for finding the derivative of a quotient of functions.
- Extend the power rule to functions with negative exponents.
- Combine the differentiation rules to find the derivative of a polynomial or rational function.

Finding derivatives of functions by using the definition of the derivative can be a lengthy and, for certain functions, a rather challenging process. For example, previously we found that

$$\frac{d}{dx}(\sqrt{x}) = \frac{1}{2\sqrt{x}}$$

by using a process that involved multiplying an expression by a conjugate prior to evaluating a limit.

The process that we could use to evaluate $\frac{d}{dx}(\sqrt[3]{x})$ using the definition, while similar, is more complicated.

In this section, we develop rules for finding derivatives that allow us to bypass this process. We begin with the basics.

The Basic Rules

The functions $f(x) = c$ and $g(x) = x^n$ where n is a positive integer are the building blocks from which all polynomials and rational functions are constructed. To find derivatives of polynomials and rational functions efficiently without resorting to the limit definition of the derivative, we must first develop formulas for differentiating these basic functions.

The Constant Rule

We first apply the limit definition of the derivative to find the derivative of the constant function, $f(x) = c$. For this function, both $f(x) = c$ and $f(x+h) = c$, so we obtain the following result:

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

The rule for differentiating constant functions is called the **constant rule**. It states that the derivative of a constant function is zero; that is, since a constant function is a horizontal line, the slope, or the rate of change, of a constant function is 0. We restate this rule in the following theorem.

The Constant Rule

Let c be a constant. If $f(x) = c$, then $f'(x) = 0$.

Alternatively, we may express this rule as

$$\frac{d}{dx}(c) = 0.$$

Example 3.3.1: Applying the Constant Rule

Find the derivative of $f(x) = 8$.

Solution

This is just a one-step application of the rule: $f'(8) = 0$.

Exercise 3.3.1

Find the derivative of $g(x) = -3$.

Hint

Use the preceding example as a guide

Answer

0

The Power Rule

We have shown that

$$\frac{d}{dx}\left(x^2\right)=2x \quad \text{and} \quad \frac{d}{dx}\left(x^{1/2}\right)=\frac{1}{2}x^{-1/2}.$$

At this point, you might see a pattern beginning to develop for derivatives of the form $\frac{d}{dx}(x^n)$. We continue our examination of derivative formulas by differentiating power functions of the form $f(x)=x^n$ where n is a positive integer. We develop formulas for derivatives of this type of function in stages, beginning with positive integer powers. Before stating and proving the general rule for derivatives of functions of this form, we take a look at a specific case, $\frac{d}{dx}(x^3)$. As we go through this derivation, pay special attention to the portion of the expression in boldface, as the technique used in this case is essentially the same as the technique used to prove the general case.

Example 3.3.2: Differentiating x^3

Find $\frac{d}{dx}\left(x^3\right)$.

Solution:

$\frac{d}{dx}\left(x^3\right)=\lim_{h\to 0}\frac{(x+h)^3-x^3}{h}$	
$=\lim_{h\to 0}\frac{x^3+3x^2h+3xh^2+h^3-x^3}{h}$	Notice that the first term in the expansion of $(x+h)^3$ is x^3 and the second term is $3x^2h$. All other terms contain powers of h that are two or greater
$=\lim_{h\to 0}\frac{3x^2h+3xh^2+h^3}{h}$	In this step the x^3 terms have been cancelled, leaving only terms containing h.
$=\lim_{h\to 0}\frac{h(3x^2+3xh+h^2)}{h}$	Factor out the common factor of h.
$=\lim_{h\to 0}(3x^2+3xh+h^2)$	After cancelling the common factor of h,the only term not containing h is $3x^2$.
$=3x^2$	Let h go to 0.

Exercise 3.3.2

Find $\frac{d}{dx}\left(x^4\right)$.

Hint

Use $(x+h)^4=x^4+4x^3h+6x^2h^2+4xh^3+h^4$ and follow the procedure outlined in the preceding example.

Answer

$\frac{d}{dx}\left(x^4\right)=4x^3$

As we shall see, the procedure for finding the derivative of the general form $f(x)=x^n$ is very similar. Although it is often unwise to draw general conclusions from specific examples, we note that when we differentiate $f(x)=x^3$, the power on x becomes the coefficient of x^2 in the derivative and the power on x in the derivative decreases by 1. The following theorem states that the **power rule** holds for all positive integer powers of x. We will eventually extend this result to negative integer powers. Later, we will see that this rule may also be extended first to rational powers of x and then to arbitrary powers of x. Be aware, however, that this rule does not apply to functions in which a constant is raised to a variable power, such as $f(x)=3^x$.

The Power Rule

Let n be a positive integer. If $f(x)=x^n$,then

$$f'(x)=nx^{n-1}.$$

Alternatively, we may express this rule as

$$\frac{d}{dx}(x^n) = nx^{n-1}.$$

Proof

For $f(x) = x^n$ where n is a positive integer, we have

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

Since

$$(x+h)^n = x^n + nx^{n-1}h + \binom{n}{2}x^{n-2}h^2 + \binom{n}{3}x^{n-3}h^3 + \ldots + nxh^{n-1} + h^n,$$

we see that

$$(x+h)^n - x^n = nx^{n-1}h + \binom{n}{2}x^{n-2}h^2 + \binom{n}{3}x^{n-3}h^3 + \ldots + nxh^{n-1} + h^n.$$

Next, divide both sides by h:

$$\frac{(x+h)^n - x^n}{h} = \frac{nx^{n-1}h + \binom{n}{2}x^{n-2}h^2 + \binom{n}{3}x^{n-3}h^3 + \ldots + nxh^{n-1} + h^n}{h}.$$

Thus,

$$\frac{(x+h)^n - x^n}{h} = nx^{n-1} + \binom{n}{2}x^{n-2}h + \binom{n}{3}x^{n-3}h^2 + \ldots + nxh^{n-2} + h^{n-1}.$$

Finally,

$$f'(x) = \lim_{h \to 0}(nx^{n-1} + \binom{n}{2}x^{n-2}h + \binom{n}{3}x^{n-3}h^2 + \ldots + nxh^{n-2} + h^{n-1})$$
$$= nx^{n-1}.$$

□

Example 3.3.3: Applying the Power Rule

Find the derivative of the function $f(x) = x^{10}$ by applying the power rule.

Solution

Using the power rule with $n = 10$, we obtain

$$f'(x) = 10x^{10-1} = 10x^9.$$

Exercise 3.3.3

Find the derivative of $f(x) = x^7$.

Hint

Use the power rule with $n = 7$.

Answer

$f'(x) = 7x^6$

The Sum, Difference, and Constant Multiple Rules

We find our next differentiation rules by looking at derivatives of sums, differences, and constant multiples of functions. Just as when we work with functions, there are rules that make it easier to find derivatives of functions that we add, subtract, or multiply by a constant. These rules are summarized in the following theorem.

Sum, Difference, and Constant Multiple Rules

Let $f(x)$ and $g(x)$ be differentiable functions and k be a constant. Then each of the following equations holds.

Sum Rule. The derivative of the sum of a function f and a function g is the same as the sum of the derivative of f and the derivative of g.

$$\frac{d}{dx}\big(f(x) + g(x)\big) = \frac{d}{dx}\big(f(x)\big) + \frac{d}{dx}\big(g(x)\big);$$

that is,

$$\text{for } s(x) = f(x) + g(x), \quad s'(x) = f'(x) + g'(x).$$

Difference Rule. The derivative of the difference of a function f and a function g is the same as the difference of the derivative of f and the derivative of g :

$$\frac{d}{dx}(f(x) - g(x)) = \frac{d}{dx}(f(x)) - \frac{d}{dx}(g(x));$$

that is,

$$\text{for } d(x) = f(x) - g(x), \quad d'(x) = f'(x) - g'(x).$$

Constant Multiple Rule. The derivative of a constant k multiplied by a function f is the same as the constant multiplied by the derivative:

$$\frac{d}{dx}\big(kf(x)\big) = k\frac{d}{dx}\big(f(x)\big);$$

that is,

$$\text{for } m(x) = kf(x), \quad m'(x) = kf'(x).$$

Proof

We provide only the proof of the sum rule here. The rest follow in a similar manner.

For differentiable functions $f(x)$ and $g(x)$, we set $s(x) = f(x) + g(x)$. Using the limit definition of the derivative we have

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

By substituting $s(x+h) = f(x+h) + g(x+h)$ and $s(x) = f(x) + g(x)$, we obtain

$$s'(x) = \lim_{h \to 0} \frac{\big(f(x+h) + g(x+h)\big) - \big(f(x) + g(x)\big)}{h}.$$

Rearranging and regrouping the terms, we have

$$s'(x) = \lim_{h \to 0} \left(\frac{f(x+h) - f(x)}{h} + \frac{g(x+h) - g(x)}{h} \right).$$

We now apply the sum law for limits and the definition of the derivative to obtain

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

□

Example 3.3.4: Applying the Constant Multiple Rule

Find the derivative of $g(x) = 3x^2$ and compare it to the derivative of $f(x) = x^2$.

Solution

We use the power rule directly:

$$g'(x) = \frac{d}{dx}(3x^2) = 3\frac{d}{dx}(x^2) = 3(2x) = 6x.$$

Since $f(x) = x^2$ has derivative $f'(x) = 2x$, we see that the derivative of $g(x)$ is 3 times the derivative of $f(x)$. This relationship is illustrated in Figure 3.3.1.

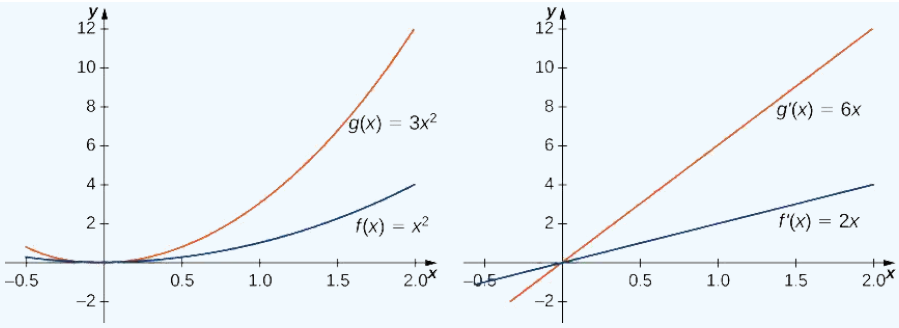


Figure 3.3.1: The derivative of $g(x)$ is 3 times the derivative of $f(x)$.

Example 3.3.5: Applying Basic Derivative Rules

Find the derivative of $f(x) = 2x^5 + 7$.

Solution

We begin by applying the rule for differentiating the sum of two functions, followed by the rules for differentiating constant multiples of functions and the rule for differentiating powers. To better understand the sequence in which the differentiation rules are applied, we use Leibniz notation throughout the solution:

$$\begin{aligned} f'(x) &= \frac{d}{dx}\left(2x^5 + 7\right) \\ &= \frac{d}{dx}(2x^5) + \frac{d}{dx}(7) && \text{Apply the sum rule.} \\ &= 2\frac{d}{dx}(x^5) + \frac{d}{dx}(7) && \text{Apply the constant multiple rule.} \\ &= 2(5x^4) + 0 && \text{Apply the power rule and the constant rule.} \\ &= 10x^4 && \text{Simplify.} \end{aligned}$$

Exercise 3.3.4

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

Hint

Use the preceding example as a guide.

Answer

$f'(x) = 6x^2 - 12x$.

Example 3.3.6: Finding the Equation of a Tangent Line

Find the equation of the line tangent to the graph of $f(x) = x^2 - 4x + 6$ at $x = 1$

Solution

To find the equation of the tangent line, we need a point and a slope. To find the point, compute

$$f(1) = 1^2 - 4(1) + 6 = 3.$$

This gives us the point $(1, 3)$. Since the slope of the tangent line at 1 is $f'(1)$, we must first find $f'(x)$. Using the definition of a derivative, we have

$$f'(x) = 2x - 4$$

so the slope of the tangent line is $f'(1) = -2$. Using the point-slope formula, we see that the equation of the tangent line is

$$y - 3 = -2(x - 1).$$

Putting the equation of the line in slope-intercept form, we obtain

$$y = -2x + 5.$$

Exercise 3.3.5

Find the equation of the line tangent to the graph of $f(x) = 3x^2 - 11$ at $x = 2$. Use the point-slope form.

Hint

Use the preceding example as a guide.

Answer

$y = 12x - 23$

The Product Rule

Now that we have examined the basic rules, we can begin looking at some of the more advanced rules. The first one examines the derivative of the product of two functions. Although it might be tempting to assume that the derivative of the product is the product of the derivatives, similar to the sum and difference rules, the **product rule** does not follow this pattern. To see why we cannot use this pattern, consider the function $f(x) = x^2$, whose derivative is $f\,'(x) = 2x$ and not $\frac{d}{dx}(x) \cdot \frac{d}{dx}(x) = 1 \cdot 1 = 1$.

Product Rule

Let $f(x)$ and $g(x)$ be differentiable functions. Then

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

That is,

$$\text{if } p(x) = f(x)g(x), \quad \text{then } p\,'(x) = f\,'(x)g(x) + g\,'(x)f(x).$$

This means that the derivative of a product of two functions is the derivative of the first function times the second function plus the derivative of the second function times the first function.

Proof

We begin by assuming that $f(x)$ and $g(x)$ are differentiable functions. At a key point in this proof we need to use the fact that, since $g(x)$ is differentiable, it is also continuous. In particular, we use the fact that since $g(x)$ is continuous, $\lim_{h \to 0} g(x+h) = g(x)$.

By applying the limit definition of the derivative to $p(x) = f(x)g(x)$, we obtain

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

By adding and subtracting $f(x)g(x+h)$ in the numerator, we have

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

After breaking apart this quotient and applying the sum law for limits, the derivative becomes

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

Rearranging, we obtain

$$\begin{aligned} p\,'(x) &= \lim_{h \to 0} \left(\frac{f(x+h) - f(x)}{h} \cdot g(x+h) \right) + \lim_{h \to 0} \left(\frac{g(x+h) - g(x)}{h} \cdot f(x) \right) \\ &= \left(\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \right) \cdot \left(\lim_{h \to 0} g(x+h) \right) + \left(\lim_{h \to 0} \frac{g(x+h) - g(x)}{h} \right) \cdot f(x) \end{aligned}$$

By using the continuity of $g(x)$, the definition of the derivatives of $f(x)$ and $g(x)$, and applying the limit laws, we arrive at the product rule,

$$p\,'(x) = f\,'(x)g(x) + g\,'(x)f(x).$$

□

Example 3.3.7: Applying the Product Rule to Constant Functions

For $p(x) = f(x)g(x)$, use the product rule to find $p'(2)$ if $f(2) = 3,\ f'(2) = -4,\ g(2) = 1$, and $g'(2) = 6$.

Solution

Since $p(x) = f(x)g(x)$, $p'(x) = f'(x)g(x) + g'(x)f(x)$, and hence

$p'(2) = f'(2)g(2) + g'(2)f(2) = (-4)(1) + (6)(3) = 14.$

Example 3.3.8: Applying the Product Rule to Binomials

For $p(x) = (x^2 + 2)(3x^3 - 5x),$ find $p'(x)$ by applying the product rule. Check the result by first finding the product and then differentiating.

Solution

If we set $f(x) = x^2 + 2$ and $g(x) = 3x^3 - 5x$, then $f'(x) = 2x$ and $g'(x) = 9x^2 - 5$. Thus,

$p'(x) = f'(x)g(x) + g'(x)f(x) = (2x)(3x^3 - 5x) + (9x^2 - 5)(x^2 + 2).$

Simplifying, we have

$$p'(x) = 15x^4 + 3x^2 - 10.$$

To check, we see that $p(x) = 3x^5 + x^3 - 10x$ and, consequently, $p'(x) = 15x^4 + 3x^2 - 10.$

Exercise 3.3.6

Use the product rule to obtain the derivative of $p(x) = 2x^5(4x^2 + x)$.

Hint

Set $f(x) = 2x^5$ and $g(x) = 4x^2 + x$ and use the preceding example as a guide.

Answer

$p'(x) = 10x^4(4x^2 + x) + (8x + 1)(2x^5) = 56x^6 + 12x^5.$

The Quotient Rule

Having developed and practiced the product rule, we now consider differentiating quotients of functions. As we see in the following theorem, the derivative of the quotient is not the quotient of the derivatives; rather, it is the derivative of the function in the numerator times the function in the denominator minus the derivative of the function in the denominator times the function in the numerator, all divided by the square of the function in the denominator. In order to better grasp why we cannot simply take the quotient of the derivatives, keep in mind that

$$\frac{d}{dx}(x^2) = 2x, \text{ not } \frac{\frac{d}{dx}(x^3)}{\frac{d}{dx}(x)} = \frac{3x^2}{1} = 3x^2.$$

The Quotient Rule

Let $f(x)$ and $g(x)$ be differentiable functions. Then

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{\frac{d}{dx}(f(x)) \cdot g(x) - \frac{d}{dx}(g(x)) \cdot f(x)}{(g(x))^2}.$$

That is, if

$$q(x) = \frac{f(x)}{g(x)}$$

then

$$q'(x) = \frac{f'(x)g(x) - g'(x)f(x)}{(g(x))^2}.$$

The proof of the quotient rule is very similar to the proof of the product rule, so it is omitted here. Instead, we apply this new rule for finding derivatives in the next example.

Example 3.3.9: Applying the Quotient Rule

Use the quotient rule to find the derivative of $q(x) = \dfrac{5x^2}{4x+3}$.

Solution

Let $f(x) = 5x^2$ and $g(x) = 4x+3$. Thus, $f'(x) = 10x$ and $g'(x) = 4$.

Substituting into the quotient rule, we have

$$q'(x) = \frac{f'(x)g(x) - g'(x)f(x)}{(g(x))^2} = \frac{10x(4x+3) - 4(5x^2)}{(4x+3)^2}.$$

Simplifying, we obtain

$$q'(x) = \frac{20x^2 + 30x}{(4x+3)^2}$$

Exercise 3.3.7

Find the derivative of $h(x) = \dfrac{3x+1}{4x-3}$.

Hint

Apply the quotient rule with $f(x) = 3x+1$ and $g(x) = 4x-3$.

Answer

$h'(x) = -\dfrac{13}{(4x-3)^2}$.

It is now possible to use the quotient rule to extend the power rule to find derivatives of functions of the form x^k where k is a negative integer.

Extended Power Rule

If k is a negative integer, then

$$\frac{d}{dx}(x^k) = kx^{k-1}.$$

Proof

If k is a negative integer, we may set $n = -k$, so that n is a positive integer with $k = -n$. Since for each positive integer n, $x^{-n} = \dfrac{1}{x^n}$, we may now apply the quotient rule by setting $f(x) = 1$ and $g(x) = x^n$. In this case, $f'(x) = 0$ and $g'(x) = nx^{n-1}$. Thus,

$$\frac{d}{dx}(x^{-n}) = \frac{0(x^n) - 1(nx^{n-1})}{(x^n)^2}.$$

Simplifying, we see that

$$\begin{aligned}\frac{d}{dx}(x^{-n}) &= \frac{-nx^{n-1}}{x^{2n}} \\ &= -nx^{(n-1)-2n} \\ &= -nx^{-n-1}.\end{aligned}$$

Finally, observe that since $k = -n$, by substituting we have

$$\frac{d}{dx}(x^k) = kx^{k-1}.$$

□

Example 3.3.10: Using the Extended Power Rule

Find $\dfrac{d}{dx}(x^{-4})$.

Solution

By applying the extended power rule with $k = -4$, we obtain

$$\frac{d}{dx}(x^{-4}) = -4x^{-4-1} = -4x^{-5}.$$

Example 3.3.11: Using the Extended Power Rule and the Constant Multiple Rule

Use the extended power rule and the constant multiple rule to find $f(x) = \frac{6}{x^2}$.

Solution

It may seem tempting to use the quotient rule to find this derivative, and it would certainly not be incorrect to do so. However, it is far easier to differentiate this function by first rewriting it as $f(x) = 6x^{-2}$.

$$\begin{aligned} f'(x) = \frac{d}{dx}\left(\frac{6}{x^2}\right) = \frac{d}{dx}(6x^{-2}) & \quad \text{Rewrite } \frac{6}{x^2} \text{ as } 6x^{-2}. \\ = 6\frac{d}{dx}(x^{-2}) & \quad \text{Apply the constant multiple rule.} \\ = 6(-2x^{-3}) & \quad \text{Use the extended power rule to differentiate } x^{-2}. \\ = -12x^{-3} & \quad \text{Simplify.} \end{aligned}$$

Exercise 3.3.8

Find the derivative of $g(x) = \frac{1}{x^7}$ using the extended power rule.

Hint

Rewrite $g(x) = \frac{1}{x^7} = x^{-7}$. Use the extended power rule with $k = -7$.

Answer

$g'(x) = -7x^{-8}$.

Combining Differentiation Rules

As we have seen throughout the examples in this section, it seldom happens that we are called on to apply just one differentiation rule to find the derivative of a given function. At this point, by combining the differentiation rules, we may find the derivatives of any polynomial or rational function. Later on we will encounter more complex combinations of differentiation rules. A good rule of thumb to use when applying several rules is to apply the rules in reverse of the order in which we would evaluate the function.

Example 3.3.12: Combining Differentiation Rules

For $k(x) = 3h(x) + x^2 g(x)$, find $k'(x)$.

Solution: Finding this derivative requires the sum rule, the constant multiple rule, and the product rule.

$k'(x) = \frac{d}{dx}\left(3h(x) + x^2 g(x)\right) = \frac{d}{dx}\left(3h(x)\right) + \frac{d}{dx}\left(x^2 g(x)\right)$	Apply the sum rule.
$= 3\frac{d}{dx}\left(h(x)\right) + \left(\frac{d}{dx}(x^2)g(x) + \frac{d}{dx}(g(x))x^2\right)$	Apply the constant multiple rule to differentiate $3h(x)$ and the product rule to differentiate $x^2 g(x)$.
$= 3h'(x) + 2xg(x) + g'(x)x^2$	

Example 3.3.13: Extending the Product Rule

For $k(x) = f(x)g(x)h(x)$, express $k'(x)$ in terms of $f(x), g(x), h(x)$, and their derivatives.

Solution

We can think of the function $k(x)$ as the product of the function $f(x)g(x)$ and the function $h(x)$. That is, $k(x) = (f(x)g(x)) \cdot h(x)$. Thus,

$$\begin{aligned} k'(x) = \frac{d}{dx}\left(f(x)g(x)\right) \cdot h(x) + \frac{d}{dx}\left(h(x)\right) \cdot \left(f(x)g(x)\right). & \quad \text{Apply the product rule to the product of } f(x)g(x) \text{ and } h(x). \\ = \left(f'(x)g(x) + g'(x)f(x)\right)h(x) + h'(x)f(x)g(x) & \quad \text{Apply the product rule to } f(x)g(x) \\ = f'(x)g(x)h(x) + f(x)g'(x)h(x) + f(x)g(x)h'(x). & \quad \text{Simplify.} \end{aligned}$$

Example 3.3.14: Combining the Quotient Rule and the Product Rule

For $h(x)=\dfrac{2x^3k(x)}{3x+2}$, find $h'(x)$.

Solution

This procedure is typical for finding the derivative of a rational function.

$$h'(x)=\frac{\frac{d}{dx}(2x^3k(x))\cdot(3x+2)-\frac{d}{dx}(3x+2)\cdot(2x^3k(x))}{(3x+2)^2}$$ Apply the quotient rule.

$$=\frac{(6x^2k(x)+k'(x)\cdot 2x^3)(3x+2)-3(2x^3k(x))}{(3x+2)^2}$$ Apply the product rule to find $\frac{d}{dx}(2x^3k(x))$. Use $\frac{d}{dx}(3x+2)=3$.

$$=\frac{-6x^3k(x)+18x^3k(x)+12x^2k(x)+6x^4k'(x)+4x^3k'(x)}{(3x+2)^2}$$ Simplify

Exercise 3.3.9

Find $\dfrac{d}{dx}(3f(x)-2g(x))$.

Hint

Apply the difference rule and the constant multiple rule.

Answer

$3f'(x)-2g'(x)$.

Example 3.3.15: Determining Where a Function Has a Horizontal Tangent

Determine the values of x for which $f(x)=x^3-7x^2+8x+1$ has a horizontal tangent line.

Solution

To find the values of x for which $f(x)$ has a horizontal tangent line, we must solve $f'(x)=0$.

Since $f'(x)=3x^2-14x+8=(3x-2)(x-4)$,

we must solve $(3x-2)(x-4)=0$. Thus we see that the function has horizontal tangent lines at $x=\dfrac{2}{3}$ and $x=4$ as shown in the following graph.

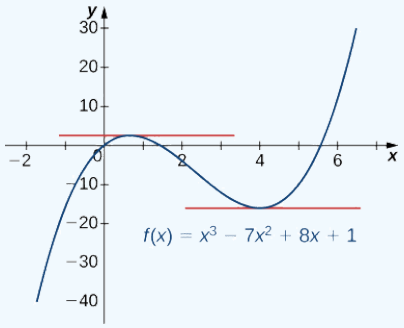


Figure 3.3.2: This function has horizontal tangent lines at $x=2/3$ and $x=4$.

Example 3.3.16: Finding a Velocity

The position of an object on a coordinate axis at time t is given by $s(t)=\dfrac{t}{t^2+1}$. What is the initial velocity of the object?

Solution

Since the initial velocity is $v(0)=s'(0)$, begin by finding $s'(t)$ by applying the quotient rule:

$$s'(t)=\frac{1(t^2+1)-2t(t)}{(t^2+1)^2}=\frac{1-t^2}{(t^2+1)^2}\ .$$

After evaluating, we see that $v(0)=1$.

Exercise 3.3.10

Find the values of x for which the line tangent to the graph of $f(x) = 4x^2 - 3x + 2$ has a tangent line parallel to the line $y = 2x + 3$.

Hint

Solve $f'(x) = 2$.

Answer

$\frac{5}{8}$

Formula One Grandstands

Formula One car races can be very exciting to watch and attract a lot of spectators. Formula One track designers have to ensure sufficient grandstand space is available around the track to accommodate these viewers. However, car racing can be dangerous, and safety considerations are paramount. The grandstands must be placed where spectators will not be in danger should a driver lose control of a car (Figure 3.3.3).



Figure 3.3.3: The grandstand next to a straightaway of the Circuit de Barcelona-Catalunya race track, located where the spectators are not in danger.

Safety is especially a concern on turns. If a driver does not slow down enough before entering the turn, the car may slide off the racetrack. Normally, this just results in a wider turn, which slows the driver down. But if the driver loses control completely, the car may fly off the track entirely, on a path tangent to the curve of the racetrack.

Suppose you are designing a new Formula One track. One section of the track can be modeled by the function $f(x) = x^3 + 3x^2 + x$ (Figure 3.3.4). The current plan calls for grandstands to be built along the first straightaway and around a portion of the first curve. The plans call for the front corner of the grandstand to be located at the point $(-1.9, 2.8)$. We want to determine whether this location puts the spectators in danger if a driver loses control of the car.

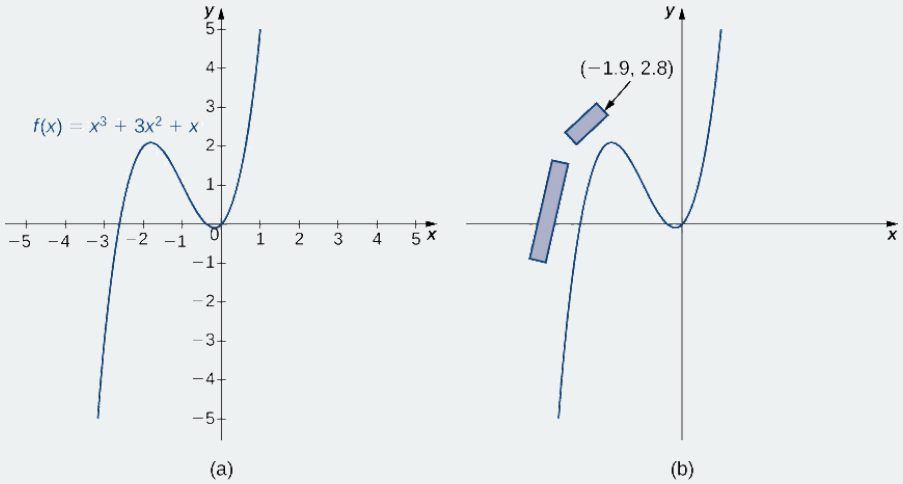


Figure 3.3.4: (a) One section of the racetrack can be modeled by the function $f(x) = x^3 + 3x^2 + x$. (b) The front corner of the grandstand is located at $(-1.9, 2.8)$.

1. Physicists have determined that drivers are most likely to lose control of their cars as they are coming into a turn, at the point where the slope of the tangent line is 1. Find the (x, y) coordinates of this point near the turn.
2. Find the equation of the tangent line to the curve at this point.
3. To determine whether the spectators are in danger in this scenario, find the x-coordinate of the point where the tangent line crosses the line $y = 2.8$. Is this point safely to the right of the grandstand? Or are the spectators in danger?

4. What if a driver loses control earlier than the physicists project? Suppose a driver loses control at the point $(-2.5, 0.625)$. What is the slope of the tangent line at this point?
5. If a driver loses control as described in part 4, are the spectators safe?
6. Should you proceed with the current design for the grandstand, or should the grandstands be moved?

Key Concepts

- The derivative of a constant function is zero.
- The derivative of a power function is a function in which the power on x becomes the coefficient of the term and the power on x in the derivative decreases by 1.
- The derivative of a constant c multiplied by a function f is the same as the constant multiplied by the derivative.
- The derivative of the sum of a function f and a function g is the same as the sum of the derivative of f and the derivative of g.
- The derivative of the difference of a function f and a function g is the same as the difference of the derivative of f and the derivative of g.
- The derivative of a product of two functions is the derivative of the first function times the second function plus the derivative of the second function times the first function.
- The derivative of the quotient of two functions is the derivative of the first function times the second function minus the derivative of the second function times the first function, all divided by the square of the second function.
- We used the limit definition of the derivative to develop formulas that allow us to find derivatives without resorting to the definition of the derivative. These formulas can be used singly or in combination with each other.

Glossary

constant multiple rule

the derivative of a constant c multiplied by a function f is the same as the constant multiplied by the derivative: $\frac{d}{dx}\big(cf(x)\big) = cf\,'(x)$

constant rule

the derivative of a constant function is zero: $\frac{d}{dx}(c) = 0$, where c is a constant

difference rule

the derivative of the difference of a function f and a function g is the same as the difference of the derivative of f and the derivative of g: $\frac{d}{dx}\big(f(x) - g(x)\big) = f\,'(x) - g'(x)$

power rule

the derivative of a power function is a function in which the power on x becomes the coefficient of the term and the power on x in the derivative decreases by 1: If n is an integer, then $\frac{d}{dx}(x^n) = nx^{n-1}$

product rule

the derivative of a product of two functions is the derivative of the first function times the second function plus the derivative of the second function times the first function: $\frac{d}{dx}\big(f(x)g(x)\big) = f\,'(x)g(x) + g'(x)f(x)$

quotient rule

the derivative of the quotient of two functions is the derivative of the first function times the second function minus the derivative of the second function times the first function, all divided by the square of the second function: $\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f\,'(x)g(x) - g'(x)f(x)}{\big(g(x)\big)^2}$

sum rule

the derivative of the sum of a function f and a function g is the same as the sum of the derivative of f and the derivative of g: $\frac{d}{dx}\big(f(x) + g(x)\big) = f\,'(x) + g'(x)$

3.3: Differentiation Rules is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **3.3: Differentiation Rules** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

3.4: The Derivative as a Rate of Change

Learning Objectives

- Determine a new value of a quantity from the old value and the amount of change.
- Calculate the average rate of change and explain how it differs from the instantaneous rate of change.
- Apply rates of change to displacement, velocity, and acceleration of an object moving along a straight line.
- Predict the future population from the present value and the population growth rate.
- Use derivatives to calculate marginal cost and revenue in a business situation.

In this section we look at some applications of the derivative by focusing on the interpretation of the derivative as the rate of change of a function. These applications include **acceleration** and velocity in physics, **population growth rates** in biology, and marginal functions in economics.

Amount of Change Formula

One application for derivatives is to estimate an unknown value of a function at a point by using a known value of a function at some given point together with its rate of change at the given point. If $f(x)$ is a function defined on an interval $[a, a+h]$, then the **amount of change** of $f(x)$ over the interval is the change in the y values of the function over that interval and is given by

$$f(a+h)-f(a).$$

The **average rate of change** of the function f over that same interval is the ratio of the amount of change over that interval to the corresponding change in the x values. It is given by

$$\frac{f(a+h)-f(a)}{h}.$$

As we already know, the instantaneous rate of change of $f(x)$ at a is its derivative

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

For small enough values of h, $f\,'(a)\approx\frac{f(a+h)-f(a)}{h}$. We can then solve for $f(a+h)$ to get the amount of change formula:

$$f(a+h)\approx f(a)+f\,'(a)h. \tag{3.4.1}$$

We can use this formula if we know only $f(a)$and $f\,'(a)$ and wish to estimate the value of $f(a+h)$. For example, we may use the current population of a city and the rate at which it is growing to estimate its population in the near future. As we can see in Figure 3.4.1, we are approximating $f(a+h)$ by the y coordinate at a+h on the line tangent to $f(x)$ at $x=a$. Observe that the accuracy of this estimate depends on the value of h as well as the value of $f\,'(a)$.

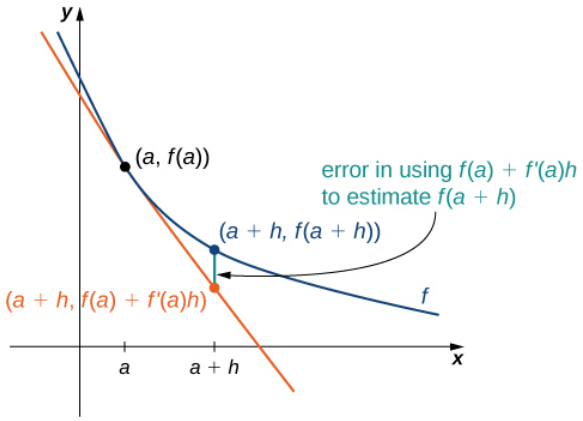


Figure 3.4.1: The new value of a changed quantity equals the original value plus the rate of change times the interval of change: $f(a+h)\approx f(a)+f\,'(a)h$.

Example 3.4.1: Estimating the Value of a Function

If $f(3)=2$ and $f'(3)=5$, estimate $f(3.2)$.

Solution

Begin by finding h. We have $h=3.2-3=0.2$. Thus,

$$f(3.2)=f(3+0.2)\approx f(3)+(0.2)f'(3)=2+0.2(5)=3.$$

Exercise 3.4.1

Given $f(10)=-5$ and $f'(10)=6$, estimate $f(10.1)$.

Hint

Use the same process as in the preceding example.

Answer

-4.4

Motion along a Line

Another use for the derivative is to analyze motion along a line. We have described velocity as the rate of change of position. If we take the derivative of the velocity, we can find the acceleration, or the rate of change of velocity. It is also important to introduce the idea of **speed**, which is the magnitude of velocity. Thus, we can state the following mathematical definitions.

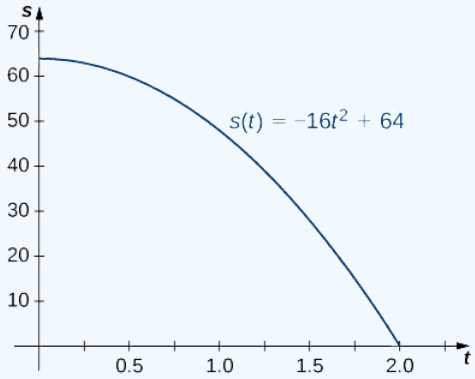
Definition

Let $s(t)$ be a function giving the position of an object at time t.

- The velocity of the object at time t is given by $v(t)=s'(t)$.
- The speed of the object at time t is given by $|v(t)|$.
- The acceleration of the object at t is given by $a(t)=v'(t)=s''(t)$.

Example 3.4.2: Comparing Instantaneous Velocity and Average Velocity

A ball is dropped from a height of 64 feet. Its height above ground (in feet) t seconds later is given by $s(t)=-16t^2+64$.



a. What is the instantaneous velocity of the ball when it hits the ground?
b. What is the average velocity during its fall?

Solution

The first thing to do is determine how long it takes the ball to reach the ground. To do this, set $s(t)=0$. Solving $-16t^2+64=0$, we get $t=2$, so it takes 2 seconds for the ball to reach the ground.

a. The instantaneous velocity of the ball as it strikes the ground is $v(2)$. Since $v(t) = s'(t) = -32t$, we obtain $v(t) = -64$ ft/s.

b. The average velocity of the ball during its fall is

$$v_{ave} = \frac{s(2)-s(0)}{2-0} = \frac{0-64}{2} = -32 \text{ ft/s.}$$

Example 3.4.3: Interpreting the Relationship between $v(t)$ and $a(t)$

A particle moves along a coordinate axis in the positive direction to the right. Its position at time t is given by $s(t) = t^3 - 4t + 2$. Find $v(1)$ and $a(1)$ and use these values to answer the following questions.

a. Is the particle moving from left to right or from right to left at time $t = 1$?

b. Is the particle speeding up or slowing down at time $t = 1$?

Solution

Begin by finding $v(t)$ and $a(t)$.

$v(t) = s'(t) = 3t^2 - 4$ and $a(t) = v'(t) = s''(t) = 6t$.

Evaluating these functions at $t = 1$, we obtain $v(1) = -1$ and $a(1) = 6$.

a. Because $v(1) < 0$, the particle is moving from right to left.

b. Because $v(1) < 0$ and $a(1) > 0$, velocity and acceleration are acting in opposite directions. In other words, the particle is being accelerated in the direction opposite the direction in which it is traveling, causing $|v(t)|$ to decrease. The particle is slowing down.

Example 3.4.4: Position and Velocity

The position of a particle moving along a coordinate axis is given by $s(t) = t^3 - 9t^2 + 24t + 4,\ t \geq 0$.

a. Find $v(t)$.

b. At what time(s) is the particle at rest?

c. On what time intervals is the particle moving from left to right? From right to left?

d. Use the information obtained to sketch the path of the particle along a coordinate axis.

Solution

a. The velocity is the derivative of the position function:

$$v(t) = s'(t) = 3t^2 - 18t + 24.$$

b. The particle is at rest when $v(t) = 0$, so set $3t^2 - 18t + 24 = 0$. Factoring the left-hand side of the equation produces $3(t-2)(t-4) = 0$. Solving, we find that the particle is at rest at $t = 2$ and $t = 4$.

c. The particle is moving from left to right when $v(t) > 0$ and from right to left when $v(t) < 0$. Figure 3.4.2 gives the analysis of the sign of $v(t)$ for $t \geq 0$, but it does not represent the axis along which the particle is moving.



Figure 3.4.2: The sign of $v(t)$ determines the direction of the particle.

- Since $3t^2 - 18t + 24 > 0$ on $[0, 2) \cup (4, +\infty)$, the particle is moving from left to right on these intervals.
- Since $3t^2 - 18t + 24 < 0$ on $(2, 4)$, the particle is moving from right to left on this interval.

d. Before we can sketch the graph of the particle, we need to know its position at the time it starts moving $(t = 0)$ and at the times that it changes direction $(t = 2, 4)$. We have $s(0) = 4$, $s(2) = 24$, and $s(4) = 20$. This means that the particle begins on the coordinate axis at 4 and changes direction at 24 and 20 on the coordinate axis. The path of the particle is shown on a coordinate axis in Figure 3.4.3.

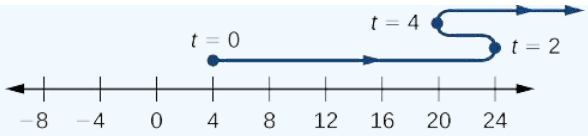


Figure 3.4.3: The path of the particle can be determined by analyzing $v(t)$.

Exercise 3.4.2

A particle moves along a coordinate axis. Its position at time t is given by $s(t) = t^2 - 5t + 1$. Is the particle moving from right to left or from left to right at time $t = 3$?

Hint

Find $v(3)$ and look at the sign.

Answer

left to right

Population Change

In addition to analyzing velocity, speed, acceleration, and position, we can use derivatives to analyze various types of populations, including those as diverse as bacteria colonies and cities. We can use a current population, together with a growth rate, to estimate the size of a population in the future. The population growth rate is the rate of change of a population and consequently can be represented by the derivative of the size of the population.

Definition

If $P(t)$ is the number of entities present in a population, then the population growth rate of $P(t)$ is defined to be $P'(t)$.

Example 3.4.5: Estimating a Population

The population of a city is tripling every 5 years. If its current population is 10,000, what will be its approximate population 2 years from now?

Solution

Let $P(t)$ be the population (in thousands) t years from now. Thus, we know that $P(0) = 10$ and based on the information, we anticipate $P(5) = 30$. Now estimate $P'(0)$, the current growth rate, using

$$P'(0) \approx \frac{P(5) - P(0)}{5 - 0} = \frac{30 - 10}{5} = 4 \, .$$

By applying Equation 3.4.1 to $P(t)$, we can estimate the population 2 years from now by writing

$$P(2) \approx P(0) + (2)P'(0) \approx 10 + 2(4) = 18 \, ;$$

thus, in 2 years the population will be 18,000.

Exercise 3.4.3

The current population of a mosquito colony is known to be 3,000; that is, $P(0) = 3,000$. If $P'(0) = 100$, estimate the size of the population in 3 days, where t is measured in days.

Hint

Use $P(3) \approx P(0) + 3P'(0)$

Answer

3,300

Changes in Cost and Revenue

In addition to analyzing motion along a line and population growth, derivatives are useful in analyzing changes in cost, revenue, and profit. The concept of a marginal function is common in the fields of business and economics and implies the use of derivatives. The marginal cost is the derivative of the cost function. The marginal revenue is the derivative of the revenue function. The marginal profit is the derivative of the profit function, which is based on the cost function and the revenue function.

Definition

- If $C(x)$ is the cost of producing x items, then the **marginal cost** $MC(x)$ is $MC(x) = C'(x)$.
- If $R(x)$ is the revenue obtained from selling x items, then the marginal revenue $MR(x)$ is $MR(x) = R'(x)$.
- If $P(x) = R(x) - C(x)$ is the profit obtained from selling x items, then the **marginal profit** $MP(x)$ is defined to be $MP(x) = P'(x) = MR(x) - MC(x) = R'(x) - C'(x)$.

We can roughly approximate

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

by choosing an appropriate value for h. Since x represents objects, a reasonable and small value for h is 1. Thus, by substituting $h = 1$, we get the approximation $MC(x) = C'(x) \approx C(x+1) - C(x)$. Consequently, $C'(x)$ for a given value of x can be thought of as the change in cost associated with producing one additional item. In a similar way, $MR(x) = R'(x)$ approximates the revenue obtained by selling one additional item, and $MP(x) = P'(x)$ approximates the profit obtained by producing and selling one additional item.

Example 3.4.6: Applying Marginal Revenue

Assume that the number of barbeque dinners that can be sold, x, can be related to the price charged, p, by the equation $p(x) = 9 - 0.03x, 0 \le x \le 300$.

In this case, the revenue in dollars obtained by selling x barbeque dinners is given by

$$R(x) = xp(x) = x(9 - 0.03x) = -0.03x^2 + 9x \quad \text{for } 0 \le x \le 300\,.$$

Use the marginal revenue function to estimate the revenue obtained from selling the 101st barbeque dinner. Compare this to the actual revenue obtained from the sale of this dinner.

Solution

First, find the marginal revenue function: $MR(x) = R'(x) = -0.06x + 9$.

Next, use $R'(100)$ to approximate $R(101) - R(100)$, the revenue obtained from the sale of the 101st dinner. Since $R'(100) = 3$, the revenue obtained from the sale of the 101st dinner is approximately \$3.

The actual revenue obtained from the sale of the 101st dinner is

$$R(101) - R(100) = 602.97 - 600 = 2.97, \text{ or } \$2.97.$$

The marginal revenue is a fairly good estimate in this case and has the advantage of being easy to compute.

Exercise 3.4.4

Suppose that the profit obtained from the sale of x fish-fry dinners is given by $P(x) = -0.03x^2 + 8x - 50$. Use the marginal profit function to estimate the profit from the sale of the 101st fish-fry dinner.

Hint

Use $P'(100)$ to approximate $P(101) - P(100)$.

Answer

\$2

Key Concepts

- Using $f(a+h) \approx f(a)+f'(a)h$, it is possible to estimate $f(a+h)$ given $f'(a)$ and $f(a)$.
- The rate of change of position is velocity, and the rate of change of velocity is acceleration. Speed is the absolute value, or magnitude, of velocity.
- The population growth rate and the present population can be used to predict the size of a future population.
- Marginal cost, marginal revenue, and marginal profit functions can be used to predict, respectively, the cost of producing one more item, the revenue obtained by selling one more item, and the profit obtained by producing and selling one more item.

Glossary

acceleration

is the rate of change of the velocity, that is, the derivative of velocity

amount of change

the amount of a function $f(x)$ over an interval $[x, x+h] is f(x+h)-f(x)$

average rate of change

is a function $f(x)$ over an interval $[x, x+h]$ is $\frac{f(x+h)-f(a)}{b-a}$

marginal cost

is the derivative of the cost function, or the approximate cost of producing one more item

marginal revenue

is the derivative of the revenue function, or the approximate revenue obtained by selling one more item

marginal profit

is the derivative of the profit function, or the approximate profit obtained by producing and selling one more item

population growth rate

is the derivative of the population with respect to time

speed

is the absolute value of velocity, that is, $|v(t)|$ is the speed of an object at time t whose velocity is given by $v(t)$

3.4: The Derivative as a Rate of Change is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **3.4: Derivatives as Rates of Change** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

3.5: Derivatives of Trigonometric Functions

Learning Objectives

- Find the derivatives of the sine and cosine function.
- Find the derivatives of the standard trigonometric functions.
- Calculate the higher-order derivatives of the sine and cosine.

One of the most important types of motion in physics is simple harmonic motion, which is associated with such systems as an object with mass oscillating on a spring. Simple harmonic motion can be described by using either sine or cosine functions. In this section we expand our knowledge of derivative formulas to include derivatives of these and other trigonometric functions. We begin with the derivatives of the sine and cosine functions and then use them to obtain formulas for the derivatives of the remaining four trigonometric functions. Being able to calculate the derivatives of the sine and cosine functions will enable us to find the velocity and acceleration of simple harmonic motion.

Derivatives of the Sine and Cosine Functions

We begin our exploration of the derivative for the sine function by using the formula to make a reasonable guess at its derivative. Recall that for a function $f(x)$,

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

Consequently, for values of h very close to 0,

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}.$$

We see that by using $h = 0.01$,

$$\frac{d}{dx}(\sin x) \approx \frac{\sin(x+0.01) - \sin x}{0.01}$$

By setting

$$D(x) = \frac{\sin(x+0.01) - \sin x}{0.01}$$

and using a graphing utility, we can get a graph of an approximation to the derivative of $\sin x$ (Figure 3.5.1).

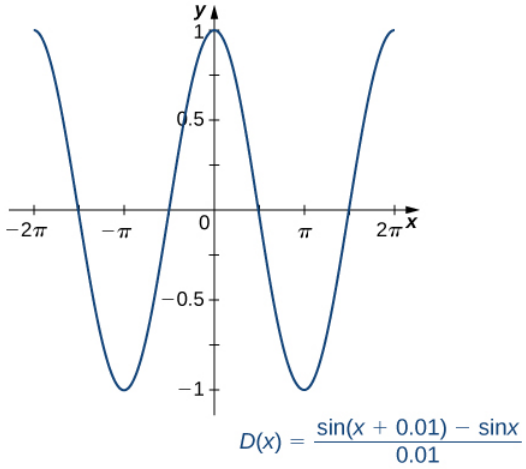


Figure 3.5.1: The graph of the function $D(x)$ looks a lot like a cosine curve.

Upon inspection, the graph of $D(x)$ appears to be very close to the graph of the cosine function. Indeed, we will show that

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

If we were to follow the same steps to approximate the derivative of the cosine function, we would find that

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

The Derivatives of $\sin x$ and $\cos x$

The derivative of the sine function is the cosine and the derivative of the cosine function is the negative sine.

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

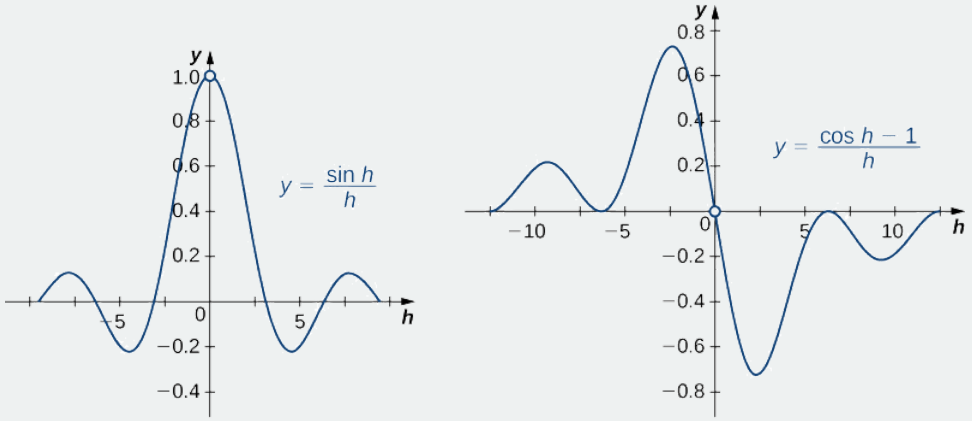
$$\frac{d}{dx}(\cos x) = -\sin x$$

Proof

Because the proofs for $\frac{d}{dx}(\sin x) = \cos x$ and $\frac{d}{dx}(\cos x) = -\sin x$ use similar techniques, we provide only the proof for $\frac{d}{dx}(\sin x) = \cos x$. Before beginning, recall two important trigonometric limits:

$$\lim_{h \to 0} \frac{\sin h}{h} = 1 \text{ and } \lim_{h \to 0} \frac{\cos h - 1}{h} = 0.$$

The graphs of $y = \frac{\sin h}{h}$ and $y = \frac{\cos h - 1}{h}$ are shown in Figure 3.5.2.



(a) (b)

Figure 3.5.2: These graphs show two important limits needed to establish the derivative formulas for the sine and cosine functions.

We also recall the following trigonometric identity for the sine of the sum of two angles:

$$\sin(x + h) = \sin x \cos h + \cos x \sin h.$$

Now that we have gathered all the necessary equations and identities, we proceed with the proof.

$$\begin{aligned}
\frac{d}{dx}(\sin x) &= \lim_{h\to 0}\frac{\sin(x+h)-\sin x}{h} && \text{Apply the definition of the derivative.}\\
&= \lim_{h\to 0}\frac{\sin x\cos h+\cos x\sin h-\sin x}{h} && \text{Use trig identity for the sine of the sum of two angles.}\\
&= \lim_{h\to 0}\left(\frac{\sin x\cos h-\sin x}{h}+\frac{\cos x\sin h}{h}\right) && \text{Regroup.}\\
&= \lim_{h\to 0}\left(\sin x\left(\frac{\cos h-1}{h}\right)+(\cos x)\left(\frac{\sin h}{h}\right)\right) && \text{Factor out } \sin x \text{ and } \cos x\\
&= (\sin x)\lim_{h\to 0}\left(\frac{\cos h-1}{h}\right)+(\cos x)\lim_{h\to 0}\left(\frac{\sin h}{h}\right) && \text{Factor } \sin x \text{ and } \cos x \text{ out of limits.}\\
&= (\sin x)(0)+(\cos x)(1) && \text{Apply trig limit formulas.}\\
&= \cos x && \text{Simplify.}
\end{aligned}$$

□

Figure 3.5.3 shows the relationship between the graph of $f(x)=\sin x$ and its derivative $f'(x)=\cos x$. Notice that at the points where $f(x)=\sin x$ has a horizontal tangent, its derivative $f'(x)=\cos x$ takes on the value zero. We also see that where f $(x)=\sin x$ is increasing, $f'(x)=\cos x>0$ and where $f(x)=\sin x$ is decreasing, $f'(x)=\cos x<0$.

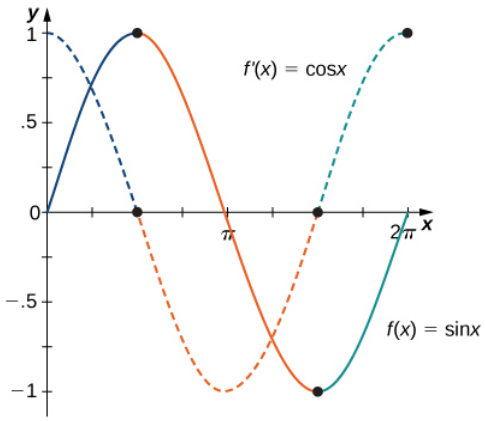


Figure 3.5.3: Where $f(x)$ has a maximum or a minimum, $f'(x)=0$ that is, $f'(x)=0$ where $f(x)$ has a horizontal tangent. These points are noted with dots on the graphs

Example 3.5.1: Differentiating a Function Containing $\sin x$

Find the derivative of $f(x)=5x^3\sin x$.

Solution

Using the product rule, we have

$$\begin{aligned}
f'(x) &= \frac{d}{dx}(5x^3)\cdot\sin x+\frac{d}{dx}(\sin x)\cdot 5x^3\\
&= 15x^2\cdot\sin x+\cos x\cdot 5x^3.
\end{aligned}$$

After simplifying, we obtain

$$f'(x)=15x^2\sin x+5x^3\cos x.$$

? Exercise 3.5.1

Find the derivative of $f(x) = \sin x \cos x$.

Hint

Don't forget to use the product rule.

Answer

$$f'(x) = \cos^2 x - \sin^2 x$$

✓ Example 3.5.2: Finding the Derivative of a Function Containing cos x

Find the derivative of $g(x) = \dfrac{\cos x}{4x^2}$.

Solution

By applying the quotient rule, we have

$$g'(x) = \frac{(-\sin x)4x^2 - 8x(\cos x)}{(4x^2)^2}.$$

Simplifying, we obtain

$$g'(x) = \frac{-4x^2 \sin x - 8x \cos x}{16x^4} = \frac{-x \sin x - 2 \cos x}{4x^3}.$$

? Exercise 3.5.2

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

Hint

Use the quotient rule.

Answer

$$f'(x) = \frac{\cos x + x \sin x}{\cos^2 x}$$

✓ Example 3.5.3: An Application to Velocity

A particle moves along a coordinate axis in such a way that its position at time t is given by $s(t) = 2 \sin t - t \quad \text{for } 0 \le t \le 2\pi$. At what times is the particle at rest?

Solution

To determine when the particle is at rest, set $s'(t) = v(t) = 0$. Begin by finding $s'(t)$. We obtain

$$s'(t) = 2 \cos t - 1,$$

so we must solve

$$2 \cos t - 1 = 0 \text{ for } 0 \le t \le 2\pi.$$

The solutions to this equation are $t = \dfrac{\pi}{3}$ and $t = \dfrac{5\pi}{3}$. Thus the particle is at rest at times $t = \dfrac{\pi}{3}$ and $t = \dfrac{5\pi}{3}$.

Exercise 3.5.3

A particle moves along a coordinate axis. Its position at time t is given by $s(t) = \sqrt{3}t + 2\cos t$ for $0 \le t \le 2\pi$. At what times is the particle at rest?

Hint

Use the previous example as a guide.

Answer

$t = \dfrac{\pi}{3}, \quad t = \dfrac{2\pi}{3}$

Derivatives of Other Trigonometric Functions

Since the remaining four trigonometric functions may be expressed as quotients involving sine, cosine, or both, we can use the quotient rule to find formulas for their derivatives.

Example 3.5.4: The Derivative of the Tangent Function

Find the derivative of $f(x) = \tan x$.

Solution

Start by expressing $\tan x$ as the quotient of $\sin x$ and $\cos x$:

$$f(x) = \tan x = \frac{\sin x}{\cos x}.$$

Now apply the quotient rule to obtain

$$f'(x) = \frac{\cos x \cos x - (-\sin x)\sin x}{(\cos x)^2}.$$

Simplifying, we obtain

$$f'(x) = \frac{\cos^2 x + \sin^2 x}{\cos^2 x}.$$

Recognizing that $\cos^2 x + \sin^2 x = 1$, by the Pythagorean theorem, we now have

$$f'(x) = \frac{1}{\cos^2 x}$$

Finally, use the identity $\sec x = \dfrac{1}{\cos x}$ to obtain

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

Exercise 3.5.4

Find the derivative of $f(x) = \cot x$.

Hint

Rewrite $\cot x$ as $\dfrac{\cos x}{\sin x}$ and use the quotient rule.

Answer

$f'(x) = -\csc^2 x$

The derivatives of the remaining trigonometric functions may be obtained by using similar techniques. We provide these formulas in the following theorem.

Derivatives of $\tan x$, $\cot x$, $\sec x$, and $\csc x$

The derivatives of the remaining trigonometric functions are as follows:

$$\frac{d}{dx}(\tan x) = \sec^2 x \tag{3.5.1}$$

$$\frac{d}{dx}(\cot x) = -\csc^2 x \tag{3.5.2}$$

$$\frac{d}{dx}(\sec x) = \sec x \tan x \tag{3.5.3}$$

$$\frac{d}{dx}(\csc x) = -\csc x \cot x. \tag{3.5.4}$$

Example 3.5.5: Finding the Equation of a Tangent Line

Find the equation of a line tangent to the graph of $f(x) = \cot x$ at $x = \frac{\pi}{4}$.

Solution

To find the equation of the tangent line, we need a point and a slope at that point. To find the point, compute

$$f\left(\tfrac{\pi}{4}\right) = \cot\tfrac{\pi}{4} = 1\,.$$

Thus the tangent line passes through the point $\left(\frac{\pi}{4}, 1\right)$. Next, find the slope by finding the derivative of $f(x) = \cot x$ and evaluating it at $\frac{\pi}{4}$:

$$f\,'(x) = -\csc^2 x \text{ and } f\,'\left(\tfrac{\pi}{4}\right) = -\csc^2\left(\tfrac{\pi}{4}\right) = -2\,.$$

Using the point-slope equation of the line, we obtain

$$y - 1 = -2\left(x - \tfrac{\pi}{4}\right)$$

or equivalently,

$$y = -2x + 1 + \tfrac{\pi}{2}\,.$$

Example 3.5.6: Finding the Derivative of Trigonometric Functions

Find the derivative of $f(x) = \csc x + x\tan x$.

Solution

To find this derivative, we must use both the sum rule and the product rule. Using the sum rule, we find

$$f\,'(x) = \frac{d}{dx}(\csc x) + \frac{d}{dx}(x\tan x).$$

In the first term, $\frac{d}{dx}(\csc x) = -\csc x \cot x$, and by applying the product rule to the second term we obtain

$$\frac{d}{dx}(x\tan x) = (1)(\tan x) + (\sec^2 x)(x).$$

Therefore, we have

$$f\,'(x) = -\csc x \cot x + \tan x + x\sec^2 x\,.$$

Exercise 3.5.5

Find the derivative of $f(x) = 2\tan x - 3\cot x$.

Hint

Use the rule for differentiating a constant multiple and the rule for differentiating a difference of two functions.

Answer

$f'(x) = 2\sec^2 x + 3\csc^2 x$

? Exercise 3.5.6

Find the slope of the line tangent to the graph of $f(x) = \tan x$ at $x = \dfrac{\pi}{6}$.

Hint

Evaluate the derivative at $x = \dfrac{\pi}{6}$.

Answer

$\dfrac{4}{3}$

Higher-Order Derivatives

The higher-order derivatives of $\sin x$ and $\cos x$ follow a repeating pattern. By following the pattern, we can find any higher-order derivative of $\sin x$ and $\cos x$.

✓ Example 3.5.7: Finding Higher-Order Derivatives of $y = \sin x$

Find the first four derivatives of $y = \sin x$.

Solution

Each step in the chain is straightforward:

$$\begin{aligned} y &= \sin x \\ \frac{dy}{dx} &= \cos x \\ \frac{d^2y}{dx^2} &= -\sin x \\ \frac{d^3y}{dx^3} &= -\cos x \\ \frac{d^4y}{dx^4} &= \sin x \end{aligned}$$

Analysis

Once we recognize the pattern of derivatives, we can find any higher-order derivative by determining the step in the pattern to which it corresponds. For example, every fourth derivative of $\sin x$ equals $\sin x$, so

$$\frac{d^4}{dx^4}(\sin x) = \frac{d^8}{dx^8}(\sin x) = \frac{d^{12}}{dx^{12}}(\sin x) = \ldots = \frac{d^{4n}}{dx^{4n}}(\sin x) = \sin x$$

$$\frac{d^5}{dx^5}(\sin x) = \frac{d^9}{dx^9}(\sin x) = \frac{d^{13}}{dx^{13}}(\sin x) = \ldots = \frac{d^{4n+1}}{dx^{4n+1}}(\sin x) = \cos x.$$

? Exercise 3.5.7

For $y = \cos x$, find $\dfrac{d^4y}{dx^4}$.

Hint

See the previous example.

Answer

$\cos x$

Example 3.5.8: Using the Pattern for Higher-Order Derivatives of $y = \sin x$

Find $\dfrac{d^{74}}{dx^{74}}(\sin x)$.

Solution

We can see right away that for the 74th derivative of $\sin x$, $74 = 4(18) + 2$, so

$$\frac{d^{74}}{dx^{74}}(\sin x) = \frac{d^{72+2}}{dx^{72+2}}(\sin x) = \frac{d^2}{dx^2}(\sin x) = -\sin x.$$

Exercise 3.5.8

For $y = \sin x$, find $\dfrac{d^{59}}{dx^{59}}(\sin x)$.

Hint

$$\frac{d^{59}}{dx^{59}}(\sin x) = \frac{d^{4\cdot 14+3}}{dx^{4\cdot 14+3}}(\sin x)$$

Answer

$-\cos x$

Example 3.5.9: An Application to Acceleration

A particle moves along a coordinate axis in such a way that its position at time t is given by $s(t) = 2 - \sin t$. Find $v(\pi/4)$ and $a(\pi/4)$. Compare these values and decide whether the particle is speeding up or slowing down.

Solution

First find $v(t) = s'(t)$

$$v(t) = s'(t) = -\cos t.$$

Thus,

$$v\left(\tfrac{\pi}{4}\right) = -\frac{1}{\sqrt{2}} = -\frac{\sqrt{2}}{2}.$$

Next, find $a(t) = v'(t)$. Thus, $a(t) = v'(t) = \sin t$ and we have

$$a\left(\tfrac{\pi}{4}\right) = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}.$$

Since $v\left(\tfrac{\pi}{4}\right) = -\dfrac{\sqrt{2}}{2} < 0$ and $a\left(\tfrac{\pi}{4}\right) = \dfrac{\sqrt{2}}{2} > 0$, we see that velocity and acceleration are acting in opposite directions; that is, the object is being accelerated in the direction opposite to the direction in which it is traveling. Consequently, the particle is slowing down.

Exercise 3.5.9

A block attached to a spring is moving vertically. Its position at time t is given by $s(t) = 2\sin t$. Find $v\left(\tfrac{5\pi}{6}\right)$ and $a\left(\tfrac{5\pi}{6}\right)$. Compare these values and decide whether the block is speeding up or slowing down.

Hint

Use Example 3.5.9as a guide.

Answer

$v\left(\frac{5\pi}{6}\right) = -\sqrt{3} < 0$ and $a\left(\frac{5\pi}{6}\right) = -1 < 0$. The block is speeding up.

Key Concepts

- We can find the derivatives of $\sin x$ and $\cos x$ by using the definition of derivative and the limit formulas found earlier. The results are

$$\frac{d}{dx}(\sin x) = \cos x \quad \text{and} \quad \frac{d}{dx}(\cos x) = -\sin x.$$

- With these two formulas, we can determine the derivatives of all six basic trigonometric functions.

Key Equations

- **Derivative of sine function**

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

- **Derivative of cosine function**

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

- **Derivative of tangent function**

$$\frac{d}{dx}(\tan x) = \sec^2 x$$

- **Derivative of cotangent function**

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

- **Derivative of secant function**

$$\frac{d}{dx}(\sec x) = \sec x \tan x$$

- **Derivative of cosecant function**

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

3.5: Derivatives of Trigonometric Functions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **3.5: Derivatives of Trigonometric Functions** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

3.6: The Chain Rule

Learning Objectives

- State the chain rule for the composition of two functions.
- Apply the chain rule together with the power rule.
- Apply the chain rule and the product/quotient rules correctly in combination when both are necessary.
- Recognize the chain rule for a composition of three or more functions.
- Describe the proof of the chain rule.

We have seen the techniques for differentiating basic functions ($x^n, \sin x, \cos x$, etc.) as well as sums, differences, products, quotients, and constant multiples of these functions. However, these techniques do not allow us to differentiate compositions of functions, such as $h(x) = \sin(x^3)$ or $k(x) = \sqrt{3x^2+1}$. In this section, we study the rule for finding the derivative of the composition of two or more functions.

Deriving the Chain Rule

When we have a function that is a composition of two or more functions, we could use all of the techniques we have already learned to differentiate it. However, using all of those techniques to break down a function into simpler parts that we are able to differentiate can get cumbersome. Instead, we use the *chain rule*, which states that the derivative of a composite function is the derivative of the outer function evaluated at the inner function times the derivative of the inner function.

To put this rule into context, let's take a look at an example: $h(x) = \sin(x^3)$. We can think of the derivative of this function with respect to x as the rate of change of $\sin(x^3)$ relative to the change in x. Consequently, we want to know how $\sin(x^3)$ changes as x changes. We can think of this event as a chain reaction: As x changes, x^3 changes, which leads to a change in $\sin(x^3)$. This chain reaction gives us hints as to what is involved in computing the derivative of $\sin(x^3)$. First of all, a change in x forcing a change in x^3 suggests that somehow the derivative of x^3 is involved. In addition, the change in x^3 forcing a change in $\sin(x^3)$ suggests that the derivative of $\sin(u)$ with respect to u, where $u = x^3$, is also part of the final derivative.

We can take a more formal look at the derivative of $h(x) = \sin(x^3)$ by setting up the limit that would give us the derivative at a specific value a in the domain of $h(x) = \sin(x^3)$.

$$h'(a) = \lim_{x \to a} \frac{\sin(x^3) - \sin(a^3)}{x - a}$$

This expression does not seem particularly helpful; however, we can modify it by multiplying and dividing by the expression $x^3 - a^3$ to obtain

$$h'(a) = \lim_{x \to a} \frac{\sin(x^3) - \sin(a^3)}{x^3 - a^3} \cdot \frac{x^3 - a^3}{x - a}.$$

From the definition of the derivative, we can see that the second factor is the derivative of x^3 at $x = a$. That is,

$$\lim_{x \to a} \frac{x^3 - a^3}{x - a} = \frac{d}{dx}(x^3)\Big|_{x=a} = 3a^2.$$

However, it might be a little more challenging to recognize that the first term is also a derivative. We can see this by letting $u = x^3$ and observing that as $x \to a, u \to a^3$:

$$\begin{aligned} \lim_{x \to a} \frac{\sin(x^3) - \sin(a^3)}{x^3 - a^3} &= \lim_{u \to a^3} \frac{\sin u - \sin(a^3)}{u - a^3} \\ &= \frac{d}{du}(\sin u)\Big|_{u=a^3} \\ &= \cos(a^3) \end{aligned}.$$

Thus, $h'(a) = \cos(a^3) \cdot 3a^2$.

In other words, if $h(x) = \sin(x^3)$, then $h'(x) = \cos(x^3) \cdot 3x^2$. Thus, if we think of $h(x) = \sin(x^3)$ as the composition $(f \circ g)(x) = f\big(g(x)\big)$ where $f(x) = \sin x$ and $g(x) = x^3$, then the derivative of $h(x) = \sin(x^3)$ is the product of the derivative of $g(x) = x^3$ and the derivative of the function $f(x) = \sin x$ evaluated at the function $g(x) = x^3$. At this point, we anticipate that for $h(x) = \sin\big(g(x)\big)$, it is quite likely that $h'(x) = \cos\big(g(x)\big)g'(x)$. As we determined above, this is the case for $h(x) = \sin(x^3)$.

Now that we have derived a special case of the chain rule, we state the general case and then apply it in a general form to other composite functions. An informal proof is provided at the end of the section.

Rule: The Chain Rule

Let f and g be functions. For all x in the domain of g for which g is differentiable at x and f is differentiable at $g(x)$, the derivative of the composite function

$$h(x) = (f \circ g)(x) = f\big(g(x)\big)$$

is given by

$$h'(x) = f'\big(g(x)\big) \cdot g'(x).$$

Alternatively, if y is a function of u, and u is a function of x, then

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}.$$

Problem-Solving Strategy: Applying the Chain Rule

1. To differentiate $h(x) = f\big(g(x)\big)$, begin by identifying $f(x)$ and $g(x)$.
2. Find $f'(x)$ and evaluate it at $g(x)$ to obtain $f'\big(g(x)\big)$.
3. Find $g'(x)$.
4. Write $h'(x) = f'\big(g(x)\big) \cdot g'(x)$.

Note: When applying the chain rule to the composition of two or more functions, keep in mind that we work our way from the outside function in. It is also useful to remember that the derivative of the composition of two functions can be thought of as having two parts; the derivative of the composition of three functions has three parts; and so on. Also, remember that we never evaluate a derivative at a derivative.

The Chain and Power Rules Combined

We can now apply the chain rule to composite functions, but note that we often need to use it with other rules. For example, to find derivatives of functions of the form $h(x) = \big(g(x)\big)^n$, we need to use the chain rule combined with the power rule. To do so, we can think of $h(x) = \big(g(x)\big)^n$ as $f\big(g(x)\big)$ where $f(x) = x^n$. Then $f'(x) = nx^{n-1}$. Thus, $f'\big(g(x)\big) = n\big(g(x)\big)^{n-1}$. This leads us to the derivative of a power function using the chain rule,

$$h'(x) = n\big(g(x)\big)^{n-1} \cdot g'(x)$$

Rule: Power Rule for Composition of Functions (General Power Rule)

For all values of x for which the derivative is defined, if

$$h(x) = \big(g(x)\big)^n,$$

Then

$$h'(x) = n\big(g(x)\big)^{n-1} \cdot g'(x). \tag{3.6.1}$$

Example 3.6.1: Using the Chain and Power Rules

Find the derivative of $h(x) = \dfrac{1}{(3x^2+1)^2}$.

Solution

First, rewrite $h(x) = \dfrac{1}{(3x^2+1)^2} = (3x^2+1)^{-2}$.

Applying the power rule with $g(x) = 3x^2+1$, we have

$$h'(x) = -2(3x^2+1)^{-3} \cdot 6x\,.$$

Rewriting back to the original form gives us

$$h'(x) = \frac{-12x}{(3x^2+1)^3}$$

Exercise 3.6.1

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

Hint

Use the General Power Rule (Equation 3.6.1) with $g(x) = 2x^3+2x-1$.

Answer

$h'(x) = 4(2x^3+2x-1)^3(6x^2+2) = 8(3x^2+1)(2x^3+2x-1)^3$

Example 3.6.2: Using the Chain and Power Rules with a Trigonometric Function

Find the derivative of $h(x) = \sin^3 x$.

Solution

First recall that $\sin^3 x = (\sin x)^3$, so we can rewrite $h(x) = \sin^3 x$ as $h(x) = (\sin x)^3$.

Applying the power rule with $g(x) = \sin x$, we obtain

$$h'(x) = 3(\sin x)^2 \cos x = 3\sin^2 x \cos x.$$

Example 3.6.3: Finding the Equation of a Tangent Line

Find the equation of a line tangent to the graph of $h(x) = \dfrac{1}{(3x-5)^2}$ at $x = 2$.

Solution

Because we are finding an equation of a line, we need a point. The x-coordinate of the point is 2. To find the y-coordinate, substitute 2 into $h(x)$. Since $h(2) = \dfrac{1}{(3(2)-5)^2} = 1$, the point is $(2, 1)$.

For the slope, we need $h'(2)$. To find $h'(x)$, first we rewrite $h(x) = (3x-5)^{-2}$ and apply the power rule to obtain

$$h'(x) = -2(3x-5)^{-3}(3) = -6(3x-5)^{-3}\,.$$

By substituting, we have $h'(2) = -6(3(2)-5)^{-3} = -6$.

Therefore, the line has equation $y-1 = -6(x-2)$. Rewriting, the equation of the line is $y = -6x+13$.

? Exercise 3.6.2

Find the equation of the line tangent to the graph of $f(x) = (x^2 - 2)^3$ at $x = -2$.

Hint

Use the preceding example as a guide.

Answer

$y = -48x - 88$

Combining the Chain Rule with Other Rules

Now that we can combine the chain rule and the power rule, we examine how to combine the chain rule with the other rules we have learned. In particular, we can use it with the formulas for the derivatives of trigonometric functions or with the product rule.

✓ Example 3.6.4: Using the Chain Rule on a General Cosine Function

Find the derivative of $h(x) = \cos\big(g(x)\big)$.

Solution

Think of $h(x) = \cos\big(g(x)\big)$ as $f\big(g(x)\big)$ where $f(x) = \cos x$. Since $f'(x) = -\sin x$, we have $f'\big(g(x)\big) = -\sin\big(g(x)\big)$. Then we do the following calculation.

$$\begin{aligned} h'(x) &= f'\big(g(x)\big) \cdot g'(x) && \text{Apply the chain rule.} \\ &= -\sin\big(g(x)\big) \cdot g'(x) && \text{Substitute } f'\big(g(x)\big) = -\sin\big(g(x)\big). \end{aligned}$$

Thus, the derivative of $h(x) = \cos\big(g(x)\big)$ is given by $h'(x) = -\sin\big(g(x)\big) \cdot g'(x)$.

In the following example we apply the rule that we have just derived.

✓ Example 3.6.5: Using the Chain Rule on a Cosine Function

Find the derivative of $h(x) = \cos(5x^2)$.

Solution

Let $g(x) = 5x^2$. Then $g'(x) = 10x$. Using the result from the previous example,

$$h'(x) = -\sin(5x^2) \cdot 10x = -10x \sin(5x^2)$$

✓ Example 3.6.6: Using the Chain Rule on Another Trigonometric Function

Find the derivative of $h(x) = \sec(4x^5 + 2x)$.

Solution

Apply the chain rule to $h(x) = \sec\big(g(x)\big)$ to obtain

$$h'(x) = \sec(g(x)) \tan\big(g(x)\big) \cdot g'(x).$$

In this problem, $g(x) = 4x^5 + 2x,$ so we have $g'(x) = 20x^4 + 2.$ Therefore, we obtain

$$h'(x) = \sec(4x^5 + 2x) \tan(4x^5 + 2x)(20x^4 + 2) = (20x^4 + 2)\sec(4x^5 + 2x) \tan(4x^5 + 2x).$$

? Exercise 3.6.3

Find the derivative of $h(x) = \sin(7x + 2)$.

Hint

Apply the chain rule to $h(x)=\sin\big(g(x)\big)$ first and then use $g(x)=7x+2$.

Answer

$h'(x)=7\cos(7x+2)$

At this point we provide a list of derivative formulas that may be obtained by applying the chain rule in conjunction with the formulas for derivatives of trigonometric functions. Their derivations are similar to those used in the examples above. For convenience, formulas are also given in Leibniz's notation, which some students find easier to remember. (We discuss the chain rule using Leibniz's notation at the end of this section.) It is not absolutely necessary to memorize these as separate formulas as they are all applications of the chain rule to previously learned formulas.

Using the Chain Rule with Trigonometric Functions

For all values of x for which the derivative is defined,

$\frac{d}{dx}\Big(\sin(g(x))\Big)=\cos(g(x))\cdot g'(x)$	$\frac{d}{dx}\Big(\sin u\Big)=\cos u\cdot\frac{du}{dx}$
$\frac{d}{dx}\Big(\cos(g(x))\Big)=-\sin(g(x))\cdot g'(x)$	$\frac{d}{dx}\Big(\cos u\Big)=-\sin u\cdot\frac{du}{dx}$
$\frac{d}{dx}\Big(\tan(g(x))\Big)=\sec^2(g(x))\cdot g'(x)$	$\frac{d}{dx}\Big(\tan u\Big)=\sec^2 u\cdot\frac{du}{dx}$
$\frac{d}{dx}\Big(\cot(g(x))\Big)=-\csc^2(g(x))\cdot g'(x)$	$\frac{d}{dx}\Big(\cot u\Big)=-\csc^2 u\cdot\frac{du}{dx}$
$\frac{d}{dx}\Big(\sec(g(x))\Big)=\sec(g(x))\tan(g(x))\cdot g'(x)$	$\frac{d}{dx}\Big(\sec u\Big)=\sec u\tan u\cdot\frac{du}{dx}$
$\frac{d}{dx}\Big(\csc(g(x))\Big)=-\csc(g(x))\cot(g(x))\cdot g'(x)$	$\frac{d}{dx}\Big(\csc u\Big)=-\csc u\cot u\cdot\frac{du}{dx}.$

Example 3.6.7: Combining the Chain Rule with the Product Rule

Find the derivative of $h(x)=(2x+1)^5(3x-2)^7$.

Solution

First apply the product rule, then apply the chain rule to each term of the product.

$$\begin{aligned}
h'(x)&=\frac{d}{dx}\big((2x+1)^5\big)\cdot(3x-2)^7+\frac{d}{dx}\big((3x-2)^7\big)\cdot(2x+1)^5 && \text{Apply the product rule.}\\
&=5(2x+1)^4\cdot 2\cdot(3x-2)^7+7(3x-2)^6\cdot 3\cdot(2x+1)^5 && \text{Apply the chain rule.}\\
&=10(2x+1)^4(3x-2)^7+21(3x-2)^6(2x+1)^5 && \text{Simplify.}\\
&=(2x+1)^4(3x-2)^6(10(3x-2)+21(2x+1)) && \text{Factor out } (2x+1)^4(3x-2)^6\\
&=(2x+1)^4(3x-2)^6(72x+1) && \text{Simplify.}
\end{aligned}$$

Exercise 3.6.4

Find the derivative of $h(x)=\dfrac{x}{(2x+3)^3}$.

Hint

Start out by applying the quotient rule. Remember to use the chain rule to differentiate the denominator.

Answer

$$h'(x) = \frac{3-4x}{(2x+3)^4}$$

Composites of Three or More Functions

We can now combine the chain rule with other rules for differentiating functions, but when we are differentiating the composition of three or more functions, we need to apply the chain rule more than once. If we look at this situation in general terms, we can generate a formula, but we do not need to remember it, as we can simply apply the chain rule multiple times.

In general terms, first we let

$$k(x) = h\Big(f\big(g(x)\big)\Big).$$

Then, applying the chain rule once we obtain

$$k'(x) = \frac{d}{dx}\Big(h\big(f\big(g(x)\big)\big)\Big) = h'\Big(f\big(g(x)\big)\Big) \cdot \frac{d}{dx}\Big(f\big(g(x)\big)\Big).$$

Applying the chain rule again, we obtain

$$k'(x) = h'\Big(f\big(g(x)\big)\Big) \cdot f'\big(g(x)\big) \cdot g'(x).$$

Rule: Chain Rule for a Composition of Three Functions

For all values of x for which the function is differentiable, if

$$k(x) = h\Big(f\big(g(x)\big)\Big),$$

then

$$k'(x) = h'\Big(f\big(g(x)\big)\Big) \cdot f'\big(g(x)\big) \cdot g'(x).$$

In other words, we are applying the chain rule twice.

Notice that the derivative of the composition of three functions has three parts. (Similarly, the derivative of the composition of four functions has four parts, and so on.) Also, *remember, we can always work from the outside in, taking one derivative at a time.*

Example 3.6.8: Differentiating a Composite of Three Functions

Find the derivative of $k(x) = \cos^4(7x^2+1)$.

Solution

First, rewrite $k(x)$ as

$$k(x) = \big(\cos(7x^2+1)\big)^4.$$

Then apply the chain rule several times.

$$\begin{aligned} k'(x) &= 4(\cos(7x^2+1))^3 \cdot \frac{d}{dx}\big(\cos(7x^2+1)\big) && \text{Apply the chain rule.} \\ &= 4(\cos(7x^2+1))^3(-\sin(7x^2+1)) \cdot \frac{d}{dx}\big(7x^2+1\big) && \text{Apply the chain rule.} \\ &= 4(\cos(7x^2+1))^3(-\sin(7x^2+1))(14x) && \text{Apply the chain rule.} \\ &= -56x\sin(7x^2+1)\cos^3(7x^2+1) && \text{Simplify} \end{aligned}$$

Exercise 3.6.5

Find the derivative of $h(x) = \sin^6(x^3)$.

Hint

Rewrite $h(x) = \sin^6(x^3) = \left(\sin(x^3)\right)^6$ and use Example 3.6.8as a guide.

Answer

$h'(x) = 18x^2 \sin^5(x^3)\cos(x^3)$

Example 3.6.9: Using the Chain Rule in a Velocity Problem

A particle moves along a coordinate axis. Its position at time t is given by $s(t) = \sin(2t) + \cos(3t)$. What is the velocity of the particle at time $t = \dfrac{\pi}{6}$?

Solution

To find $v(t)$, the velocity of the particle at time t, we must differentiate $s(t)$. Thus,

$$v(t) = s'(t) = 2\cos(2t) - 3\sin(3t).$$

Proof of Chain Rule

At this point, we present a very informal proof of the chain rule. For simplicity's sake we ignore certain issues: For example, we assume that $g(x) \neq g(a)$ for $x \neq a$ in some open interval containing a. We begin by applying the limit definition of the derivative to the function $h(x)$ to obtain $h'(a)$:

$$h'(a) = \lim_{x \to a} \frac{f\big(g(x)\big) - f\big(g(a)\big)}{x - a}.$$

Rewriting, we obtain

$$h'(a) = \lim_{x \to a} \frac{f\big(g(x)\big) - f\big(g(a)\big)}{g(x) - g(a)} \cdot \frac{g(x) - g(a)}{x - a}.$$

Although it is clear that

$$\lim_{x \to a} \frac{g(x) - g(a)}{x - a} = g'(a),$$

it is not obvious that

$$\lim_{x \to a} \frac{f\big(g(x)\big) - f\big(g(a)\big)}{g(x) - g(a)} = f'\big(g(a)\big).$$

To see that this is true, first recall that since g is differentiable at a, g is also continuous at a. Thus,

$$\lim_{x \to a} g(x) = g(a).$$

Next, make the substitution $y = g(x)$and $b = g(a)$ and use change of variables in the limit to obtain

$$\lim_{x \to a} \frac{f\big(g(x)\big) - f\big(g(a)\big)}{g(x) - g(a)} = \lim_{y \to b} \frac{f(y) - f(b)}{y - b} = f'(b) = f'\big(g(a)\big).$$

Finally,

$$h'(a) = \lim_{x \to a} \frac{f\big(g(x)\big) - f\big(g(a)\big)}{g(x) - g(a)} \cdot \frac{g(x) - g(a)}{x - a} = f'\big(g(a)\big) \cdot g'(a).$$

□

Example 3.6.10: Using the Chain Rule with Functional Values

Let $h(x) = f\big(g(x)\big)$. If $g(1) = 4, g'(1) = 3$, and $f'(4) = 7$, find $h'(1)$.

Solution

Use the chain rule, then substitute.

$$\begin{aligned} h'(1) &= f'\big(g(1)\big) \cdot g'(1) && \text{Apply the chain rule.} \\ &= f'(4) \cdot 3 && \text{Substitute } g(1) = 4 \text{ and } g'(1) = 3. \\ &= 7 \cdot 3 && \text{Substitute } f'(4) = 7. \\ &= 21 && \text{Simplify.} \end{aligned}$$

Exercise 3.6.6

Given $h(x) = f(g(x))$. If $g(2) = -3, g'(2) = 4,$ and $f'(-3) = 7$, find $h'(2)$.

Hint

Follow Example 3.6.10

Answer

28

The Chain Rule Using Leibniz's Notation

As with other derivatives that we have seen, we can express the chain rule using Leibniz's notation. This notation for the chain rule is used heavily in physics applications.

For $h(x) = f(g(x))$, let $u = g(x)$ and $y = h(x) = f(u)$. Thus,

$$h'(x) = \frac{dy}{dx}$$

$$f'(g(x)) = f'(u) = \frac{dy}{du}$$

and

$$g'(x) = \frac{du}{dx}.$$

Consequently,

$$\frac{dy}{dx} = h'(x) = f'\big(g(x)\big) \cdot g'(x) = \frac{dy}{du} \cdot \frac{du}{dx}.$$

Rule: Chain Rule Using Leibniz's Notation

If y is a function of u, and u is a function of x, then

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}.$$

Example 3.6.11: Taking a Derivative Using Leibniz's Notation I

Find the derivative of $y = \left(\frac{x}{3x+2}\right)^5$.

Solution

First, let $u = \frac{x}{3x+2}$. Thus, $y = u^5$. Next, find $\frac{du}{dx}$ and $\frac{dy}{du}$. Using the quotient rule,

$$\frac{du}{dx} = \frac{2}{(3x+2)^2}$$

and

$$\frac{dy}{du} = 5u^4.$$

Finally, we put it all together.

$$\begin{aligned}
\frac{dy}{dx} &= \frac{dy}{du}\cdot\frac{du}{dx} && \text{Apply the chain rule.}\\
&= 5u^4\cdot\frac{2}{(3x+2)^2} && \text{Substitute } \frac{dy}{du} = 5u^4 \text{ and } \frac{du}{dx} = \frac{2}{(3x+2)^2}.\\
&= 5\left(\frac{x}{3x+2}\right)^4\cdot\frac{2}{(3x+2)^2} && \text{Substitute } u = \frac{x}{3x+2}.\\
&= \frac{10x^4}{(3x+2)^6} && \text{Simplify.}
\end{aligned}$$

It is important to remember that, when using the Leibniz form of the chain rule, the final answer must be expressed entirely in terms of the original variable given in the problem.

Example 3.6.12: Taking a Derivative Using Leibniz's Notation II

Find the derivative of $y = \tan(4x^2 - 3x + 1)$.

Solution

First, let $u = 4x^2 - 3x + 1$. Then $y = \tan u$. Next, find $\frac{du}{dx}$ and $\frac{dy}{du}$:

$$\frac{du}{dx} = 8x - 3 \text{ and } \frac{dy}{du} = \sec^2 u.$$

Finally, we put it all together.

$$\begin{aligned}
\frac{dy}{dx} &= \frac{dy}{du}\cdot\frac{du}{dx} && \text{Apply the chain rule.}\\
&= \sec^2 u\cdot(8x-3) && \text{Use } \frac{du}{dx} = 8x-3 \text{ and } \frac{dy}{du} = \sec^2 u.\\
&= \sec^2(4x^2-3x+1)\cdot(8x-3) && \text{Substitute } u = 4x^2-3x+1.
\end{aligned}$$

Exercise 3.6.7

Use Leibniz's notation to find the derivative of $y = \cos(x^3)$. Make sure that the final answer is expressed entirely in terms of the variable x.

Hint

Let $u = x^3$.

Answer

$\frac{dy}{dx} = -3x^2 \sin(x^3)$.

Key Concepts

- The chain rule allows us to differentiate compositions of two or more functions. It states that for $h(x) = f\big(g(x)\big)$,

$$h'(x) = f'\big(g(x)\big) \cdot g'(x).$$

  In Leibniz's notation this rule takes the form

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}\,.$$

- We can use the chain rule with other rules that we have learned, and we can derive formulas for some of them.
- The chain rule combines with the power rule to form a new rule:

$$\text{If } h(x) = \big(g(x)\big)^n, \text{ then } h'(x) = n\big(g(x)\big)^{n-1} \cdot g'(x)\,.$$

- When applied to the composition of three functions, the chain rule can be expressed as follows: If $h(x) = f\Big(g\big(k(x)\big)\Big)$, then $h'(x) = f'\Big(g\big(k(x)\big)\Big) \cdot g'\big(k(x)\big) \cdot k'(x)$.

Key Equations

- **The chain rule**

$$h'(x) = f'\big(g(x)\big) \cdot g'(x)$$

- **The power rule for functions**

$$h'(x) = n\big(g(x)\big)^{n-1} \cdot g'(x)$$

Glossary

chain rule

the chain rule defines the derivative of a composite function as the derivative of the outer function evaluated at the inner function times the derivative of the inner function

This page titled 3.6: The Chain Rule is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin "Jed" Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **3.6: The Chain Rule** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

3.7: Implicit Differentiation

Learning Objectives

- Find the derivative of a complicated function by using implicit differentiation.
- Use implicit differentiation to determine the equation of a tangent line.

We have already studied how to find equations of tangent lines to functions and the rate of change of a function at a specific point. In all these cases we had the explicit equation for the function and differentiated these functions explicitly. Suppose instead that we want to determine the equation of a tangent line to an arbitrary curve or the rate of change of an arbitrary curve at a point. In this section, we solve these problems by finding the derivatives of functions that define y implicitly in terms of x.

Implicit Differentiation

In most discussions of math, if the dependent variable y is a function of the independent variable x, we express y in terms of x. If this is the case, we say that y is an explicit function of x. For example, when we write the equation $y = x^2 + 1$, we are defining y explicitly in terms of x. On the other hand, if the relationship between the function y and the variable x is expressed by an equation where y is not expressed entirely in terms of x, we say that the equation defines y implicitly in terms of x. For example, the equation $y - x^2 = 1$ defines the function $y = x^2 + 1$ implicitly.

Implicit differentiation allows us to find slopes of tangents to curves that are clearly not functions (they fail the vertical line test). We are using the idea that portions of y are functions that satisfy the given equation, but that y is not actually a function of x.

In general, an equation defines a function implicitly if the function satisfies that equation. An equation may define many different functions implicitly. For example, the functions

$$y = \sqrt{25 - x^2}$$

and

$$y = \begin{cases} \sqrt{25 - x^2}, & \text{if } -5 \le x < 0 \\ -\sqrt{25 - x^2}, & \text{if } 0 \le x \le 5 \end{cases}$$

which are illustrated in Figure 3.7.1, are just two of the many functions defined implicitly by the equation $x^2 + y^2 = 25$.

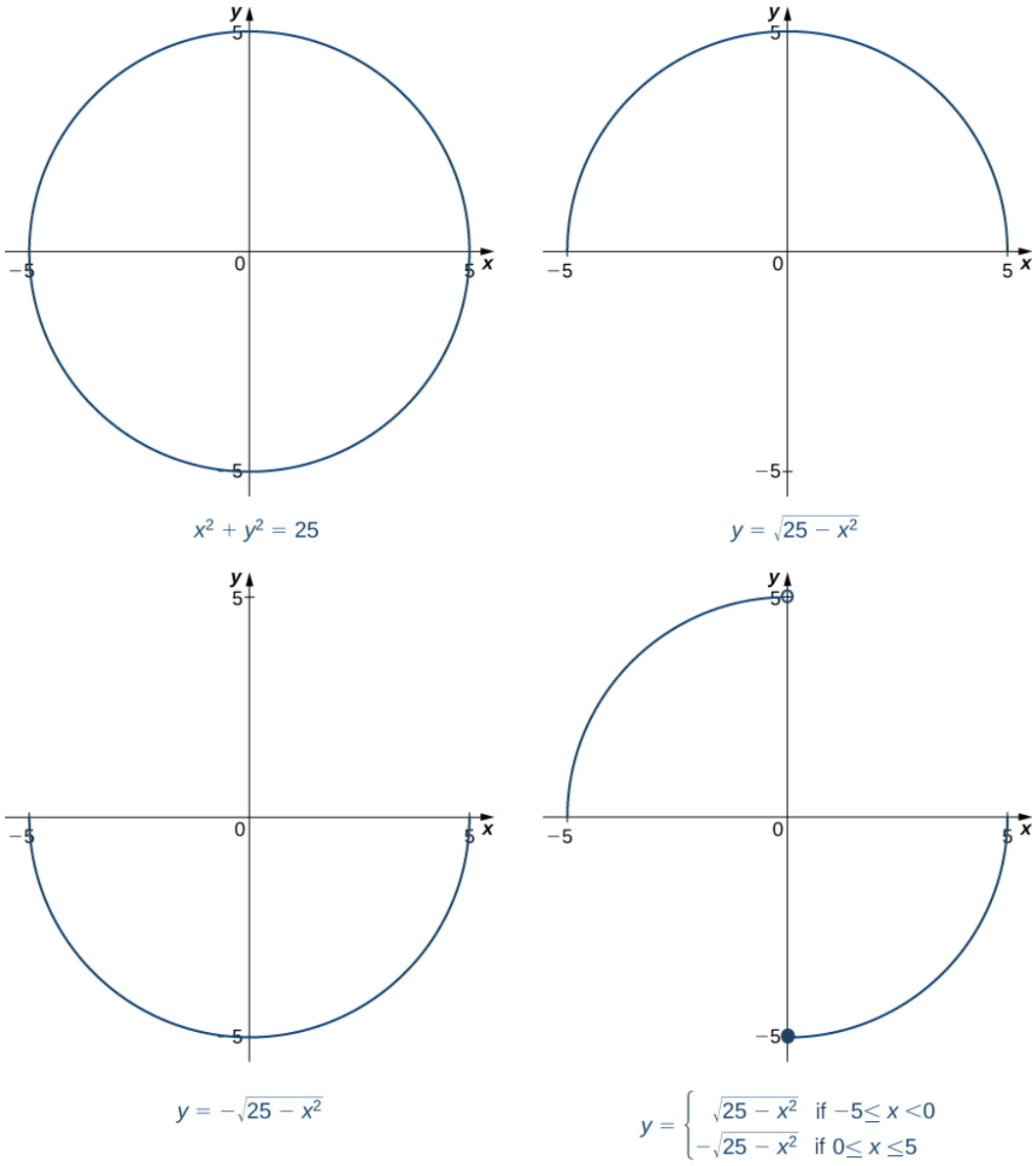


Figure 3.7.1:The equation $x^2 + y^2 = 25$ defines many functions implicitly.

If we want to find the slope of the line tangent to the graph of $x^2 + y^2 = 25$ at the point $(3, 4)$, we could evaluate the derivative of the function $y = \sqrt{25 - x^2}$ at $x = 3$. On the other hand, if we want the slope of the tangent line at the point $(3, -4)$, we could use the derivative of $y = -\sqrt{25 - x^2}$. However, it is not always easy to solve for a function defined implicitly by an equation. Fortunately, the technique of **implicit differentiation** allows us to find the derivative of an implicitly defined function without ever solving for the function explicitly. The process of finding $\dfrac{dy}{dx}$ using implicit differentiation is described in the following problem-solving strategy.

Problem-Solving Strategy: Implicit Differentiation

To perform *implicit differentiation* on an equation that defines a function y implicitly in terms of a variable x, use the following steps:

1. Take the derivative of both sides of the equation. Keep in mind that y is a function of x. Consequently, whereas

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

and

$$\frac{d}{dx}(\sin y) = \cos y \cdot \frac{dy}{dx}$$

because we must use the chain rule to differentiate $\sin y$ with respect to x.

2. Rewrite the equation so that all terms containing dy/dx are on the left and all terms that do not contain dy/dx are on the right.
3. Factor out dy/dx on the left.
4. Solve for dy/dx by dividing both sides of the equation by an appropriate algebraic expression.

Example 3.7.1: Using Implicit Differentiation

Assuming that y is defined implicitly by the equation $x^2 + y^2 = 25$, find $\frac{dy}{dx}$.

Solution

Follow the steps in the problem-solving strategy.

$\frac{d}{dx}(x^2 + y^2) = \frac{d}{dx}(25)$	Step 1. Differentiate both sides of the equation.
$\frac{d}{dx}(x^2) + \frac{d}{dx}(y^2) = 0$	Step 1.1. Use the sum rule on the left. On the right $\frac{d}{dx}(25) = 0$.
$2x + 2y\frac{dy}{dx} = 0$	Step 1.2. Take the derivatives, so $\frac{d}{dx}(x^2) = 2x$ and $\frac{d}{dx}(y^2) = 2y\frac{dy}{dx}$.
$2y\frac{dy}{dx} = -2x$	Step 2. Keep the terms with $\frac{dy}{dx}$ on the left. Move the remaining terms to the right.
$\frac{dy}{dx} = -\frac{x}{y}$	Step 4. Divide both sides of the equation by $2y$. (Step 3 does not apply in this case.)

Analysis

Note that the resulting expression for $\frac{dy}{dx}$ is in terms of both the independent variable x and the dependent variable y. Although in some cases it may be possible to express $\frac{dy}{dx}$ in terms of x only, it is generally not possible to do so.

Example 3.7.2: Using Implicit Differentiation and the Product Rule

Assuming that y is defined implicitly by the equation $x^3 \sin y + y = 4x + 3$, find $\frac{dy}{dx}$.

Solution

$\frac{d}{dx}(x^3 \sin y + y) = \frac{d}{dx}(4x + 3)$	Step 1: Differentiate both sides of the equation.
$\frac{d}{dx}(x^3 \sin y) + \frac{d}{dx}(y) = 4$	Step 1.1: Apply the sum rule on the left. On the right, $\frac{d}{dx}(4x + 3) = 4$.
$\left(\frac{d}{dx}(x^3) \cdot \sin y + \frac{d}{dx}(\sin y) \cdot x^3\right) + \frac{dy}{dx} = 4$	Step 1.2: Use the product rule to find $\frac{d}{dx}(x^3 \sin y)$. Observe that $\frac{d}{dx}(y) = \frac{dy}{dx}$.
$3x^2 \sin y + (\cos y\frac{dy}{dx}) \cdot x^3 + \frac{dy}{dx} = 4$	Step 1.3: We know $\frac{d}{dx}(x^3) = 3x^2$. Use the chain rule to obtain $\frac{d}{dx}(\sin y) = \cos y\frac{dy}{dx}$.

$x^3 \cos y \frac{dy}{dx} + \frac{dy}{dx} = 4 - 3x^2 \sin y$	Step 2: Keep all terms containing $\frac{dy}{dx}$ on the left. Move all other terms to the right.
$\frac{dy}{dx}(x^3 \cos y + 1) = 4 - 3x^2 \sin y$	Step 3: Factor out $\frac{dy}{dx}$ on the left.
$\frac{dy}{dx} = \frac{4 - 3x^2 \sin y}{x^3 \cos y + 1}$	Step 4: Solve for $\frac{dy}{dx}$ by dividing both sides of the equation by $x^3 \cos y + 1$.

Example 3.7.3: Using Implicit Differentiation to Find a Second Derivative

Find $\frac{d^2y}{dx^2}$ if $x^2 + y^2 = 25$.

Solution

In Example 3.7.1, we showed that $\frac{dy}{dx} = -\frac{x}{y}$. We can take the derivative of both sides of this equation to find $\frac{d^2y}{dx^2}$.

$$\frac{d^2y}{dx^2} = \frac{d}{dy}\left(-\frac{x}{y}\right) \qquad \text{Differentiate both sides of } \frac{dy}{dx} = -\frac{x}{y}.$$

$$= -\frac{\left(1 \cdot y - x\frac{dy}{dx}\right)}{y^2} \qquad \text{Use the quotient rule to find } \frac{d}{dy}\left(-\frac{x}{y}\right).$$

$$= \frac{-y + x\frac{dy}{dx}}{y^2} \qquad \text{Simplify.}$$

$$= \frac{-y + x\left(-\frac{x}{y}\right)}{y^2} \qquad \text{Substitute } \frac{dy}{dx} = -\frac{x}{y}.$$

$$= \frac{-y^2 - x^2}{y^3} \qquad \text{Simplify.}$$

At this point we have found an expression for $\frac{d^2y}{dx^2}$. If we choose, we can simplify the expression further by recalling that $x^2 + y^2 = 25$ and making this substitution in the numerator to obtain $\frac{d^2y}{dx^2} = -\frac{25}{y^3}$.

Exercise 3.7.1

Find $\frac{dy}{dx}$ for y defined implicitly by the equation $4x^5 + \tan y = y^2 + 5x$.

Hint

Follow the problem solving strategy, remembering to apply the chain rule to differentiate $\tan y$ and y^2.

Answer

$$\frac{dy}{dx} = \frac{5 - 20x^4}{\sec^2 y - 2y}$$

Finding Tangent Lines Implicitly

Now that we have seen the technique of implicit differentiation, we can apply it to the problem of finding equations of tangent lines to curves described by equations.

Example 3.7.4: Finding a Tangent Line to a Circle

Find the equation of the line tangent to the curve $x^2+y^2=25$ at the point $(3,-4)$.

Solution

Although we could find this equation without using implicit differentiation, using that method makes it much easier. In Example 3.7.1, we found $\frac{dy}{dx}=-\frac{x}{y}$.

The slope of the tangent line is found by substituting $(3,-4)$ into this expression. Consequently, the slope of the tangent line is $\frac{dy}{dx}\Big|_{(3,-4)}=-\frac{3}{-4}=\frac{3}{4}$.

Using the point $(3,-4)$ and the slope $\frac{3}{4}$ in the point-slope equation of the line, we obtain the equation $y=\frac{3}{4}x-\frac{25}{4}$ (Figure 3.7.2).

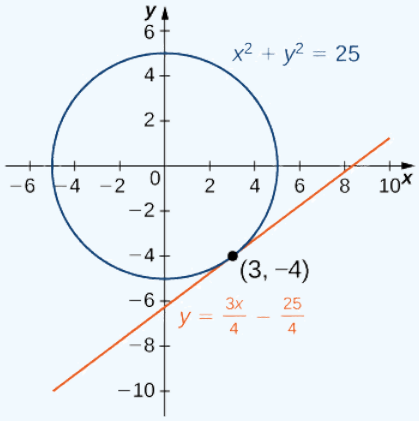


Figure 3.7.2: The line $y=\frac{3}{4}x-\frac{25}{4}$ is tangent to $x^2+y^2=25$ at the point (3, −4).

Example 3.7.5: Finding the Equation of the Tangent Line to a Curve

Find the equation of the line tangent to the graph of $y^3+x^3-3xy=0$ at the point $\left(\frac{3}{2},\frac{3}{2}\right)$ (Figure 3.7.3). This curve is known as the folium (or leaf) of Descartes.

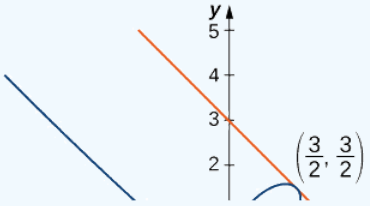


Figure 3.7.3: Finding the tangent line to the folium of Descartes at $\left(\frac{3}{2}, \frac{3}{2}\right)$.

Solution

Begin by finding $\frac{dy}{dx}$.

$$\frac{d}{dx}\left(y^3 + x^3 - 3xy\right) = \frac{d}{dx}(0)$$
$$3y^2\frac{dy}{dx} + 3x^2 - \left(3y + 3x\frac{dy}{dx}\right) = 0$$
$$3y^2\frac{dy}{dx} + 3x^2 - 3y - 3x\frac{dy}{dx} = 0$$
$$\left(3y^2 - 3x\right)\frac{dy}{dx} = 3y - 3x^2$$
$$\frac{dy}{dx} = \frac{3y - 3x^2}{3y^2 - 3x}.$$

Next, substitute $\left(\frac{3}{2}, \frac{3}{2}\right)$ into $\frac{dy}{dx} = \frac{3y - 3x^2}{3y^2 - 3x}$ to find the slope of the tangent line:

$$\left.\frac{dy}{dx}\right|_{\left(\frac{3}{2}, \frac{3}{2}\right)} = -1.$$

Finally, substitute into the point-slope equation of the line to obtain

$$y = -x + 3.$$

Example 3.7.6: Applying Implicit Differentiation

In a simple video game, a rocket travels in an elliptical orbit whose path is described by the equation $4x^2 + 25y^2 = 100$. The rocket can fire missiles along lines tangent to its path. The object of the game is to destroy an incoming asteroid traveling along the positive x-axis toward $(0, 0)$. If the rocket fires a missile when it is located at $\left(3, \frac{8}{5}\right)$, where will it intersect the x-axis?

Solution

To solve this problem, we must determine where the line tangent to the graph of

$4x^2 + 25y^2 = 100$ at $\left(3, \frac{8}{5}\right)$ intersects the x-axis. Begin by finding $\frac{dy}{dx}$ implicitly.

Differentiating, we have

$$8x+50y\frac{dy}{dx}=0.$$

Solving for $\frac{dy}{dx}$,

we have

$$\frac{dy}{dx}=-\frac{4x}{25y}.$$

The slope of the tangent line is $\left.\frac{dy}{dx}\right|_{\left(3,\frac{8}{5}\right)}=-\frac{3}{10}$. The equation of the tangent line is $y=-\frac{3}{10}x+\frac{5}{2}$. To determine where the line intersects the x-axis, solve $0=-\frac{3}{10}x+\frac{5}{2}$. The solution is $x=\frac{25}{3}$. The missile intersects the x-axis at the point $\left(\frac{25}{3},0\right)$.

? Exercise 3.7.2

Find the equation of the line tangent to the hyperbola $x^2-y^2=16$ at the point $(5,3)$.

Hint

$\frac{dy}{dx}=\frac{x}{y}$

Answer

$y=\frac{5}{3}x-\frac{16}{3}$

Key Concepts

- We use implicit differentiation to find derivatives of implicitly defined functions (functions defined by equations).
- By using implicit differentiation, we can find the equation of a tangent line to the graph of a curve.

Glossary

implicit differentiation

is a technique for computing $\frac{dy}{dx}$ for a function defined by an equation, accomplished by differentiating both sides of the equation (remembering to treat the variable y as a function) and solving for $\frac{dy}{dx}$

This page titled 3.7: Implicit Differentiation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **3.8: Implicit Differentiation** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

3.8: Derivatives of Inverse and Logarithmic Functions

Learning Objectives

- Calculate the derivative of an inverse function.
- Find the derivative of exponential functions.
- Find the derivative of logarithmic functions.
- Use logarithmic differentiation to determine the derivative of a function.

So far, we have learned how to differentiate a variety of functions, including trigonometric, inverse, and implicit functions. In this section, we explore derivatives of exponential and logarithmic functions and also inverses. As we discussed in Introduction to Functions and Graphs, exponential functions play an important role in modeling population growth and the decay of radioactive materials. Logarithmic functions can help rescale large quantities and are particularly helpful for rewriting complicated expressions.

The Derivative of an Inverse Function

We begin by considering a function and its inverse. If $f(x)$ is both invertible and differentiable, it seems reasonable that the inverse of $f(x)$ is also differentiable. Figure 3.8.1 shows the relationship between a function $f(x)$ and its inverse $f^{-1}(x)$. Look at the point $\big(a,\ f^{-1}(a)\big)$ on the graph of $f^{-1}(x)$ having a tangent line with a slope of

$$\big(f^{-1}\big)'(a) = \frac{p}{q}.$$

This point corresponds to a point $\big(f^{-1}(a),\ a\big)$ on the graph of $f(x)$ having a tangent line with a slope of

$$f'\big(f^{-1}(a)\big) = \frac{q}{p}.$$

Thus, if $f^{-1}(x)$ is differentiable at a, then it must be the case that

$$\big(f^{-1}\big)'(a) = \frac{1}{f'\big(f^{-1}(a)\big)}.$$

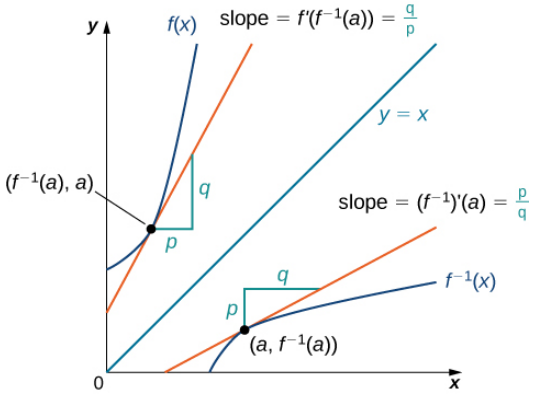


Figure 3.8.1: The tangent lines of a function and its inverse are related; so, too, are the derivatives of these functions.

We may also derive the formula for the derivative of the inverse by first recalling that $x = f\big(f^{-1}(x)\big)$. Then by differentiating both sides of this equation (using the chain rule on the right), we obtain

$$1 = f'\big(f^{-1}(x)\big)\big(f^{-1}\big)'(x).$$

Solving for $\big(f^{-1}\big)'(x)$, we obtain

$$\big(f^{-1}\big)'(x) = \frac{1}{f'\big(f^{-1}(x)\big)}.$$

We summarize this result in the following theorem.

Inverse Function Theorem

Let $f(x)$ be a function that is both invertible and differentiable. Let $y = f^{-1}(x)$ be the inverse of $f(x)$. For all x satisfying $f'\big(f^{-1}(x)\big) \neq 0$,

$$\frac{dy}{dx} = \frac{d}{dx}\left(f^{-1}(x)\right) = \left(f^{-1}\right)'(x) = \frac{1}{f'\left(f^{-1}(x)\right)}. \tag{3.8.1}$$

Alternatively, if $y = g(x)$ is the inverse of $f(x)$, then

$$g'(x) = \frac{1}{f'\left(g(x)\right)}. \tag{3.8.2}$$

Example 3.8.1: Applying the Inverse Function Theorem

Use the inverse function theorem to find the derivative of $g(x) = \dfrac{x+2}{x}$. Compare the resulting derivative to that obtained by differentiating the function directly.

Solution

The inverse of $g(x) = \dfrac{x+2}{x}$ is $f(x) = \dfrac{2}{x-1}$.

We will use Equation 3.8.2 and begin by finding $f'(x)$. Thus,

$$f'(x) = \frac{-2}{(x-1)^2}$$

and

$$f'\left(g(x)\right) = \frac{-2}{(g(x)-1)^2} = \frac{-2}{\left(\dfrac{x+2}{x} - 1\right)^2} = -\frac{x^2}{2}.$$

Finally,

$$g'(x) = \frac{1}{f'\left(g(x)\right)} = -\frac{2}{x^2}.$$

We can verify that this is the correct derivative by applying the quotient rule to $g(x)$ to obtain

$$g'(x) = -\frac{2}{x^2}.$$

Exercise 3.8.1

Use the inverse function theorem to find the derivative of $g(x) = \dfrac{1}{x+2}$. Compare the result obtained by differentiating $g(x)$ directly.

Hint

Use the preceding example as a guide.

Answer

$$g'(x) = -\frac{1}{(x+2)^2}$$

Example 3.8.2: Applying the Inverse Function Theorem

Use the inverse function theorem to find the derivative of $g(x) = \sqrt[3]{x}$.

Solution

The function $g(x) = \sqrt[3]{x}$ is the inverse of the function $f(x) = x^3$. Since $g'(x) = \dfrac{1}{f'\left(g(x)\right)}$, begin by finding $f'(x)$. Thus,

$$f'(x) = 3x^2$$

and

$$f'\left(g(x)\right) = 3\left(\sqrt[3]{x}\right)^2 = 3x^{2/3}$$

Finally,

$$g'(x) = \frac{1}{3x^{2/3}}.$$

If we were to differentiate $g(x)$ directly, using the power rule, we would first rewrite $g(x) = \sqrt[3]{x}$ as a power of x to get,

$$g(x) = x^{1/3}$$

Then we would differentiate using the power rule to obtain

$$g'(x) = \tfrac{1}{3}x^{-2/3} = \frac{1}{3x^{2/3}}.$$

? Exercise 3.8.2

Find the derivative of $g(x) = \sqrt[5]{x}$ by applying the inverse function theorem.

Hint

$g(x)$ is the inverse of $f(x) = x^5$.

Answer

$g(x) = \frac{1}{5}x^{-4/5}$

Derivative of the Exponential Function

Just as when we found the derivatives of other functions, we can find the derivatives of exponential and logarithmic functions using formulas. As we develop these formulas, we need to make certain basic assumptions. The proofs that these assumptions hold are beyond the scope of this course.

First of all, we begin with the assumption that the function $B(x) = b^x$, $b > 0$, is defined for every real number and is continuous. In previous courses, the values of exponential functions for all rational numbers were defined—beginning with the definition of b^n, where n is a positive integer—as the product of b multiplied by itself n times. Later, we defined $b^0 = 1, b^{-n} = \frac{1}{b^n}$, for a positive integer n, and $b^{s/t} = (\sqrt[t]{b})^s$ for positive integers s and t. These definitions leave open the question of the value of b^r where r is an arbitrary real number. By assuming the continuity of $B(x) = b^x, b > 0$, we may interpret b^r as $\lim\limits_{x \to r} b^x$ where the values of x as we take the limit are rational. For example, we may view 4^π as the number satisfying

$$4^3 < 4^\pi < 4^4, \quad 4^{3.1} < 4^\pi < 4^{3.2}, \quad 4^{3.14} < 4^\pi < 4^{3.15},$$
$$4^{3.141} < 4^\pi < 4^{3.142}, \quad 4^{3.1415} < 4^\pi < 4^{3.1416}, \quad \ldots.$$

As we see in the following table, $4^\pi \approx 77.88$.

x	4^x	x	4^x
4^3	64	$4^{3.141593}$	77.8802710486
$4^{3.1}$	73.5166947198	$4^{3.1416}$	77.8810268071
$4^{3.14}$	77.7084726013	$4^{3.142}$	77.9242251944
$4^{3.141}$	77.8162741237	$4^{3.15}$	78.7932424541
$4^{3.1415}$	77.8702309526	$4^{3.2}$	84.4485062895
$4^{3.14159}$	77.8799471543	4^4	256

Approximating a Value of 4^π

We also assume that for $B(x) = b^x$, $b > 0$, the value $B'(0)$ of the derivative exists. In this section, we show that by making this one additional assumption, it is possible to prove that the function $B(x)$ is differentiable everywhere.

We make one final assumption: that there is a unique value of $b > 0$ for which $B'(0) = 1$. We define e to be this unique value, as we did in Introduction to Functions and Graphs. Figure 3.8.2 provides graphs of the functions $y = 2^x$, $y = 3^x$, $y = 2.7^x$, and $y = 2.8^x$. A visual estimate of the slopes of the tangent lines to these functions at 0 provides evidence that the value of e lies somewhere between 2.7 and 2.8. The function $E(x) = e^x$ is called the **natural exponential function**. Its inverse, $L(x) = \log_e x = \ln x$ is called the **natural logarithmic function.**

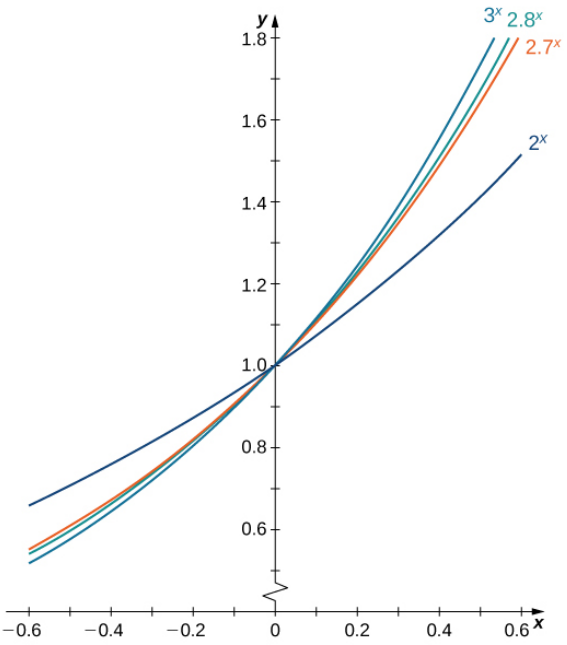


Figure 3.8.2: The graph of $E(x) = e^x$ is between $y = 2^x$ and $y = 3^x$.

For a better estimate of e, we may construct a table of estimates of $B'(0)$ for functions of the form $B(x) = b^x$. Before doing this, recall that

$$B'(0) = \lim_{x \to 0} \frac{b^x - b^0}{x - 0} = \lim_{x \to 0} \frac{b^x - 1}{x} \approx \frac{b^x - 1}{x}$$

for values of x very close to zero. For our estimates, we choose $x = 0.00001$ and $x = -0.00001$

to obtain the estimate

$$\frac{b^{-0.00001} - 1}{-0.00001} < B'(0) < \frac{b^{0.00001} - 1}{0.00001}.$$

See the following table.

Table : Estimating the value of e

b	$\frac{b^{-0.00001}-1}{-0.00001} < B'(0) < \frac{b^{0.00001}-1}{0.00001}$.	b	$\frac{b^{-0.00001}-1}{-0.00001} < B'(0) < \frac{b^{0.00001}-1}{0.00001}$.
2	$0.693145 < B'(0) < 0.69315$	2.7183	$1.000002 < B'(0) < 1.000012$
2.7	$0.993247 < B'(0) < 0.993257$	2.719	$1.000259 < B'(0) < 1.000269$
2.71	$0.996944 < B'(0) < 0.996954$	2.72	$1.000627 < B'(0) < 1.000637$
2.718	$0.999891 < B'(0) < 0.999901$	2.8	$1.029614 < B'(0) < 1.029625$
2.7182	$0.999965 < B'(0) < 0.999975$	3	$1.098606 < B'(0) < 1.098618$

The evidence from the table suggests that $2.7182 < e < 2.7183$.

The graph of $E(x) = e^x$ together with the line $y = x + 1$ are shown in Figure 3.8.2. This line is tangent to the graph of $E(x) = e^x$ at $x = 0$.

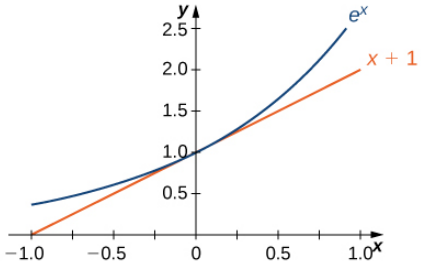


Figure 3.8.2: The tangent line to $E(x) = e^x$ at $x = 0$ has slope 1.

Now that we have laid out our basic assumptions, we begin our investigation by exploring the derivative of $B(x) = b^x$, $b > 0$. Recall that we have assumed that $B'(0)$ exists. By applying the limit definition to the derivative we conclude that

$$B'(0) = \lim_{h \to 0} \frac{b^{0+h} - b^0}{h} = \lim_{h \to 0} \frac{b^h - 1}{h}$$

Turning to $B'(x)$, we obtain the following.

$$\begin{aligned}
B'(x) &= \lim_{h \to 0} \frac{b^{x+h} - b^x}{h} && \text{Apply the limit definition of the derivative.} \\
&= \lim_{h \to 0} \frac{b^x b^h - b^x}{h} && \text{Note that } b^{x+h} = b^x b^h. \\
&= \lim_{h \to 0} \frac{b^x (b^h - 1)}{h} && \text{Factor out } b^x. \\
&= b^x \lim_{h \to 0} \frac{b^h - 1}{h} && \text{Apply a property of limits.} \\
&= b^x B'(0) && \text{Use } B'(0) = \lim_{h \to 0} \frac{b^{0+h} - b^0}{h} = \lim_{h \to 0} \frac{b^h - 1}{h}.
\end{aligned}$$

We see that on the basis of the assumption that $B(x) = b^x$ is differentiable at 0, $B(x)$ is not only differentiable everywhere, but its derivative is

$$B'(x) = b^x B'(0).$$

For $E(x) = e^x$, $E'(0) = 1$. Thus, we have $E'(x) = e^x$. (The value of $B'(0)$ for an arbitrary function of the form $B(x) = b^x$, $b > 0$, will be derived later.)

Derivative of the Natural Exponential Function

Let $E(x) = e^x$ be the natural exponential function. Then

$$E'(x) = e^x.$$

In general,

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

Example 3.8.3: Derivative of an Exponential Function

Find the derivative of $f(x) = e^{\tan(2x)}$.

Solution:

Using the derivative formula and the chain rule,

$$f'(x) = e^{\tan(2x)} \frac{d}{dx}\left(\tan(2x)\right) = e^{\tan(2x)} \sec^2(2x) \cdot 2$$

Example 3.8.4: Combining Differentiation Rules

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

Solution

Use the derivative of the natural exponential function, the quotient rule, and the chain rule.

$$\begin{aligned}
y' &= \frac{(e^{x^2} \cdot 2)x \cdot x - 1 \cdot e^{x^2}}{x^2} && \text{Apply the quotient rule.} \\
&= \frac{e^{x^2}(2x^2 - 1)}{x^2} && \text{Simplify.}
\end{aligned}$$

Exercise 3.8.4

Find the derivative of $h(x) = xe^{2x}$.

Hint

Don't forget to use the product rule.

Answer

$h'(x) = e^{2x} + 2xe^{2x}$

Example 3.8.5: Applying the Natural Exponential Function

A colony of mosquitoes has an initial population of 1000. After t days, the population is given by $A(t) = 1000e^{0.3t}$. Show that the ratio of the rate of change of the population, $A'(t)$, to the population, $A(t)$ is constant.

Solution

First find $A'(t)$. By using the chain rule, we have $A'(t) = 300e^{0.3t}$. Thus, the ratio of the rate of change of the population to the population is given by

$$\frac{A'(t)}{A(t)} = \frac{300e^{0.3t}}{1000e^{0.3t}} = 0.3.$$

The ratio of the rate of change of the population to the population is the constant 0.3.

Exercise 3.8.5

If $A(t) = 1000e^{0.3t}$ describes the mosquito population after t days, as in the preceding example, what is the rate of change of $A(t)$ after 4 days?

Hint

Find $A'(4)$.

Answer

996

Derivative of the Logarithmic Function

Now that we have the derivative of the natural exponential function, we can use implicit differentiation to find the derivative of its inverse, the natural logarithmic function.

Definition: The Derivative of the Natural Logarithmic Function

If $x > 0$ and $y = \ln x$, then

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

More generally, let $g(x)$ be a differentiable function. For all values of x for which $g'(x) > 0$, the derivative of $h(x) = \ln(g(x))$ is given by

$$h'(x) = \frac{1}{g(x)}g'(x).$$

Proof

If $x > 0$ and $y = \ln x$, then $e^y = x$. Differentiating both sides of this equation results in the equation

$$e^y \frac{dy}{dx} = 1.$$

Solving for $\frac{dy}{dx}$ yields

$$\frac{dy}{dx} = \frac{1}{e^y}.$$

Finally, we substitute $x = e^y$ to obtain

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

We may also derive this result by applying the inverse function theorem, as follows. Since $y = g(x) = \ln x$ is the inverse of $f(x) = e^x$, by applying the inverse function theorem we have

$$\frac{dy}{dx} = \frac{1}{f'(g(x))} = \frac{1}{e^{\ln x}} = \frac{1}{x}.$$

Using this result and applying the chain rule to $h(x) = \ln(g(x))$ yields

$$h'(x) = \frac{1}{g(x)} g'(x). \tag{3.8.3}$$

□

The graph of $y = \ln x$ and its derivative $\frac{dy}{dx} = \frac{1}{x}$ are shown in Figure 3.8.3.

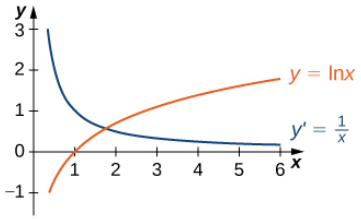


Figure 3.8.3: The function $y = \ln x$ is increasing on $(0, +\infty)$. Its derivative $y' = \frac{1}{x}$ is greater than zero on $(0, +\infty)$

Example 3.8.6: Taking a Derivative of a Natural Logarithm

Find the derivative of $f(x) = \ln(x^3 + 3x - 4)$.

Solution

Use Equation 3.8.3 directly.

$$f'(x) = \frac{1}{x^3 + 3x - 4} \cdot (3x^2 + 3) \qquad \text{Use } g(x) = x^3 + 3x - 4 \text{ in } h'(x) = \frac{1}{g(x)} g'(x).$$

$$= \frac{3x^2 + 3}{x^3 + 3x - 4} \qquad \text{Rewrite.}$$

Example 3.8.7: Using Properties of Logarithms in a Derivative

Find the derivative of $f(x) = \ln\left(\frac{x^2 \sin x}{2x + 1}\right)$.

Solution

At first glance, taking this derivative appears rather complicated. However, by using the properties of logarithms prior to finding the derivative, we can make the problem much simpler.

$$f(x) = \ln\left(\frac{x^2 \sin x}{2x + 1}\right) = 2\ln x + \ln(\sin x) - \ln(2x + 1) \qquad \text{Apply properties of logarithms.}$$

$$f'(x) = \frac{2}{x} + \cot x - \frac{2}{2x + 1} \qquad \text{Apply sum rule and } h'(x) = \frac{1}{g(x)} g'(x).$$

Exercise 3.8.7

Differentiate: $f(x) = \ln(3x + 2)^5$.

Hint

Use a property of logarithms to simplify before taking the derivative.

Answer

$$f'(x) = \frac{15}{3x+2}$$

Now that we can differentiate the natural logarithmic function, we can use this result to find the derivatives of $y = \log_b x$ and $y = b^x$ for $b > 0$, $b \neq 1$.

Derivatives of General Exponential and Logarithmic Functions

Let $b > 0, b \neq 1,$ and let $g(x)$ be a differentiable function.

i. If $y = \log_b x$, then

$$\frac{dy}{dx} = \frac{1}{x \ln b}.$$

More generally, if $h(x) = \log_b(g(x))$, then for all values of x for which $g(x) > 0$,

$$h'(x) = \frac{g'(x)}{g(x) \ln b}. \tag{3.8.4}$$

ii. If $y = b^x$, then

$$\frac{dy}{dx} = b^x \ln b.$$

More generally, if $h(x) = b^{g(x)}$, then

$$h'(x) = b^{g(x)} g'(x) \ln b \tag{3.8.5}$$

Proof

If $y = \log_b x,$ then $b^y = x$. It follows that $\ln(b^y) = \ln x$. Thus $y \ln b = \ln x$. Solving for y, we have $y = \frac{\ln x}{\ln b}$. Differentiating and keeping in mind that $\ln b$ is a constant, we see that

$$\frac{dy}{dx} = \frac{1}{x \ln b}.$$

The derivative in Equation 3.8.4 now follows from the chain rule.

If $y = b^x$. then $\ln y = x \ln b$. Using implicit differentiation, again keeping in mind that $\ln b$ is constant, it follows that $\frac{1}{y}\frac{dy}{dx} = \ln b$. Solving for $\frac{dy}{dx}$ and substituting $y = b^x$, we see that

$$\frac{dy}{dx} = y \ln b = b^x \ln b.$$

The more general derivative (Equation 3.8.5) follows from the chain rule.

□

Example 3.8.8: Applying Derivative Formulas

Find the derivative of $h(x) = \frac{3^x}{3^x + 2}$.

Solution

Use the quotient rule and Note.

$$h'(x) = \frac{3^x \ln 3(3^x + 2) - 3^x \ln 3(3^x)}{(3^x + 2)^2} \qquad \text{Apply the quotient rule.}$$

$$= \frac{2 \cdot 3^x \ln 3}{(3x + 2)^2} \qquad \text{Simplify.}$$

Example 3.8.9: Finding the Slope of a Tangent Line

Find the slope of the line tangent to the graph of $y = \log_2(3x+1)$ at $x = 1$.

Solution

To find the slope, we must evaluate $\dfrac{dy}{dx}$ at $x = 1$. Using Equation 3.8.4, we see that

$$\frac{dy}{dx} = \frac{3}{(3x+1)\ln 2}.$$

By evaluating the derivative at $x = 1$, we see that the tangent line has slope

$$\left.\frac{dy}{dx}\right|_{x=1} = \frac{3}{4\ln 2} = \frac{3}{\ln 16}.$$

Exercise 3.8.9

Find the slope for the line tangent to $y = 3^x$ at $x = 2$.

Hint

Evaluate the derivative at $x = 2$.

Answer

$9\ln(3)$

Logarithmic Differentiation

At this point, we can take derivatives of functions of the form $y = (g(x))^n$ for certain values of n, as well as functions of the form $y = b^{g(x)}$, where $b > 0$ and $b \neq 1$. Unfortunately, we still do not know the derivatives of functions such as $y = x^x$ or $y = x^\pi$. These functions require a technique called **logarithmic differentiation**, which allows us to differentiate any function of the form $h(x) = g(x)^{f(x)}$. It can also be used to convert a very complex differentiation problem into a simpler one, such as finding the derivative of $y = \dfrac{x\sqrt{2x+1}}{e^x \sin^3 x}$. We outline this technique in the following problem-solving strategy.

Problem-Solving Strategy: Using Logarithmic Differentiation

1. To differentiate $y = h(x)$ using logarithmic differentiation, take the natural logarithm of both sides of the equation to obtain $\ln y = \ln(h(x))$.
2. Use properties of logarithms to expand $\ln(h(x))$ as much as possible.
3. Differentiate both sides of the equation. On the left we will have $\dfrac{1}{y}\dfrac{dy}{dx}$.
4. Multiply both sides of the equation by y to solve for $\dfrac{dy}{dx}$.
5. Replace y by $h(x)$.

Example 3.8.10: Using Logarithmic Differentiation

Find the derivative of $y = (2x^4+1)^{\tan x}$.

Solution

Use logarithmic differentiation to find this derivative.

$\ln y = \ln(2x^4+1)^{\tan x}$	Step 1. Take the natural logarithm of both sides.
$\ln y = \tan x \ln(2x^4+1)$	Step 2. Expand using properties of logarithms.
$\frac{1}{y}\frac{dy}{dx} = \sec^2 x \ln(2x^4+1) + \frac{8x^3}{2x^4+1} \cdot \tan x$	Step 3. Differentiate both sides. Use the product rule on the right.
$\frac{dy}{dx} = y \cdot (\sec^2 x \ln(2x^4+1) + \frac{8x^3}{2x^4+1} \cdot \tan x)$	Step 4. Multiply by y on both sides.
$\frac{dy}{dx} = (2x^4+1)^{\tan x}(\sec^2 x \ln(2x^4+1) + \frac{8x^3}{2x^4+1} \cdot \tan x)$	Step 5. Substitute $y = (2x^4+1)^{\tan x}$.

Example 3.8.11: Extending the Power Rule

Find the derivative of $y = \dfrac{x\sqrt{2x+1}}{e^x \sin^3 x}$.

Solution

This problem really makes use of the properties of logarithms and the differentiation rules given in this chapter.

$\ln y = \ln \dfrac{x\sqrt{2x+1}}{e^x \sin^3 x}$	Step 1. Take the natural logarithm of both sides.
$\ln y = \ln x + \frac{1}{2}\ln(2x+1) - x \ln e - 3 \ln \sin x$	Step 2. Expand using properties of logarithms.
$\frac{1}{y}\frac{dy}{dx} = \frac{1}{x} + \frac{1}{2x+1} - 1 - 3\frac{\cos x}{\sin x}$	Step 3. Differentiate both sides.
$\frac{dy}{dx} = y\left(\frac{1}{x} + \frac{1}{2x+1} - 1 - 3\cot x\right)$	Step 4. Multiply by y on both sides.
$\frac{dy}{dx} = \frac{x\sqrt{2x+1}}{e^x \sin^3 x}\left(\frac{1}{x} + \frac{1}{2x+1} - 1 - 3\cot x\right)$	Step 5. Substitute $y = \dfrac{x\sqrt{2x+1}}{e^x \sin^3 x}$.

Exercise 3.8.11

Use logarithmic differentiation to find the derivative of $y = x^x$.

Hint

Follow the problem solving strategy.

Answer

Solution: $\dfrac{dy}{dx} = x^x(1+\ln x)$

Exercise 3.8.12

Find the derivative of $y = (\tan x)^\pi$.

Hint

Use the power rule (since the exponent π is a constant) and the chain rule.

Answer

$y' = \pi(\tan x)^{\pi-1} \sec^2 x$

Key Concepts

- On the basis of the assumption that the exponential function $y = b^x$, $b > 0$ is continuous everywhere and differentiable at 0, this function is differentiable everywhere and there is a formula for its derivative.
- We can use a formula to find the derivative of $y = \ln x$, and the relationship $\log_b x = \dfrac{\ln x}{\ln b}$ allows us to extend our differentiation formulas to include logarithms with arbitrary bases.

- Logarithmic differentiation allows us to differentiate functions of the form $y = g(x)^{f(x)}$ or very complex functions by taking the natural logarithm of both sides and exploiting the properties of logarithms before differentiating.

Key Equations

- **Derivative of the natural exponential function**

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

- **Derivative of the natural logarithmic function**

$$\frac{d}{dx}\left(\ln g(x)\right) = \frac{1}{g(x)}g'(x)$$

- **Derivative of the general exponential function**

$$\frac{d}{dx}\left(b^{g(x)}\right) = b^{g(x)}g'(x)\ln b$$

- **Derivative of the general logarithmic function**

$$\frac{d}{dx}\left(\log_b g(x)\right) = \frac{g'(x)}{g(x)\ln b}$$

Glossary

logarithmic differentiation

is a technique that allows us to differentiate a function by first taking the natural logarithm of both sides of an equation, applying properties of logarithms to simplify the equation, and differentiating implicitly

---

This page titled 3.8: Derivatives of Inverse and Logarithmic Functions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **3.9: Derivatives of Exponential and Logarithmic Functions** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **3.7: Derivatives of Inverse Functions** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

3.9: Inverse Trigonometric Functions

Learning Objectives

- Recognize the derivatives of the standard inverse trigonometric functions.

Derivatives of Inverse Trigonometric Functions

In this section we will find derivatives of inverse trigonometric functions. These derivatives will prove invaluable in the study of integration later in this text. The derivatives of inverse trigonometric functions are quite surprising in that their derivatives are actually algebraic functions. Previously, derivatives of algebraic functions have proven to be algebraic functions and derivatives of trigonometric functions have been shown to be trigonometric functions. Here, for the first time, we see that the derivative of a function need not be of the same type as the original function.

Example 3.9.1*A*: Derivative of the Inverse Sine Function

Use the inverse function theorem to find the derivative of $g(x) = \sin^{-1} x$.

Solution

Since for x in the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, $f(x) = \sin x$ is the inverse of $g(x) = \sin^{-1} x$, begin by finding $f'(x)$. Since

$$f'(x) = \cos x$$

and

$$f'(g(x)) = \cos\left(\sin^{-1} x\right) = \sqrt{1 - x^2}$$

we see that

$$g'(x) = \frac{d}{dx}\left(\sin^{-1} x\right) = \frac{1}{f'(g(x))} = \frac{1}{\sqrt{1 - x^2}}$$

Analysis

To see that $\cos(\sin^{-1} x) = \sqrt{1 - x^2}$, consider the following argument. Set $\sin^{-1} x = \theta$. In this case, $\sin\theta = x$ where $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$. We begin by considering the case where $0 < \theta < \frac{\pi}{2}$. Since θ is an acute angle, we may construct a right triangle having acute angle θ, a hypotenuse of length 1 and the side opposite angle θ having length x. From the Pythagorean theorem, the side adjacent to angle θ has length $\sqrt{1 - x^2}$. This triangle is shown in Figure 3.9.2 Using the triangle, we see that $\cos(\sin^{-1} x) = \cos\theta = \sqrt{1 - x^2}$.

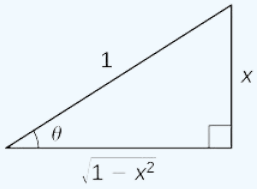


Figure 3.9.2: Using a right triangle having acute angle θ, a hypotenuse of length 1, and the side opposite angle θ having length x, we can see that $\cos(\sin^{-1} x) = \cos\theta = \sqrt{1 - x^2}$.

In the case where $-\frac{\pi}{2} < \theta < 0$, we make the observation that $0 < -\theta < \frac{\pi}{2}$ and hence

$$\cos\left(\sin^{-1} x\right) = \cos\theta = \cos(-\theta) = \sqrt{1 - x^2}.$$

Now if $\theta = \frac{\pi}{2}$ or $\theta = -\frac{\pi}{2}$, $x = 1$ or $x = -1$, and since in either case $\cos\theta = 0$ and $\sqrt{1 - x^2} = 0$, we have

$$\cos\left(\sin^{-1} x\right) = \cos\theta = \sqrt{1 - x^2}.$$

Consequently, in all cases,

$$\cos\left(\sin^{-1} x\right) = \sqrt{1 - x^2}.$$

Example 3.9.1B: Applying the Chain Rule to the Inverse Sine Function

Apply the chain rule to the formula derived in Example 3.9.4A to find the derivative of $h(x) = \sin^{-1}\left(g(x)\right)$ and use this result to find the derivative of $h(x) = \sin^{-1}(2x^3)$.

Solution

Applying the chain rule to $h(x) = \sin^{-1}\left(g(x)\right)$, we have

$$h'(x) = \frac{1}{\sqrt{1-\left(g(x)\right)^2}}g'(x).$$

Now let $g(x) = 2x^3$, so $g'(x) = 6x^2$. Substituting into the previous result, we obtain

$$\begin{aligned} h'(x) &= \frac{1}{\sqrt{1-4x^6}} \cdot 6x^2 \\ &= \frac{6x^2}{\sqrt{1-4x^6}} \end{aligned}$$

Exercise 3.9.1

Use the inverse function theorem to find the derivative of $g(x) = \tan^{-1} x$.

Hint

The inverse of $g(x)$ is $f(x) = \tan x$. Use Example 3.9.4Aas a guide.

Answer

$$g'(x) = \frac{1}{1+x^2}$$

The derivatives of the remaining inverse trigonometric functions may also be found by using the inverse function theorem. These formulas are provided in the following theorem.

Derivatives of Inverse Trigonometric Functions

$$\frac{d}{dx}\left(\sin^{-1} x\right) = \frac{1}{\sqrt{1-x^2}} \tag{3.9.1}$$

$$\frac{d}{dx}\left(\cos^{-1} x\right) = \frac{-1}{\sqrt{1-x^2}} \tag{3.9.2}$$

$$\frac{d}{dx}\left(\tan^{-1} x\right) = \frac{1}{1+x^2} \tag{3.9.3}$$

$$\frac{d}{dx}\left(\cot^{-1} x\right) = \frac{-1}{1+x^2} \tag{3.9.4}$$

$$\frac{d}{dx}\left(\sec^{-1} x\right) = \frac{1}{|x|\sqrt{x^2-1}} \tag{3.9.5}$$

$$\frac{d}{dx}\left(\csc^{-1} x\right) = \frac{-1}{|x|\sqrt{x^2-1}} \tag{3.9.6}$$

Example 3.9.2A: Applying Differentiation Formulas to an Inverse Tangent Function

Find the derivative of $f(x) = \tan^{-1}(x^2)$.

Solution

Let $g(x) = x^2$, so $g'(x) = 2x$. Substituting into Equation 3.9.3, we obtain

$$f'(x) = \frac{1}{1+(x^2)^2} \cdot (2x).$$

Simplifying, we have

$$f'(x) = \frac{2x}{1+x^4}.$$

Example 3.9.2B: Applying Differentiation Formulas to an Inverse Sine Function

Find the derivative of $h(x) = x^2 \sin^{-1} x$.

Solution

By applying the product rule, we have

$$h'(x) = 2x \sin^{-1} x + \frac{1}{\sqrt{1-x^2}} \cdot x^2$$

Exercise 3.9.2

Find the derivative of $h(x) = \cos^{-1}(3x-1)$.

Hint

Use Equation 3.9.2 with $g(x) = 3x - 1$

Answer

$$h'(x) = \frac{-3}{\sqrt{6x-9x^2}}$$

Example 3.9.3: Applying the Inverse Tangent Function

The position of a particle at time t is given by $s(t) = \tan^{-1}\left(\frac{1}{t}\right)$ for $t \geq \frac{1}{2}$. Find the velocity of the particle at time $t = 1$.

Solution

Begin by differentiating $s(t)$ in order to find $v(t)$.Thus,

$$v(t) = s'(t) = \frac{1}{1+\left(\frac{1}{t}\right)^2} \cdot \frac{-1}{t^2}.$$

Simplifying, we have

$$v(t) = -\frac{1}{t^2+1}.$$

Thus, $v(1) = -\frac{1}{2}$.

Exercise 3.9.3

Find the equation of the line tangent to the graph of $f(x) = \sin^{-1} x$ at $x = 0$.

Hint

$f'(0)$ is the slope of the tangent line.

Answer

$y = x$

3.9: Inverse Trigonometric Functions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **2.4: Continuity** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **3.7: Derivatives of Inverse Functions** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

3.10: Related Rates

Learning Objectives

- Express changing quantities in terms of derivatives.
- Find relationships among the derivatives in a given problem.
- Use the chain rule to find the rate of change of one quantity that depends on the rate of change of other quantities.

We have seen that for quantities that are changing over time, the rates at which these quantities change are given by derivatives. If two related quantities are changing over time, the rates at which the quantities change are related. For example, if a balloon is being filled with air, both the radius of the balloon and the volume of the balloon are increasing. In this section, we consider several problems in which two or more related quantities are changing and we study how to determine the relationship between the rates of change of these quantities.

Setting up Related-Rates Problems

In many real-world applications, related quantities are changing with respect to time. For example, if we consider the balloon example again, we can say that the rate of change in the volume, V, is related to the rate of change in the radius, r. In this case, we say that $\frac{dV}{dt}$ and $\frac{dr}{dt}$ are **related rates** because V is related to r. Here we study several examples of related quantities that are changing with respect to time and we look at how to calculate one rate of change given another rate of change.

Example 3.10.1: Inflating a Balloon

A spherical balloon is being filled with air at the constant rate of $2\ \text{cm}^3/\text{sec}$ (Figure 3.10.1). How fast is the radius increasing when the radius is 3 cm?

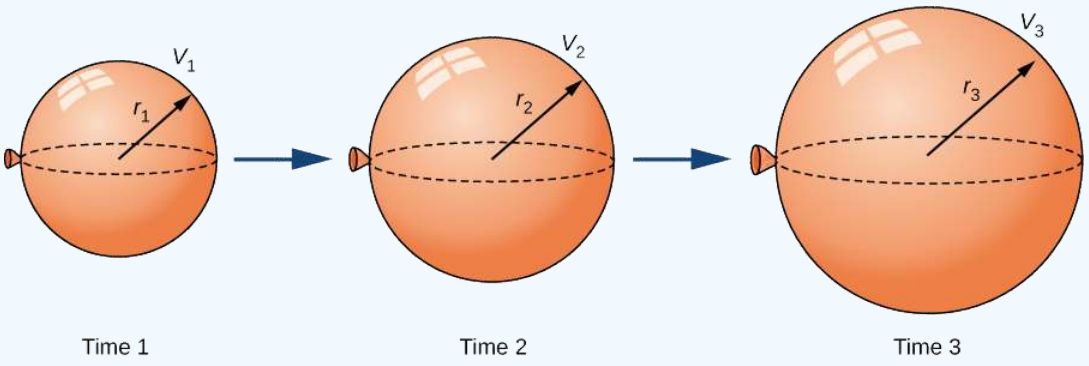


Figure 3.10.1: As the balloon is being filled with air, both the radius and the volume are increasing with respect to time.

Solution

The volume of a sphere of radius r centimeters is

$$V = \tfrac{4}{3}\pi r^3\ \text{cm}^3.$$

Since the balloon is being filled with air, both the volume and the radius are functions of time. Therefore, t seconds after beginning to fill the balloon with air, the volume of air in the balloon is

$$V(t) = \tfrac{4}{3}\pi\big[r(t)\big]^3\text{cm}^3.$$

Differentiating both sides of this equation with respect to time and applying the chain rule, we see that the rate of change in the volume is related to the rate of change in the radius by the equation

$$V'(t) = 4\pi\big[r(t)\big]^2 r'(t).$$

The balloon is being filled with air at the constant rate of $2\ \text{cm}^3/\text{sec}$, so $V'(t) = 2\ \text{cm}^3/\text{sec}$. Therefore,

$$2\ \text{cm}^3/\text{sec} = \Big(4\pi\big[r(t)\big]^2\ \text{cm}^2\Big)\cdot\Big(r'(t)\ \text{cm/s}\Big),$$

which implies

$$r'(t) = \frac{1}{2\pi\left[r(t)\right]^2} \text{ cm/sec}.$$

When the radius $r = 3$ cm,

$$r'(t) = \frac{1}{18\pi} \text{ cm/sec}.$$

Exercise 3.10.1

What is the instantaneous rate of change of the radius when $r = 6$ cm?

Hint

$$\frac{dr}{dt} = \frac{1}{2\pi r^2}$$

Answer

$\frac{1}{72\pi}$ cm/sec, or approximately 0.0044 cm/sec

Before looking at other examples, let's outline the problem-solving strategy we will be using to solve related-rates problems.

Problem-Solving Strategy: Solving a Related-Rates Problem

1. Assign symbols to all variables involved in the problem. Draw a figure if applicable.
2. State, in terms of the variables, the information that is given and the rate to be determined.
3. Find an equation relating the variables introduced in step 1.
4. Using the chain rule, differentiate both sides of the equation found in step 3 with respect to the independent variable. This new equation will relate the derivatives.
5. Substitute all known values into the equation from step 4, then solve for the unknown rate of change

Note that when solving a related-rates problem, it is crucial not to substitute known values too soon. For example, if the value for a changing quantity is substituted into an equation before both sides of the equation are differentiated, then that quantity will behave as a constant and its derivative will not appear in the new equation found in step 4. We examine this potential error in the following example.

Examples of the Process

Let's now implement the strategy just described to solve several related-rates problems. The first example involves a plane flying overhead. The relationship we are studying is between the speed of the plane and the rate at which the distance between the plane and a person on the ground is changing.

Example 3.10.2: An Airplane Flying at a Constant Elevation

An airplane is flying overhead at a constant elevation of 4000 ft. A man is viewing the plane from a position 3000 ft from the base of a radio tower. The airplane is flying horizontally away from the man. If the plane is flying at the rate of 600 ft/sec, at what rate is the distance between the man and the plane increasing when the plane passes over the radio tower?

Solution

Step 1. Draw a picture, introducing variables to represent the different quantities involved.

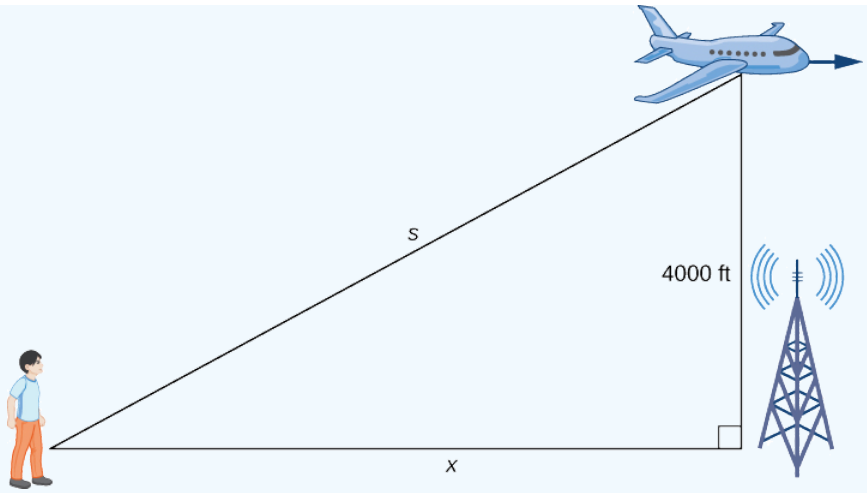


Figure 3.10.2: An airplane is flying at a constant height of 4000 ft. The distance between the person and the airplane and the person and the place on the ground directly below the airplane are changing. We denote those quantities with the variables s and x, respectively.

As shown, x denotes the distance between the man and the position on the ground directly below the airplane. The variable s denotes the distance between the man and the plane. Note that both x and s are functions of time. We do not introduce a variable for the height of the plane because it remains at a constant elevation of 4000 ft. Since an object's height above the ground is measured as the shortest distance between the object and the ground, the line segment of length 4000 ft is perpendicular to the line segment of length x feet, creating a right triangle.

Step 2. Since x denotes the horizontal distance between the man and the point on the ground below the plane, dx/dt represents the speed of the plane. We are told the speed of the plane is 600 ft/sec. Therefore, $\frac{dx}{dt} = 600$ ft/sec. Since we are asked to find the rate of change in the distance between the man and the plane when the plane is directly above the radio tower, we need to find ds/dt when $x = 3000$ ft.

Step 3. From the figure, we can use the **Pythagorean theorem** to write an equation relating x and s:

$$[x(t)]^2 + 4000^2 = [s(t)]^2.$$

Step 4. Differentiating this equation with respect to time and using the fact that the derivative of a constant is zero, we arrive at the equation

$$x\frac{dx}{dt} = s\frac{ds}{dt}.$$

Step 5. Find the rate at which the distance between the man and the plane is increasing when the plane is directly over the radio tower. That is, find $\frac{ds}{dt}$ when $x = 3000$ ft. Since the speed of the plane is 600 ft/sec, we know that $\frac{dx}{dt} = 600$ ft/sec. We are not given an explicit value for s; however, since we are trying to find $\frac{ds}{dt}$ when $x = 3000$ ft, we can use the Pythagorean theorem to determine the distance s when $x = 3000$ ft and the height is 4000 ft. Solving the equation

$$3000^2 + 4000^2 = s^2$$

for s, we have $s = 5000$ ft at the time of interest. Using these values, we conclude that ds/dt

is a solution of the equation

$$(3000)(600) = (5000) \cdot \frac{ds}{dt}.$$

Therefore,

$$\frac{ds}{dt} = \frac{3000 \cdot 600}{5000} = 360 \text{ ft/sec}.$$

Note: When solving related-rates problems, it is important not to substitute values for the variables too soon. For example, in step 3, we related the variable quantities $x(t)$ and $s(t)$ by the equation

$$[x(t)]^2 + 4000^2 = [s(t)]^2.$$

Since the plane remains at a constant height, it is not necessary to introduce a variable for the height, and we are allowed to use the constant 4000 to denote that quantity. However, the other two quantities are changing. If we mistakenly substituted $x(t) = 3000$ into the equation before differentiating, our equation would have been

$$3000^2 + 4000^2 = [s(t)]^2.$$

After differentiating, our equation would become

$$0 = s(t)\frac{ds}{dt}.$$

As a result, we would incorrectly conclude that $\frac{ds}{dt} = 0$.

? Exercise 3.10.2

What is the speed of the plane if the distance between the person and the plane is increasing at the rate of 300 ft/sec?

Hint

$\frac{ds}{dt} = 300$ ft/sec

Answer

500 ft/sec

We now return to the problem involving the rocket launch from the beginning of the chapter.

✓ Example 3.10.3: Chapter Opener - A Rocket Launch

A rocket is launched so that it rises vertically. A camera is positioned 5000 ft from the launch pad. When the rocket is 1000 ft above the launch pad, its velocity is 600 ft/sec.



Figure 3.10.3: (credit: modification of work by Steve Jurvetson, Wikimedia Commons)

Find the necessary rate of change of the camera's angle as a function of time so that it stays focused on the rocket.

Solution

Step 1. Draw a picture introducing the variables.

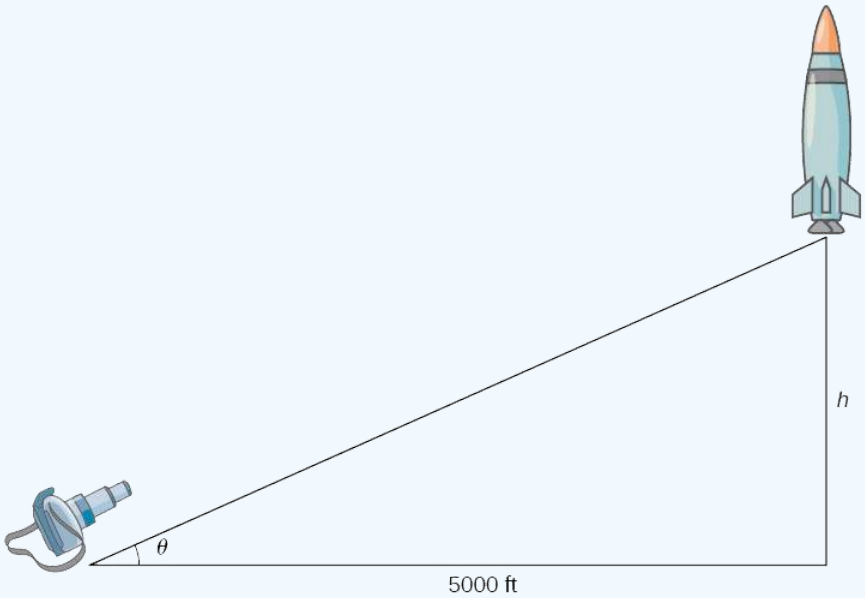


Figure 3.10.4: A camera is positioned 5000 ft from the launch pad of the rocket. The height of the rocket and the angle of the camera are changing with respect to time. We denote those quantities with the variables h and θ,respectively.

Let h denote the height of the rocket above the launch pad and θ be the angle between the camera lens and the ground.

Step 2. We are trying to find the rate of change in the angle of the camera with respect to time when the rocket is 1000 ft off the ground. That is, we need to find $\frac{d\theta}{dt}$ when $h = 1000$ ft. At that time, we know the velocity of the rocket is $\frac{dh}{dt} = 600$ ft/sec.

Step 3. Now we need to find an equation relating the two quantities that are changing with respect to time: h and θ. How can we create such an equation? Using the fact that we have drawn a right triangle, it is natural to think about trigonometric functions. Recall that $\tan\theta$ is the ratio of the length of the opposite side of the triangle to the length of the adjacent side. Thus, we have

$$\tan\theta = \frac{h}{5000}.$$

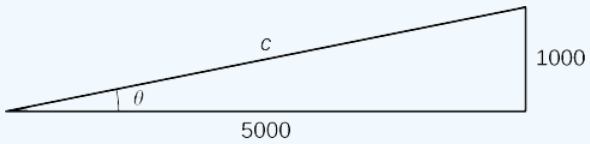
This gives us the equation

$$h = 5000\tan\theta.$$

Step 4. Differentiating this equation with respect to time t, we obtain

$$\frac{dh}{dt} = 5000\sec^2\theta\frac{d\theta}{dt}.$$

Step 5. We want to find $\frac{d\theta}{dt}$ when $h = 1000$ ft. At this time, we know that $\frac{dh}{dt} = 600$ ft/sec. We need to determine $\sec^2\theta$. Recall that $\sec\theta$ is the ratio of the length of the hypotenuse to the length of the adjacent side. We know the length of the adjacent side is 5000 ft. To determine the length of the hypotenuse, we use the Pythagorean theorem, where the length of one leg is 5000 ft, the length of the other leg is $h = 1000$ ft, and the length of the hypotenuse is c feet as shown in the following figure.



We see that

$$1000^2 + 5000^2 = c^2$$

and we conclude that the hypotenuse is

$$c = 1000\sqrt{26}\text{ ft}.$$

Therefore, when $h = 1000$, we have

$$\sec^2\theta = \left(\frac{1000\sqrt{26}}{5000}\right)^2 = \frac{26}{25}.$$

Recall from step 4 that the equation relating $\frac{d\theta}{dt}$ to our known values is

$$\frac{dh}{dt} = 5000\sec^2\theta\frac{d\theta}{dt}.$$

When $h = 1000$ ft, we know that $\frac{dh}{dt} = 600$ ft/sec and $\sec^2\theta = \frac{26}{25}$. Substituting these values into the previous equation, we arrive at the equation

$$600 = 5000\left(\tfrac{26}{25}\right)\frac{d\theta}{dt}.$$

Therefore, $\dfrac{d\theta}{dt} = \dfrac{3}{26}$ rad/sec.

Exercise 3.10.3

What rate of change is necessary for the elevation angle of the camera if the camera is placed on the ground at a distance of 4000 ft from the launch pad and the velocity of the rocket is 500 ft/sec when the rocket is 2000 ft off the ground?

Hint

Find $\frac{d\theta}{dt}$ when $h = 2000$ ft. At that time, $\frac{dh}{dt} = 500$ ft/sec.

Answer

$\frac{1}{10}$ rad/sec

In the next example, we consider water draining from a cone-shaped funnel. We compare the rate at which the level of water in the cone is decreasing with the rate at which the volume of water is decreasing.

Example 3.10.4: Water Draining from a Funnel

Water is draining from the bottom of a cone-shaped funnel at the rate of $0.03\,\text{ft}^3/\text{sec}$. The height of the funnel is 2 ft and the radius at the top of the funnel is 1 ft. At what rate is the height of the water in the funnel changing when the height of the water is $\frac{1}{2}$ ft?

Solution

Step 1: Draw a picture introducing the variables.

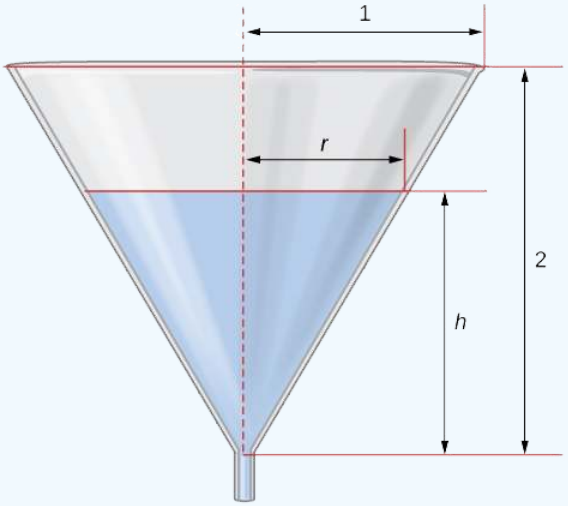


Figure 3.10.5: Water is draining from a funnel of height 2 ft and radius 1 ft. The height of the water and the radius of water are changing over time. We denote these quantities with the variables h and r, respectively.

Let h denote the height of the water in the funnel, r denote the radius of the water at its surface, and V denote the volume of the water.

Step 2: We need to determine $\frac{dh}{dt}$ when $h = \frac{1}{2}$ ft. We know that $\frac{dV}{dt} = -0.03$ ft/sec.

Step 3: The volume of water in the cone is

$$V = \tfrac{1}{3}\pi r^2 h.$$

From the figure, we see that we have similar triangles. Therefore, the ratio of the sides in the two triangles is the same. Therefore, $\frac{r}{h} = \frac{1}{2}$ or $r = \frac{h}{2}$. Using this fact, the equation for volume can be simplified to

$$V = \tfrac{1}{3}\pi\left(\tfrac{h}{2}\right)^2 h = \tfrac{\pi}{12}h^3.$$

Step 4: Applying the chain rule while differentiating both sides of this equation with respect to time t, we obtain

$$\frac{dV}{dt} = \frac{\pi}{4}h^2\frac{dh}{dt}.$$

Step 5: We want to find $\frac{dh}{dt}$ when $h = \frac{1}{2}$ ft. Since water is leaving at the rate of $0.03\,\text{ft}^3/\text{sec}$, we know that $\frac{dV}{dt} = -0.03\,\text{ft}^3/\text{sec}$. Therefore,

$$-0.03 = \frac{\pi}{4}\left(\frac{1}{2}\right)^2\frac{dh}{dt},$$

which implies

$$-0.03 = \frac{\pi}{16}\frac{dh}{dt}.$$

It follows that

$$\frac{dh}{dt} = -\frac{0.48}{\pi} = -0.153\,\text{ft/sec}.$$

? Exercise 3.10.4

At what rate is the height of the water changing when the height of the water is $\frac{1}{4}$ ft?

Hint

We need to find $\frac{dh}{dt}$ when $h = \frac{1}{4}$.

Answer

-0.61 ft/sec

Key Concepts

- To solve a related rates problem, first draw a picture that illustrates the relationship between the two or more related quantities that are changing with respect to time.
- In terms of the quantities, state the information given and the rate to be found.
- Find an equation relating the quantities.
- Use differentiation, applying the chain rule as necessary, to find an equation that relates the rates.
- Be sure not to substitute a variable quantity for one of the variables until after finding an equation relating the rates.

Glossary

related rates

are rates of change associated with two or more related quantities that are changing over time

---

This page titled 3.10: Related Rates is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **4.1: Related Rates** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

3.11: Linearization and Differentials

Learning Objectives

- Describe the linear approximation to a function at a point.
- Write the linearization of a given function.
- Draw a graph that illustrates the use of differentials to approximate the change in a quantity.
- Calculate the relative error and percentage error in using a differential approximation.

We have just seen how derivatives allow us to compare related quantities that are changing over time. In this section, we examine another application of derivatives: the ability to approximate functions locally by linear functions. Linear functions are the easiest functions with which to work, so they provide a useful tool for approximating function values. In addition, the ideas presented in this section are generalized later in the text when we study how to approximate functions by higher-degree polynomials Introduction to Power Series and Functions.

Linear Approximation of a Function at a Point

Consider a function f that is differentiable at a point $x = a$. Recall that the tangent line to the graph of f at a is given by the equation

$$y = f(a) + f'(a)(x - a).$$

For example, consider the function $f(x) = \frac{1}{x}$ at $a = 2$. Since f is differentiable at $x = 2$ and $f'(x) = -\frac{1}{x^2}$, we see that $f'(2) = -\frac{1}{4}$. Therefore, the tangent line to the graph of f at $a = 2$ is given by the equation

$$y = \frac{1}{2} - \frac{1}{4}(x - 2).$$

Figure $3.11.1a$ shows a graph of $f(x) = \frac{1}{x}$ along with the tangent line to f at $x = 2$. Note that for x near 2, the graph of the tangent line is close to the graph of f. As a result, we can use the equation of the tangent line to approximate $f(x)$ for x near 2. For example, if $x = 2.1$, the y value of the corresponding point on the tangent line is

$$y = \frac{1}{2} - \frac{1}{4}(2.1 - 2) = 0.475.$$

The actual value of $f(2.1)$ is given by

$$f(2.1) = \frac{1}{2.1} \approx 0.47619.$$

Therefore, the tangent line gives us a fairly good approximation of $f(2.1)$ (Figure $3.11.1b$). However, note that for values of x far from 2, the equation of the tangent line does not give us a good approximation. For example, if $x = 10$, the y-value of the corresponding point on the tangent line is

$$y = \frac{1}{2} - \frac{1}{4}(10 - 2) = \frac{1}{2} - 2 = -1.5,$$

whereas the value of the function at $x = 10$ is $f(10) = 0.1$.

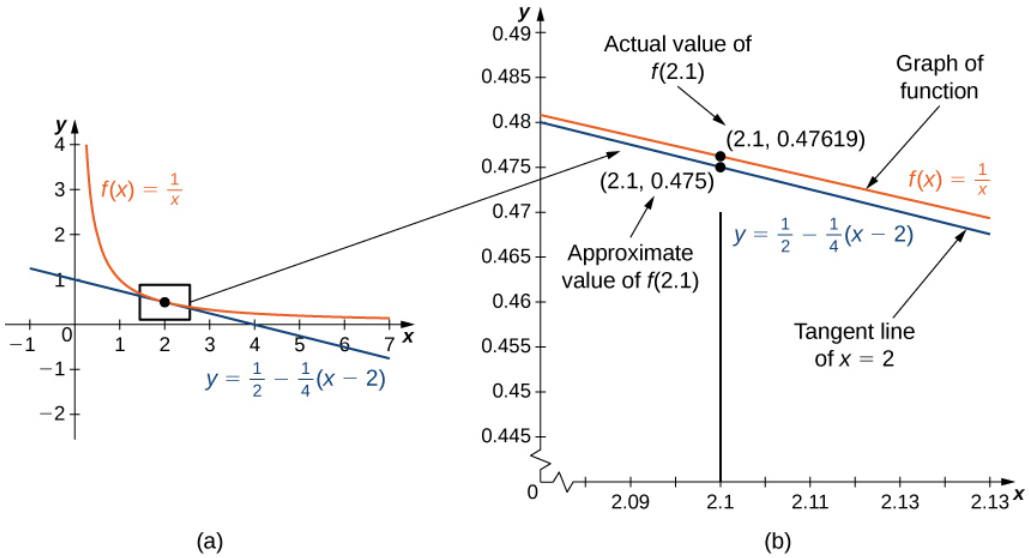


Figure 3.11.1: (a) The tangent line to $f(x) = 1/x$ at $x = 2$ provides a good approximation to f for x near 2. (b) At $x = 2.1$, the value of y on the tangent line to $f(x) = 1/x$ is 0.475. The actual value of $f(2.1)$ is $1/2.1$, which is approximately 0.47619.

In general, for a differentiable function f, the equation of the tangent line to f at $x = a$ can be used to approximate $f(x)$ for x near a. Therefore, we can write

$$f(x) \approx f(a) + f'(a)(x - a) \quad \text{for } x \text{ near } a.$$

We call the linear function

$$L(x) = f(a) + f'(a)(x - a) \tag{3.11.1}$$

the **linear approximation,** or **tangent line approximation**, of f at $x = a$. This function L is also known as the **linearization** of f at $x = a$.

To show how useful the linear approximation can be, we look at how to find the linear approximation for $f(x) = \sqrt{x}$ at $x = 9$.

✔ Example 3.11.1: Linear Approximation of $\sqrt{x}$

Find the linear approximation of $f(x) = \sqrt{x}$ at $x = 9$ and use the approximation to estimate $\sqrt{9.1}$.

Solution

Since we are looking for the linear approximation at $x = 9$, using Equation 3.11.1 we know the linear approximation is given by

$$L(x) = f(9) + f'(9)(x - 9).$$

We need to find $f(9)$ and $f'(9)$.

$$f(x) = \sqrt{x} \Rightarrow f(9) = \sqrt{9} = 3$$
$$f'(x) = \tfrac{1}{2\sqrt{x}} \Rightarrow f'(9) = \tfrac{1}{2\sqrt{9}} = \tfrac{1}{6}$$

Therefore, the linear approximation is given by Figure 3.11.2

$$L(x) = 3 + \frac{1}{6}(x - 9)$$

Using the linear approximation, we can estimate $\sqrt{9.1}$ by writing

$$\sqrt{9.1} = f(9.1) \approx L(9.1) = 3 + \frac{1}{6}(9.1 - 9) \approx 3.0167.$$

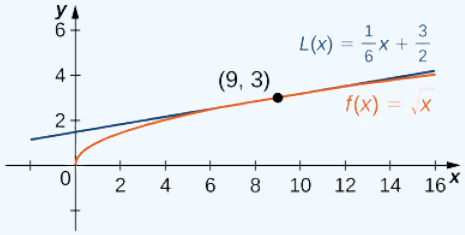


Figure 3.11.2: The local linear approximation to $f(x) = \sqrt{x}$ at $x = 9$ provides an approximation to f for x near 9.

Analysis

Using a calculator, the value of $\sqrt{9.1}$ to four decimal places is 3.0166. The value given by the linear approximation, 3.0167, is very close to the value obtained with a calculator, so it appears that using this linear approximation is a good way to estimate $\sqrt{x}$, at least for x near 9. At the same time, it may seem odd to use a linear approximation when we can just push a few buttons on a calculator to evaluate $\sqrt{9.1}$. However, how does the calculator evaluate $\sqrt{9.1}$? The calculator uses an approximation! In fact, calculators and computers use approximations all the time to evaluate mathematical expressions; they just use higher-degree approximations.

? Exercise 3.11.1

Find the local linear approximation to $f(x) = \sqrt[3]{x}$ at $x = 8$. Use it to approximate $\sqrt[3]{8.1}$ to five decimal places.

Hint

$L(x) = f(a) + f'(a)(x - a)$

Answer

$L(x) = 2 + \frac{1}{12}(x - 8)$; 2.00833

✔ Example 3.11.2: Linear Approximation of $\sin x$

Find the linear approximation of $f(x) = \sin x$ at $x = \frac{\pi}{3}$ and use it to approximate $\sin(62^\circ)$.

Solution

First we note that since $\frac{\pi}{3}$ rad is equivalent to 60°, using the linear approximation at $x = \pi/3$ seems reasonable. The linear approximation is given by

$$L(x) = f(\tfrac{\pi}{3}) + f'(\tfrac{\pi}{3})(x - \tfrac{\pi}{3}).$$

We see that

$$f(x) = \sin x \Rightarrow f(\tfrac{\pi}{3}) = \sin(\tfrac{\pi}{3}) = \tfrac{\sqrt{3}}{2}$$
$$f'(x) = \cos x \Rightarrow f'(\tfrac{\pi}{3}) = \cos(\tfrac{\pi}{3}) = \tfrac{1}{2}$$

Therefore, the linear approximation of f at $x = \pi/3$ is given by Figure 3.11.3

$$L(x) = \tfrac{\sqrt{3}}{2} + \tfrac{1}{2}(x - \tfrac{\pi}{3})$$

To estimate $\sin(62^\circ)$ using L, we must first convert 62° to radians. We have $62^\circ = \frac{62\pi}{180}$ radians, so the estimate for $\sin(62^\circ)$ is given by

$$\sin(62^\circ) = f(\tfrac{62\pi}{180}) \approx L(\tfrac{62\pi}{180}) = \tfrac{\sqrt{3}}{2} + \tfrac{1}{2}(\tfrac{62\pi}{180} - \tfrac{\pi}{3}) = \tfrac{\sqrt{3}}{2} + \tfrac{1}{2}(\tfrac{2\pi}{180}) = \tfrac{\sqrt{3}}{2} + \tfrac{\pi}{180} \approx 0.88348.$$

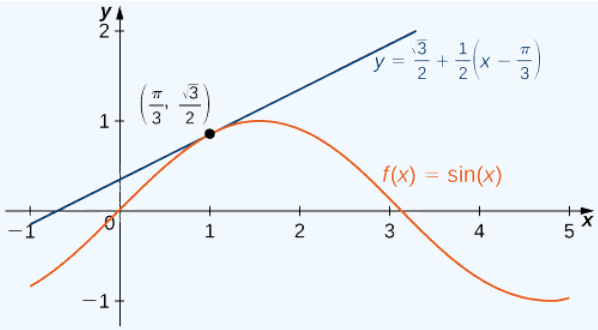


Figure 3.11.3: The linear approximation to $f(x) = \sin x$ at $x = \pi/3$ provides an approximation to $\sin x$ for x near $\pi/3$.

? Exercise 3.11.2

Find the linear approximation for $f(x) = \cos x$ at $x = \frac{\pi}{2}$.

Hint

$L(x) = f(a) + f'(a)(x - a)$

Answer

$L(x) = -x + \frac{\pi}{2}$

Linear approximations may be used in estimating roots and powers. In the next example, we find the linear approximation for $f(x) = (1 + x)^n$ at $x = 0$, which can be used to estimate roots and powers for real numbers near 1. The same idea can be extended to a function of the form $f(x) = (m + x)^n$ to estimate roots and powers near a different number m.

✓ Example 3.11.3: Approximating Roots and Powers

Find the linear approximation of $f(x) = (1 + x)^n$ at $x = 0$. Use this approximation to estimate $(1.01)^3$.

Solution

The linear approximation at $x = 0$ is given by

$$L(x) = f(0) + f'(0)(x - 0).$$

Because

$$f(x) = (1 + x)^n \Rightarrow f(0) = 1$$
$$f'(x) = n(1 + x)^{n-1} \Rightarrow f'(0) = n,$$

the linear approximation is given by Figure 3.11.4a

$$L(x) = 1 + n(x - 0) = 1 + nx$$

We can approximate $(1.01)^3$ by evaluating $L(0.01)$ when $n = 3$. We conclude that

$$(1.01)^3 = f(1.01) \approx L(1.01) = 1 + 3(0.01) = 1.03.$$

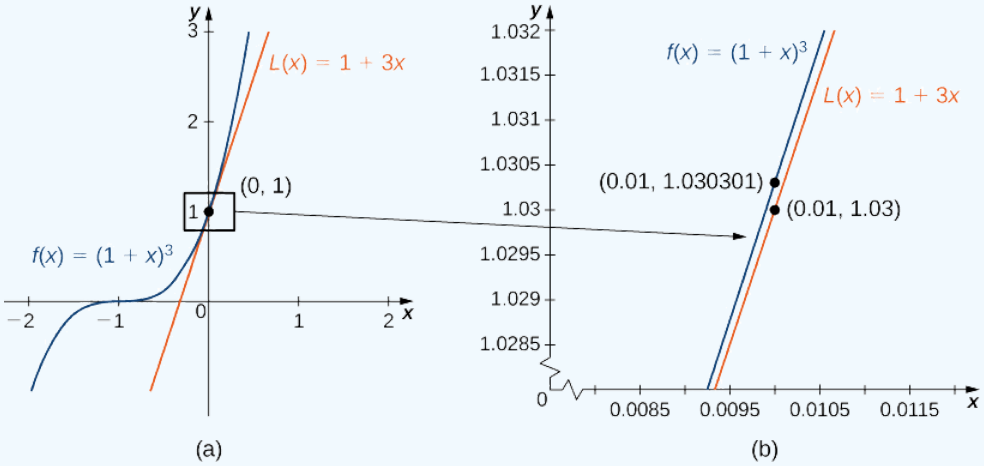


Figure 3.11.4: (a) The linear approximation of $f(x)$ at $x = 0$ is $L(x)$. (b) The actual value of 1.01^3 is 1.030301. The linear approximation of $f(x)$ at $x = 0$ estimates 1.01^3 to be 1.03.

? Exercise 3.11.3

Find the linear approximation of $f(x) = (1 + x)^4$ at $x = 0$ without using the result from the preceding example.

Hint

$f'(x) = 4(1 + x)^3$

Answer

$L(x) = 1 + 4x$

Differentials

We have seen that linear approximations can be used to estimate function values. They can also be used to estimate the amount a function value changes as a result of a small change in the input. To discuss this more formally, we define a related concept: **differentials**. Differentials provide us with a way of estimating the amount a function changes as a result of a small change in input values.

When we first looked at derivatives, we used the Leibniz notation dy/dx to represent the derivative of y with respect to x. Although we used the expressions dy and dx in this notation, they did not have meaning on their own. Here we see a meaning to the expressions dy and dx. Suppose $y = f(x)$ is a differentiable function. Let dx be an independent variable that can be assigned any nonzero real number, and define the dependent variable dy by

$$dy = f'(x)\,dx. \tag{3.11.2}$$

It is important to notice that dy is a function of both x and dx. The expressions dy and dx are called differentials. We can divide both sides of Equation 3.11.2 by dx, which yields

$$\frac{dy}{dx} = f'(x). \tag{3.11.3}$$

This is the familiar expression we have used to denote a derivative. Equation 3.11.3 is known as the **differential form** of Equation 3.11.2.

✓ Example 3.11.4: Computing Differentials

For each of the following functions, find dy and evaluate when $x = 3$ and $dx = 0.1$.

a. $y = x^2 + 2x$
b. $y = \cos x$

Solution

The key step is calculating the derivative. When we have that, we can obtain dy directly.

a. Since $f(x) = x^2 + 2x,$ we know $f'(x) = 2x + 2$, and therefore

$$dy = (2x + 2)\, dx.$$

When $x = 3$ and $dx = 0.1$,

$$dy = (2 \cdot 3 + 2)(0.1) = 0.8.$$

b. Since $f(x) = \cos x$, $f'(x) = -\sin(x)$. This gives us

$$dy = -\sin x\, dx.$$

When $x = 3$ and $dx = 0.1$,

$$dy = -\sin(3)(0.1) = -0.1 \sin(3).$$

? Exercise 3.11.4

For $y = e^{x^2}$, find dy.

Hint

$dy = f'(x)\, dx$

Answer

$dy = 2xe^{x^2} dx$

We now connect differentials to linear approximations. Differentials can be used to estimate the change in the value of a function resulting from a small change in input values. Consider a function f that is differentiable at point a. Suppose the input x changes by a small amount. We are interested in how much the output y changes. If x changes from a to $a + dx$, then the change in x is dx (also denoted Δx), and the change in y is given by

$$\Delta y = f(a + dx) - f(a).$$

Instead of calculating the exact change in y, however, it is often easier to approximate the change in y by using a linear approximation. For x near a, $f(x)$ can be approximated by the linear approximation (Equation 3.11.1)

$$L(x) = f(a) + f'(a)(x - a).$$

Therefore, if dx is small,

$$f(a + dx) \approx L(a + dx) = f(a) + f'(a)(a + dx - a).$$

That is,

$$f(a + dx) - f(a) \approx L(a + dx) - f(a) = f'(a)\, dx.$$

In other words, the actual change in the function f if x increases from a to $a + dx$ is approximately the difference between $L(a + dx)$ and $f(a)$, where $L(x)$ is the linear approximation of f at a. By definition of $L(x)$, this difference is equal to $f'(a)\, dx$. In summary,

$$\Delta y = f(a + dx) - f(a) \approx L(a + dx) - f(a) = f'(a)\, dx = dy.$$

Therefore, we can use the differential $dy = f'(a)\, dx$ to approximate the change in y if x increases from $x = a$ to $x = a + dx$. We can see this in the following graph.

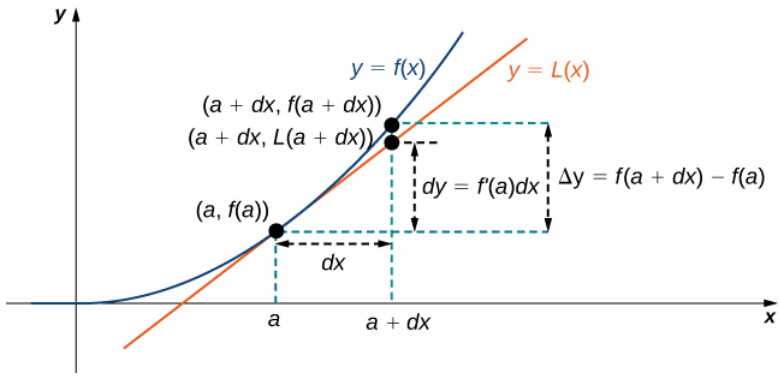


Figure 3.11.5: The differential $dy = f'(a)\,dx$ is used to approximate the actual change in y if x increases from a to $a + dx$.

We now take a look at how to use differentials to approximate the change in the value of the function that results from a small change in the value of the input. Note the calculation with differentials is much simpler than calculating actual values of functions and the result is very close to what we would obtain with the more exact calculation.

Example 3.11.5: Approximating Change with Differentials

Let $y = x^2 + 2x$. Compute Δy and dy at $x = 3$ if $dx = 0.1$.

Solution

The actual change in y if x changes from $x = 3$ to $x = 3.1$ is given by

$$\Delta y = f(3.1) - f(3) = [(3.1)^2 + 2(3.1)] - [3^2 + 2(3)] = 0.81.$$

The approximate change in y is given by $dy = f'(3)\,dx$. Since $f'(x) = 2x + 2$, we have

$$dy = f'(3)\,dx = (2(3) + 2)(0.1) = 0.8.$$

Exercise 3.11.5

For $y = x^2 + 2x$, find Δy and dy at $x = 3$ if $dx = 0.2$.

Hint

$dy = f'(3)\,dx,\ \Delta y = f(3.2) - f(3)$

Answer

$dy = 1.6,\ \Delta y = 1.64$

Calculating the Amount of Error

Any type of measurement is prone to a certain amount of error. In many applications, certain quantities are calculated based on measurements. For example, the area of a circle is calculated by measuring the radius of the circle. An error in the measurement of the radius leads to an error in the computed value of the area. Here we examine this type of error and study how differentials can be used to estimate the error.

Consider a function f with an input that is a measured quantity. Suppose the exact value of the measured quantity is a, but the measured value is $a + dx$. We say the measurement error is dx (or Δx). As a result, an error occurs in the calculated quantity $f(x)$. This type of error is known as a **propagated error** and is given by

$$\Delta y = f(a + dx) - f(a).$$

Since all measurements are prone to some degree of error, we do not know the exact value of a measured quantity, so we cannot calculate the propagated error exactly. However, given an estimate of the accuracy of a measurement, we can use differentials to approximate the propagated error Δy. Specifically, if f is a differentiable function at a,the propagated error is

$$\Delta y \approx dy = f'(a)\,dx.$$

Unfortunately, we do not know the exact value a. However, we can use the measured value $a+dx,$ and estimate

$$\Delta y \approx dy \approx f'(a+dx)\,dx.$$

In the next example, we look at how differentials can be used to estimate the error in calculating the volume of a box if we assume the measurement of the side length is made with a certain amount of accuracy.

Example 3.11.6: Volume of a Cube

Suppose the side length of a cube is measured to be 5 cm with an accuracy of 0.1 cm.

a. Use differentials to estimate the error in the computed volume of the cube.
b. Compute the volume of the cube if the side length is (i) 4.9 cm and (ii) 5.1 cm to compare the estimated error with the actual potential error.

Solution

a. The measurement of the side length is accurate to within ± 0.1 cm. Therefore,

$$-0.1 \le dx \le 0.1.$$

The volume of a cube is given by $V = x^3$, which leads to

$$dV = 3x^2 dx.$$

Using the measured side length of 5 cm, we can estimate that

$$-3(5)^2(0.1) \le dV \le 3(5)^2(0.1).$$

Therefore,

$$-7.5 \le dV \le 7.5.$$

b. If the side length is actually 4.9 cm, then the volume of the cube is

$$V(4.9) = (4.9)^3 = 117.649\text{cm}^3.$$

If the side length is actually 5.1 cm, then the volume of the cube is

$$V(5.1) = (5.1)^3 = 132.651\text{cm}^3.$$

Therefore, the actual volume of the cube is between $117.649$and 132.651. Since the side length is measured to be 5 cm, the computed volume is $V(5) = 5^3 = 125.$ Therefore, the error in the computed volume is

$$117.649 - 125 \le \Delta V \le 132.651 - 125.$$

That is,

$$-7.351 \le \Delta V \le 7.651.$$

We see the estimated error dV is relatively close to the actual potential error in the computed volume.

Exercise 3.11.6

Estimate the error in the computed volume of a cube if the side length is measured to be 6 cm with an accuracy of 0.2 cm.

Hint

$dV = 3x^2 dx$

Answer

The volume measurement is accurate to within 21.6 cm^3.

The measurement error dx $(= \Delta x)$ and the propagated error Δy are absolute errors. We are typically interested in the size of an error relative to the size of the quantity being measured or calculated. Given an absolute error Δq for a particular quantity, we define the **relative error** as $\frac{\Delta q}{q}$, where q is the actual value of the quantity. The **percentage error** is the relative error expressed as

a percentage. For example, if we measure the height of a ladder to be 63 in. when the actual height is 62 in., the absolute error is 1 in. but the relative error is $\frac{1}{62} = 0.016$, or 1.6%. By comparison, if we measure the width of a piece of cardboard to be 8.25 in. when the actual width is 8 in., our absolute error is $\frac{1}{4}$ in., whereas the relative error is $\frac{0.25}{8} = \frac{1}{32}$, or 3.1%. Therefore, the percentage error in the measurement of the cardboard is larger, even though 0.25 in. is less than 1 in.

Example 3.11.7: Relative and Percentage Error

An astronaut using a camera measures the radius of Earth as 4000 mi with an error of ± 80 mi. Let's use differentials to estimate the relative and percentage error of using this radius measurement to calculate the volume of Earth, assuming the planet is a perfect sphere.

Solution: If the measurement of the radius is accurate to within ± 80, we have

$$-80 \le dr \le 80.$$

Since the volume of a sphere is given by $V = (\frac{4}{3})\pi r^3$, we have

$$dV = 4\pi r^2 dr.$$

Using the measured radius of 4000 mi, we can estimate

$$-4\pi(4000)^2(80) \le dV \le 4\pi(4000)^2(80).$$

To estimate the relative error, consider $\dfrac{dV}{V}$. Since we do not know the exact value of the volume V, use the measured radius $r = 4000$ mi to estimate V. We obtain $V \approx (\frac{4}{3})\pi(4000)^3$. Therefore the relative error satisfies

$$\frac{-4\pi(4000)^2(80)}{4\pi(4000)^3/3} \le \frac{dV}{V} \le \frac{4\pi(4000)^2(80)}{4\pi(4000)^3/3},$$

which simplifies to

$$-0.06 \le \frac{dV}{V} \le 0.06.$$

The relative error is 0.06 and the percentage error is 6%.

Exercise 3.11.7

Determine the percentage error if the radius of Earth is measured to be 3950 mi with an error of ± 100 mi.

Hint

Use the fact that $dV = 4\pi r^2 dr$ to find dV/V.

Answer

7.6%

Key Concepts

- A differentiable function $y = f(x)$ can be approximated at a by the linear function
$$L(x) = f(a) + f'(a)(x - a).$$
- For a function $y = f(x)$, if x changes from a to $a + dx$, then
$$dy = f'(x)\,dx$$
is an approximation for the change in y. The actual change in y is
$$\Delta y = f(a + dx) - f(a).$$
- A measurement error dx can lead to an error in a calculated quantity $f(x)$. The error in the calculated quantity is known as the propagated error. The propagated error can be estimated by
$$dy \approx f'(x)\,dx.$$

- To estimate the relative error of a particular quantity q, we estimate $\frac{\Delta q}{q}$.

Key Equations

- **Linear approximation**

$L(x) = f(a) + f'(a)(x - a)$

- **A differential**

$dy = f'(x)\,dx$

Glossary

differential

the differential dx is an independent variable that can be assigned any nonzero real number; the differential dy is defined to be $dy = f'(x)\,dx$

differential form

given a differentiable function $y = f'(x),$ the equation $dy = f'(x)\,dx$ is the differential form of the derivative of y with respect to x

linear approximation

the linear function $L(x) = f(a) + f'(a)(x - a)$ is the linear approximation of f at $x = a$

percentage error

the relative error expressed as a percentage

propagated error

the error that results in a calculated quantity $f(x)$ resulting from a measurement error dx

relative error

given an absolute error Δq for a particular quantity, $\frac{\Delta q}{q}$ is the relative error.

tangent line approximation (linearization)

since the linear approximation of f at $x = a$ is defined using the equation of the tangent line, the linear approximation of f at $x = a$ is also known as the tangent line approximation to f at $x = a$

3.11: Linearization and Differentials is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **4.2: Linear Approximations and Differentials** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

CHAPTER OVERVIEW

4: Applications of Derivatives

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

4.1: Maxima and Minima
4.2: The Mean Value Theorem
4.3: The First Derivative Test
4.4: The Second Derivative Test and Curve Sketching
4.5: Indeterminate Forms and L'Hôpital's Rule
4.6: Applied Optimization Problems
4.7: Newton's Method

4: Applications of Derivatives is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

4.1: Maxima and Minima

Learning Objectives

- Define absolute extrema.
- Define local extrema.
- Explain how to find the critical points of a function over a closed interval.
- Describe how to use critical points to locate absolute extrema over a closed interval.

Given a particular function, we are often interested in determining the largest and smallest values of the function. This information is important in creating accurate graphs. Finding the maximum and minimum values of a function also has practical significance, because we can use this method to solve optimization problems, such as maximizing profit, minimizing the amount of material used in manufacturing an aluminum can, or finding the maximum height a rocket can reach. In this section, we look at how to use derivatives to find the largest and smallest values for a function.

Absolute Extrema

Consider the function $f(x) = x^2 + 1$ over the interval $(-\infty, \infty)$. As $x \to \pm\infty$, $f(x) \to \infty$. Therefore, the function does not have a largest value. However, since $x^2 + 1 \geq 1$ for all real numbers x and $x^2 + 1 = 1$ when $x = 0$, the function has a smallest value, 1, when $x = 0$. We say that 1 is the absolute minimum of $f(x) = x^2 + 1$ and it occurs at $x = 0$. We say that $f(x) = x^2 + 1$ does not have an absolute maximum (Figure 4.1.1).

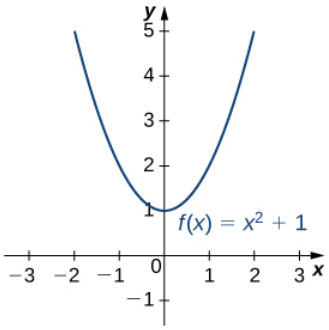


Figure 4.1.1: The given function has an absolute minimum of 1 at $x = 0$. The function does not have an absolute maximum.

Definition: Absolute Extrema

Let f be a function defined over an interval I and let $c \in I$. We say f has an *absolute maximum* on I at c if $f(c) \geq f(x)$ for all $x \in I$. We say f has an *absolute minimum* on I at c if $f(c) \leq f(x)$ for all $x \in I$. If f has an absolute maximum on I at c or an absolute minimum on I at c, we say f has *an absolute extremum* on I at c.

Before proceeding, let's note two important issues regarding this definition. First, the term absolute here does not refer to absolute value. An absolute extremum may be positive, negative, or zero. Second, if a function f has an absolute extremum over an interval I at c, the absolute extremum is $f(c)$. The real number c is a point in the domain at which the absolute extremum occurs. For example, consider the function $f(x) = 1/(x^2 + 1)$ over the interval $(-\infty, \infty)$. Since

$$f(0) = 1 \geq \frac{1}{x^2 + 1} = f(x)$$

for all real numbers x, we say f has an absolute maximum over $(-\infty, \infty)$ at $x = 0$. The absolute maximum is $f(0) = 1$. It occurs at $x = 0$, as shown in Figure 4.1.2(b).

A function may have both an absolute maximum and an absolute minimum, just one extremum, or neither. Figure 4.1.2 shows several functions and some of the different possibilities regarding absolute extrema. However, the following theorem, called the **Extreme Value Theorem**, guarantees that a continuous function f over a closed, bounded interval $[a, b]$ has both an absolute maximum and an absolute minimum.

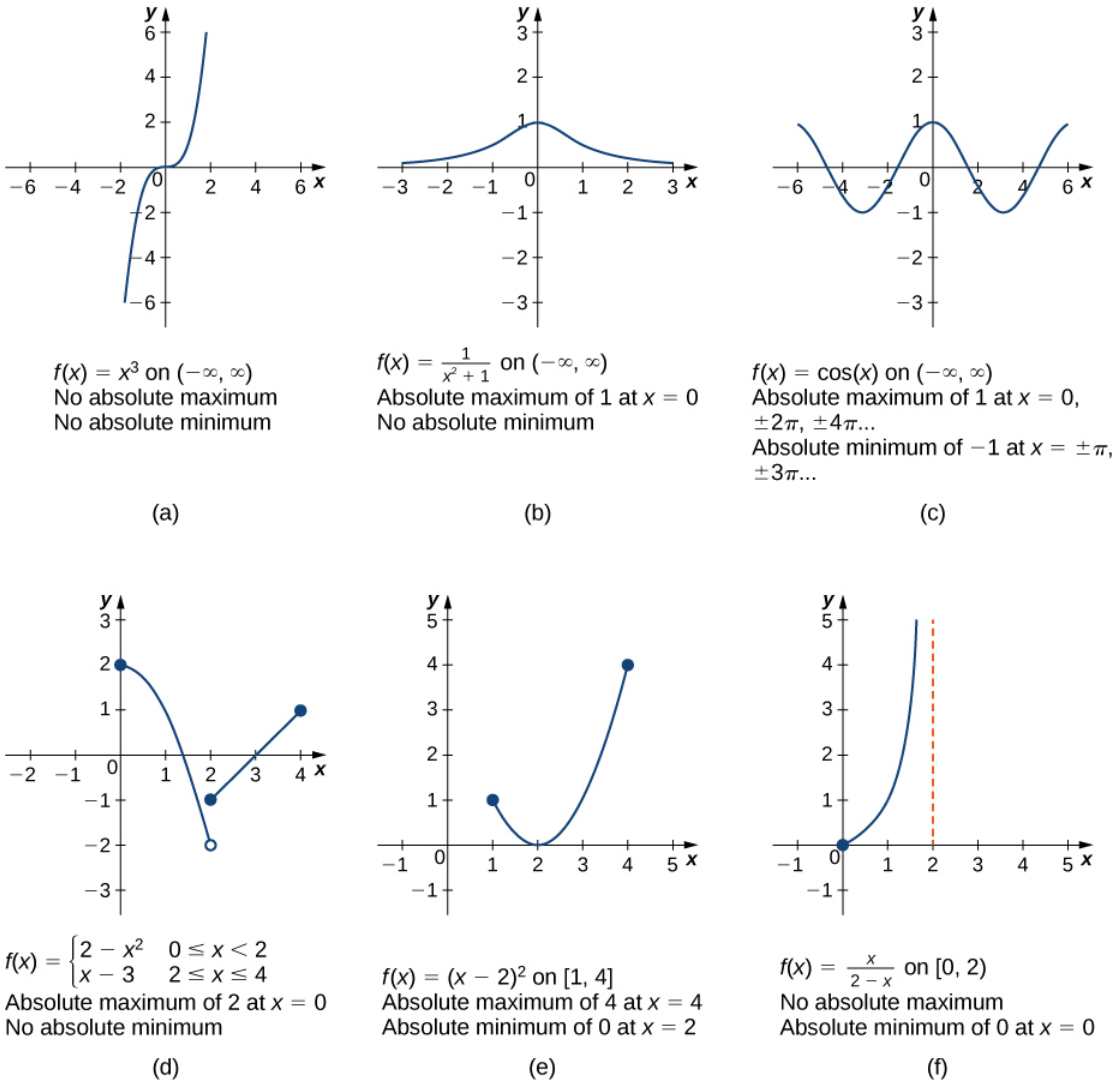


Figure 4.1.2: Graphs (a), (b), and (c) show several possibilities for absolute extrema for functions with a domain of $(-\infty, \infty)$. Graphs (d), (e), and (f) show several possibilities for absolute extrema for functions with a domain that is a bounded interval.

Theorem 4.1.1: Extreme Value Theorem

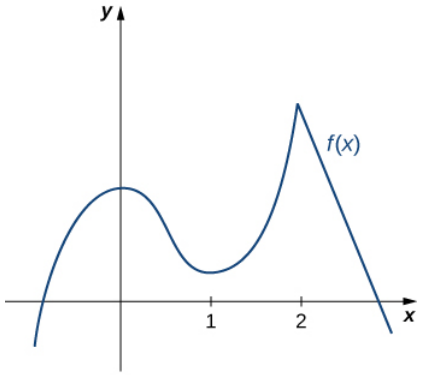
If f is a continuous function over the closed, bounded interval $[a, b]$, then there is a point in $[a, b]$ at which f has an absolute maximum over $[a, b]$ and there is a point in $[a, b]$ at which f has an absolute minimum over $[a, b]$.

The proof of the extreme value theorem is beyond the scope of this text. Typically, it is proved in a course on real analysis. There are a couple of key points to note about the statement of this theorem. For the extreme value theorem to apply, the function must be continuous over a closed, bounded interval. If the interval I is open or the function has even one point of discontinuity, the function may not have an absolute maximum or absolute minimum over I. For example, consider the functions shown in Figure $4.1.2$(d), (e), and (f). All three of these functions are defined over bounded intervals. However, the function in graph (e) is the only one that has both an absolute maximum and an absolute minimum over its domain. The extreme value theorem cannot be applied to the functions in graphs (d) and (f) because neither of these functions is continuous over a closed, bounded interval. Although the function in graph (d) is defined over the closed interval $[0, 4]$, the function is discontinuous at $x = 2$. The function has an absolute maximum over $[0, 4]$ but does not have an absolute minimum. The function in graph (f) is continuous over the half-open interval $[0, 2)$, but is not defined at $x = 2$, and therefore is not continuous over a closed, bounded interval. The function has an absolute minimum over $[0, 2)$, but does not have an absolute maximum over $[0, 2)$. These two graphs illustrate why a function over a bounded interval may fail to have an absolute maximum and/or absolute minimum.

Before looking at how to find absolute extrema, let's examine the related concept of local extrema. This idea is useful in determining where absolute extrema occur.

Local Extrema and Critical Points

Consider the function f shown in Figure 4.1.3. The graph can be described as two mountains with a valley in the middle. The absolute maximum value of the function occurs at the higher peak, at $x = 2$. However, $x = 0$ is also a point of interest. Although $f(0)$ is not the largest value of f, the value $f(0)$ is larger than $f(x)$ for all x near 0. We say f has a local maximum at $x = 0$. Similarly, the function f does not have an absolute minimum, but it does have a local minimum at $x = 1$ because $f(1)$ is less than $f(x)$ for x near 1.



$f(x)$ defined on $(-\infty, \infty)$
Local maxima at $x = 0$ and $x = 2$
Local minimum at $x = 1$

Figure 4.1.3: This function f has two local maxima and one local minimum. The local maximum at $x = 2$ is also the absolute maximum.

Definition: Local Extrema

A function f has a *local maximum* at c if there exists an open interval I containing c such that I is contained in the domain of f and $f(c) \geq f(x)$ for all $x \in I$. A function f has a *local minimum* at c if there exists an open interval I containing c such that I is contained in the domain of f and $f(c) \leq f(x)$ for all $x \in I$. A function f has a *local extremum* at c if f has a local maximum at c or f has a local minimum at c.

Note that if f has an absolute extremum at c and f is defined over an interval containing c, then $f(c)$ is also considered a local extremum. If an absolute extremum for a function f occurs at an endpoint, we do not consider that to be a local extremum, but instead refer to that as an endpoint extremum.

Given the graph of a function f, it is sometimes easy to see where a local maximum or local minimum occurs. However, it is not always easy to see, since the interesting features on the graph of a function may not be visible because they occur at a very small scale. Also, we may not have a graph of the function. In these cases, how can we use a formula for a function to determine where these extrema occur?

To answer this question, let's look at Figure 4.1.3 again. The local extrema occur at $x = 0, x = 1$, and $x = 2$. Notice that at $x = 0$ and $x = 1$, the derivative $f'(x) = 0$. At $x = 2$, the derivative $f'(x)$ does not exist, since the function f has a corner there. In fact, if f has a local extremum at a point $x = c$, the derivative $f'(c)$ must satisfy one of the following conditions: either $f'(c) = 0$ or $f'(c)$ is undefined. Such a value c is known as a critical point and it is important in finding extreme values for functions.

Definition: Critical Points

Let c be an interior point in the domain of f. We say that c is a *critical point* of f if $f'(c) = 0$ or $f'(c)$ is undefined.

As mentioned earlier, if f has a local extremum at a point $x = c$, then c must be a critical point of f. This fact is known as **Fermat's theorem.**

Theorem 4.1.2: Fermat's Theorem

If f has a local extremum at c and f is differentiable at c, then $f'(c) = 0$.

Proof

Suppose f has a local extremum at c and f is differentiable at c. We need to show that $f'(c) = 0$. To do this, we will show that $f'(c) \geq 0$ and $f'(c) \leq 0$, and therefore $f'(c) = 0$. Since f has a local extremum at c, f has a local maximum or local minimum at c. Suppose f has a local maximum at c. The case in which f has a local minimum at c can be handled similarly. There then exists an open interval I such that $f(c) \geq f(x)$ for all $x \in I$. Since f is differentiable at c, from the definition of the derivative, we know that

$$f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c}.$$

Since this limit exists, both one-sided limits also exist and equal $f'(c)$. Therefore,

$$f'(c) = \lim_{x \to c^+} \frac{f(x) - f(c)}{x - c,} \tag{4.1.1}$$

and

$$f'(c) = \lim_{x \to c^-} \frac{f(x) - f(c)}{x - c}.$$

Since $f(c)$ is a local maximum, we see that $f(x) - f(c) \leq 0$ for x near c. Therefore, for x near c, but $x > c$, we have $\frac{f(x)-f(c)}{x-c} \leq 0$. From Equation 4.1.1 we conclude that $f'(c) \leq 0$. Similarly, it can be shown that $f'(c) \geq 0$. Therefore, $f'(c) = 0$.

□

From Fermat's theorem, we conclude that if f has a local extremum at c, then either $f'(c) = 0$ or $f'(c)$ is undefined. In other words, local extrema can only occur at critical points.

Note this theorem does not claim that a function f must have a local extremum at a critical point. Rather, it states that critical points are candidates for local extrema. For example, consider the function $f(x) = x^3$. We have $f'(x) = 3x^2 = 0$ when $x = 0$. Therefore, $x = 0$ is a critical point. However, $f(x) = x^3$ is increasing over $(-\infty, \infty)$, and thus f does not have a local extremum at $x = 0$. In Figure 4.1.4, we see several different possibilities for critical points. In some of these cases, the functions have local extrema at critical points, whereas in other cases the functions do not. Note that these graphs do not show all possibilities for the behavior of a function at a critical point.

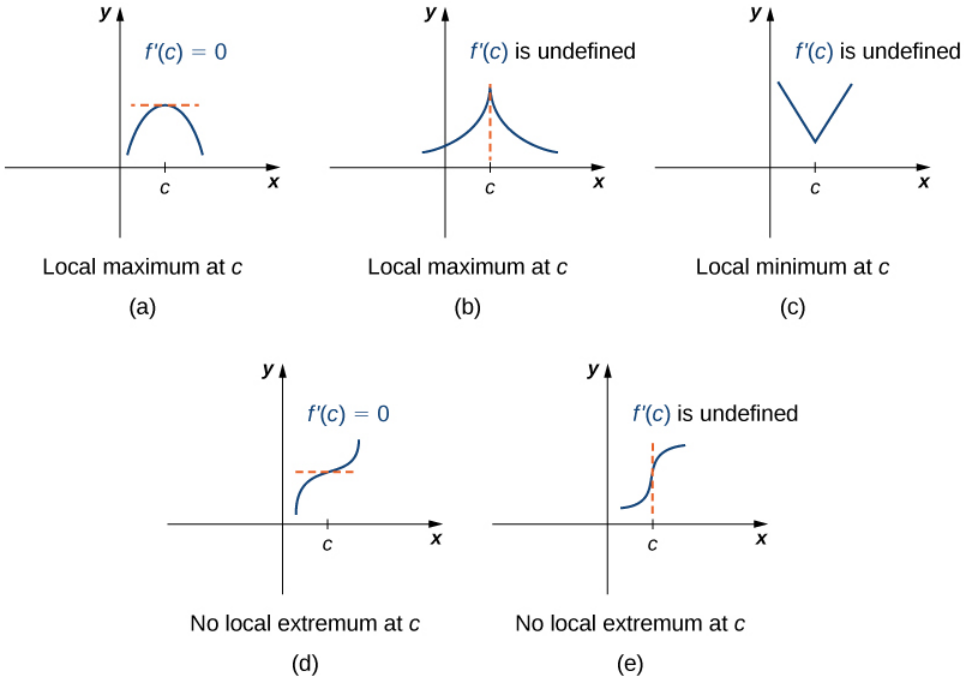


Figure 4.1.4: (a–e) A function f has a critical point at c if $f'(c) = 0$ or $f'(c)$ is undefined. A function may or may not have a local extremum at a critical point.

Later in this chapter we look at analytical methods for determining whether a function actually has a local extremum at a critical point. For now, let's turn our attention to finding critical points. We will use graphical observations to determine whether a critical point is associated with a local extremum.

Example 4.1.1: Locating Critical Points

For each of the following functions, find all critical points. Use a graphing utility to determine whether the function has a local extremum at each of the critical points.

a. $f(x) = \frac{1}{3}x^3 - \frac{5}{2}x^2 + 4x$
b. $f(x) = (x^2 - 1)^3$
c. $f(x) = \frac{4x}{1+x^2}$

Solution

a. The derivative $f'(x) = x^2 - 5x + 4$ is defined for all real numbers x. Therefore, we only need to find the values for x where $f'(x) = 0$. Since $f'(x) = x^2 - 5x + 4 = (x-4)(x-1)$, the critical points are $x = 1$ and $x = 4$. From the graph of f in Figure 4.1.5, we see that f has a local maximum at $x = 1$ and a local minimum at $x = 4$.

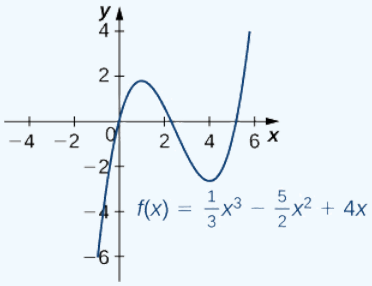


Figure 4.1.5: This function has a local maximum and a local minimum.

b. Using the chain rule, we see the derivative is

$$f'(x) = 3(x^2-1)^2(2x) = 6x(x^2-1)^2.$$

Therefore, f has critical points when $x = 0$ and when $x^2 - 1 = 0$. We conclude that the critical points are $x = 0, \pm 1$. From the graph of f in Figure 4.1.6, we see that f has a local (and absolute) minimum at $x = 0$, but does not have a local extremum

at $x = 1$ or $x = -1$.

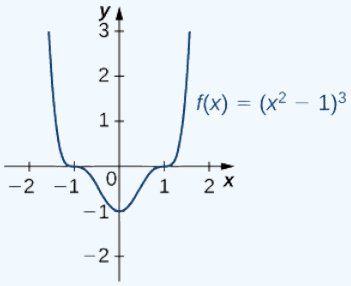


Figure 4.1.6: This function has three critical points: $x = 0$, $x = 1$, and $x = -1$. The function has a local (and absolute) minimum at $x = 0$, but does not have extrema at the other two critical points.

c. By the quotient rule, we see that the derivative is

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

The derivative is defined everywhere. Therefore, we only need to find values for x where $f'(x) = 0$. Solving $f'(x) = 0$, we see that $4 - 4x^2 = 0,$ which implies $x = \pm 1$. Therefore, the critical points are $x = \pm 1$. From the graph of f in Figure 4.1.7, we see that f has an absolute maximum at $x = 1$ and an absolute minimum at $x = -1$. Hence, f has a local maximum at $x = 1$ and a local minimum at $x = -1$. (Note that if f has an absolute extremum over an interval I at a point c that is not an endpoint of I, then f has a local extremum at c.)

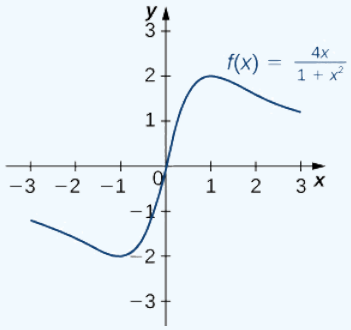


Figure 4.1.7: This function has an absolute maximum and an absolute minimum.

? Exercise 4.1.1

Find all critical points for $f(x) = x^3 - \frac{1}{2}x^2 - 2x + 1$.

Hint

Calculate $f'(x)$.

Answer

$x = \frac{-2}{3}, x = 1$

Locating Absolute Extrema

The extreme value theorem states that a continuous function over a closed, bounded interval has an absolute maximum and an absolute minimum. As shown in Figure 4.1.2, one or both of these absolute extrema could occur at an endpoint. If an absolute extremum does not occur at an endpoint, however, it must occur at an interior point, in which case the absolute extremum is a local extremum. Therefore, by Fermat's Theorem, the point c at which the local extremum occurs must be a critical point. We summarize this result in the following theorem.

Theorem 4.1.3: Location of Absolute Extrema

Let f be a continuous function over a closed, bounded interval I. The absolute maximum of f over I and the absolute minimum of f over I must occur at endpoints of I or at critical points of f in I.

With this idea in mind, let's examine a procedure for locating absolute extrema.

Problem-Solving Strategy: Locating Absolute Extrema over a Closed Interval

Consider a continuous function f defined over the closed interval $[a, b]$.

1. Evaluate f at the endpoints $x = a$ and $x = b$.
2. Find all critical points of f that lie over the interval (a, b) and evaluate f at those critical points.
3. Compare all values found in (1) and (2). From "Location of Absolute Extrema," the absolute extrema must occur at endpoints or critical points. Therefore, the largest of these values is the absolute maximum of f. The smallest of these values is the absolute minimum of f.

Now let's look at how to use this strategy to find the absolute maximum and absolute minimum values for continuous functions.

Example 4.1.2: Locating Absolute Extrema

For each of the following functions, find the absolute maximum and absolute minimum over the specified interval and state where those values occur.

a. $f(x) = -x^2 + 3x - 2$ over $[1, 3]$.
b. $f(x) = x^2 - 3x^{2/3}$ over $[0, 2]$.

Solution

a. Step 1. Evaluate f at the endpoints $x = 1$ and $x = 3$.

$$f(1) = 0 \text{ and } f(3) = -2$$

Step 2. Since $f'(x) = -2x + 3$, f' is defined for all real numbers x. Therefore, there are no critical points where the derivative is undefined. It remains to check where $f'(x) = 0$. Since $f'(x) = -2x + 3 = 0$ at $x = \frac{3}{2}$ and $\frac{3}{2}$ is in the interval $[1, 3]$, $f(\frac{3}{2})$ is a candidate for an absolute extremum of f over $[1, 3]$. We evaluate $f(\frac{3}{2})$ and find

$$f\left(\tfrac{3}{2}\right) = \tfrac{1}{4}.$$

Step 3. We set up the following table to compare the values found in steps 1 and 2.

x	$f(x)$	Conclusion
1	0	
$\frac{3}{2}$	$\frac{1}{4}$	Absolute maximum
3	-2	Absolute minimum

From the table, we find that the absolute maximum of f over the interval [1, 3] is $\frac{1}{4}$, and it occurs at $x = \frac{3}{2}$. The absolute minimum of f over the interval $[1, 3]$ is -2, and it occurs at $x = 3$ as shown in Figure 4.1.8.

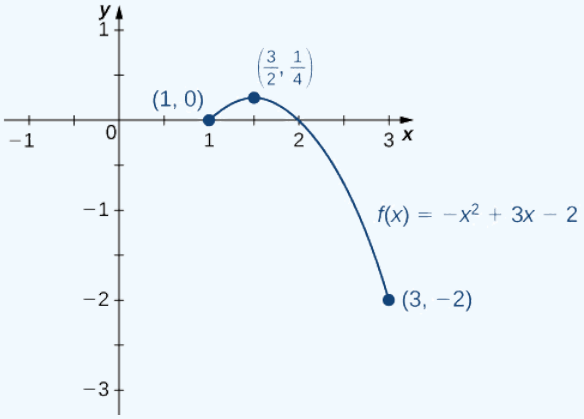


Figure 4.1.8: This function has both an absolute maximum and an absolute minimum.

b. Step 1. Evaluate f at the endpoints $x = 0$ and $x = 2$.

$$f(0) = 0 \text{ and } f(2) = 4 - 3(2)^{2/3} \approx -0.762$$

Step 2. The derivative of f is given by

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

for $x \neq 0$. The derivative is zero when $2x^{4/3} - 2 = 0$, which implies $x = \pm 1$. The derivative is undefined at $x = 0$. Therefore, the critical points of f are $x = 0, 1, -1$. The point $x = 0$ is an endpoint, so we already evaluated $f(0)$ in step 1. The point $x = -1$ is not in the interval of interest, so we need only evaluate $f(1)$. We find that

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

Step 3. We compare the values found in steps 1 and 2, in the following table.

x	$f(x)$	Conclusion
0	0	Absolute maximum
1	-2	Absolute minimum
2	-0.762	

We conclude that the absolute maximum of f over the interval $[0, 2]$ is zero, and it occurs at $x = 0$. The absolute minimum is -2, and it occurs at $x = 1$ as shown in Figure 4.1.9.

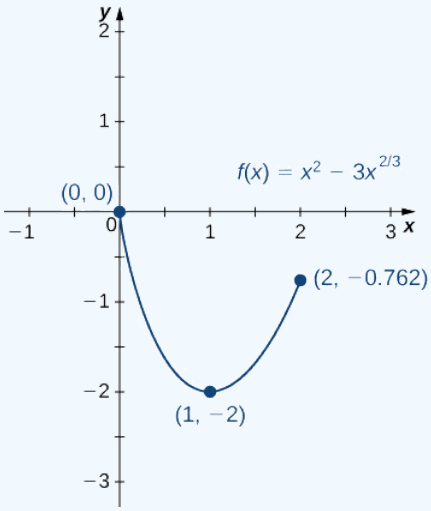


Figure 4.1.9: This function has an absolute maximum at an endpoint of the interval.

? Exercise 4.1.2

Find the absolute maximum and absolute minimum of $f(x) = x^2 - 4x + 3$ over the interval $[1, 4]$.

Hint

Look for critical points. Evaluate f at all critical points and at the endpoints.

Answer

The absolute maximum is 3 and it occurs at $x = 4$. The absolute minimum is -1 and it occurs at $x = 2$.

At this point, we know how to locate absolute extrema for continuous functions over closed intervals. We have also defined local extrema and determined that if a function f has a local extremum at a point c, then c must be a critical point of f. However, c being a critical point is not a sufficient condition for f to have a local extremum at c. Later in this chapter, we show how to determine whether a function actually has a local extremum at a critical point. First, however, we need to introduce the Mean Value Theorem, which will help as we analyze the behavior of the graph of a function.

Key Concepts

- A function may have both an absolute maximum and an absolute minimum, have just one absolute extremum, or have no absolute maximum or absolute minimum.
- If a function has a local extremum, the point at which it occurs must be a critical point. However, a function need not have a local extremum at a critical point.
- A continuous function over a closed, bounded interval has an absolute maximum and an absolute minimum. Each extremum occurs at a critical point or an endpoint.

Glossary

absolute extremum

if f has an absolute maximum or absolute minimum at c, we say f has an absolute extremum at c

absolute maximum

if $f(c) \geq f(x)$ for all x in the domain of f, we say f has an absolute maximum at c

absolute minimum

if $f(c) \leq f(x)$ for all x in the domain of f, we say f has an absolute minimum at c

critical point

if $f'(c) = 0$ or $f'(c)$ is undefined, we say that c is a critical point of f

extreme value theorem

if f is a continuous function over a finite, closed interval, then f has an absolute maximum and an absolute minimum

Fermat's theorem

if f has a local extremum at c, then c is a critical point of f

local extremum

if f has a local maximum or local minimum at c, we say f has a local extremum at c

local maximum

if there exists an interval I such that $f(c) \geq f(x)$ for all $x \in I$, we say f has a local maximum at c

local minimum

if there exists an interval I such that $f(c) \leq f(x)$ for all $x \in I$, we say f has a local minimum at c

4.1: Maxima and Minima is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **4.3: Maxima and Minima** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

4.2: The Mean Value Theorem

Learning Objectives

- Explain the meaning of Rolle's theorem.
- Describe the significance of the Mean Value Theorem.
- State three important consequences of the Mean Value Theorem.

The *Mean Value Theorem* is one of the most important theorems in calculus. We look at some of its implications at the end of this section. First, let's start with a special case of the Mean Value Theorem, called Rolle's theorem.

Rolle's Theorem

Informally, **Rolle's theorem** states that if the outputs of a differentiable function f are equal at the endpoints of an interval, then there must be an interior point c where $f'(c) = 0$. Figure $4.2.1$ illustrates this theorem.

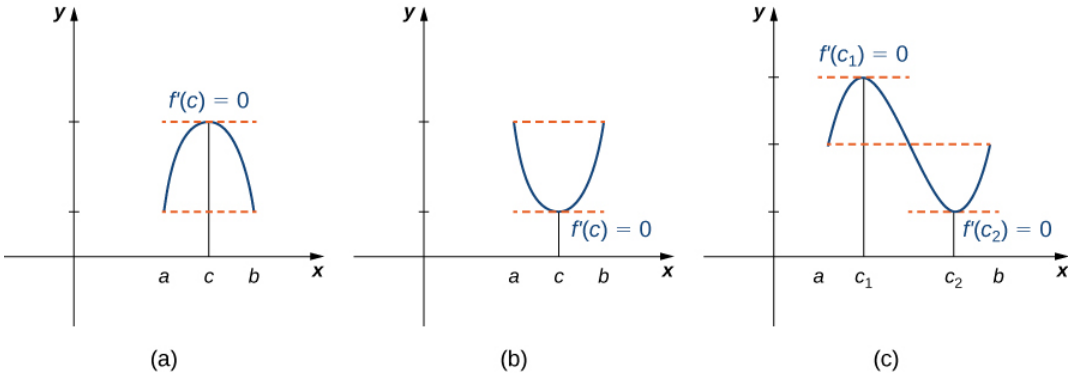


Figure 4.2.1: If a differentiable function f satisfies $f(a) = f(b)$, then its derivative must be zero at some point(s) between a and b.

Rolle's Theorem

Let f be a continuous function over the closed interval $[a, b]$ and differentiable over the open interval (a, b) such that $f(a) = f(b)$. There then exists at least one $c \in (a, b)$ such that $f'(c) = 0$.

Proof

Let $k = f(a) = f(b)$. We consider three cases:

1. $f(x) = k$ for all $x \in (a, b)$.
2. There exists $x \in (a, b)$ such that $f(x) > k$.
3. There exists $x \in (a, b)$ such that $f(x) < k$.

Case 1: If $f(x) = k$ for all $x \in (a, b)$, then $f'(x) = 0$ for all $x \in (a, b)$.

Case 2: Since f is a continuous function over the closed, bounded interval $[a, b]$, by the extreme value theorem, it has an absolute maximum. Also, since there is a point $x \in (a, b)$ such that $f(x) > k$, the absolute maximum is greater than k. Therefore, the absolute maximum does not occur at either endpoint. As a result, the absolute maximum must occur at an interior point $c \in (a, b)$. Because f has a maximum at an interior point c, and f is differentiable at c, by Fermat's theorem, $f'(c) = 0$.

Case 3: The case when there exists a point $x \in (a, b)$ such that $f(x) < k$ is analogous to case 2, with maximum replaced by minimum.

□

An important point about Rolle's theorem is that the differentiability of the function f is critical. If f is not differentiable, even at a single point, the result may not hold. For example, the function $f(x) = |x| - 1$ is continuous over $[-1, 1]$ and $f(-1) = 0 = f(1)$, but $f'(c) \neq 0$ for any $c \in (-1, 1)$ as shown in the following figure.

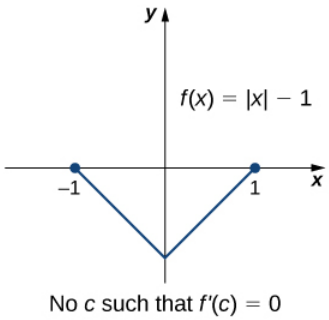


Figure 4.2.2: Since $f(x) = |x| - 1$ is not differentiable at $x = 0$, the conditions of Rolle's theorem are not satisfied. In fact, the conclusion does not hold here; there is no $c \in (-1, 1)$ such that $f'(c) = 0$.

Let's now consider functions that satisfy the conditions of Rolle's theorem and calculate explicitly the points c where $f'(c) = 0$.

Example 4.2.1: Using Rolle's Theorem

For each of the following functions, verify that the function satisfies the criteria stated in Rolle's theorem and find all values c in the given interval where $f'(c) = 0$.

a. $f(x) = x^2 + 2x$ over $[-2, 0]$
b. $f(x) = x^3 - 4x$ over $[-2, 2]$

Solution

a. Since f is a polynomial, it is continuous and differentiable everywhere. In addition, $f(-2) = 0 = f(0)$. Therefore, f satisfies the criteria of Rolle's theorem. We conclude that there exists at least one value $c \in (-2, 0)$ such that $f'(c) = 0$. Since $f'(x) = 2x + 2 = 2(x + 1)$, we see that $f'(c) = 2(c + 1) = 0$ implies $c = -1$ as shown in the following graph.

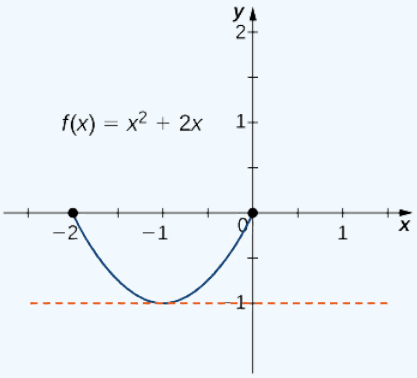


Figure 4.2.3: This function is continuous and differentiable over [−2,0], $f'(c) = 0$ when $c = -1$.

b. As in part a. f is a polynomial and therefore is continuous and differentiable everywhere. Also, $f(-2) = 0 = f(2)$. That said, f satisfies the criteria of Rolle's theorem. Differentiating, we find that $f'(x) = 3x^2 - 4$. Therefore, $f'(c) = 0$ when $x = \pm\frac{2}{\sqrt{3}}$. Both points are in the interval $[-2, 2]$, and, therefore, both points satisfy the conclusion of Rolle's theorem as shown in the following graph.

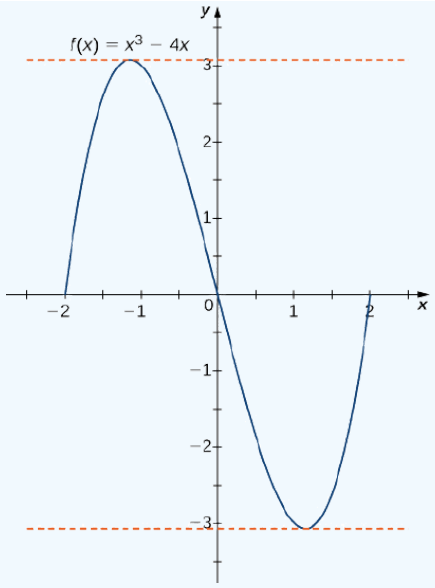


Figure 4.2.4: For this polynomial over $[-2, 2]$, $f'(c) = 0$ at $x = \pm 2/\sqrt{3}$.

Exercise 4.2.1

Verify that the function $f(x) = 2x^2 - 8x + 6$ defined over the interval $[1, 3]$ satisfies the conditions of Rolle's theorem. Find all points c guaranteed by Rolle's theorem.

Hint

Find all values c, where $f'(c) = 0$.

Answer

$c = 2$

The Mean Value Theorem and Its Meaning

Rolle's theorem is a special case of the Mean Value Theorem. In Rolle's theorem, we consider differentiable functions f that are zero at the endpoints. The Mean Value Theorem generalizes Rolle's theorem by considering functions that are not necessarily zero at the endpoints. Consequently, we can view the Mean Value Theorem as a slanted version of Rolle's theorem (Figure 4.2.5). The Mean Value Theorem states that if f is continuous over the closed interval $[a, b]$ and differentiable over the open interval (a, b), then there exists a point $c \in (a, b)$ such that the tangent line to the graph of f at c is parallel to the secant line connecting $(a, f(a))$ and $(b, f(b))$.

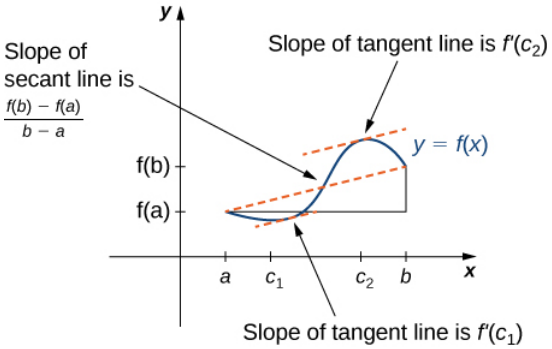


Figure 4.2.5: The Mean Value Theorem says that for a function that meets its conditions, at some point the tangent line has the same slope as the secant line between the ends. For this function, there are two values c_1 and c_2 such that the tangent line to f at c_1 and c_2 has the same slope as the secant line.

Mean Value Theorem

Let f be continuous over the closed interval $[a, b]$ and differentiable over the open interval (a, b). Then, there exists at least one point $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

Proof

The proof follows from Rolle's theorem by introducing an appropriate function that satisfies the criteria of Rolle's theorem. Consider the line connecting $(a, f(a))$ and $(b, f(b))$. Since the slope of that line is

$$\frac{f(b) - f(a)}{b - a}$$

and the line passes through the point $(a, f(a))$, the equation of that line can be written as

$$y = \frac{f(b) - f(a)}{b - a}(x - a) + f(a).$$

Let $g(x)$ denote the vertical difference between the point $(x, f(x))$ and the point (x, y) on that line. Therefore,

$$g(x) = f(x) - \left[\frac{f(b) - f(a)}{b - a}(x - a) + f(a)\right].$$

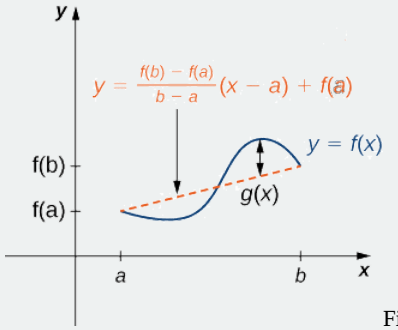


Figure 4.2.6: The value $g(x)$ is the vertical difference between the point $(x, f(x))$ and the point (x, y) on the secant line connecting $(a, f(a))$ and $(b, f(b))$.

Since the graph of f intersects the secant line when $x = a$ and $x = b$, we see that $g(a) = 0 = g(b)$. Since f is a differentiable function over (a, b), g is also a differentiable function over (a, b). Furthermore, since f is continuous over $[a, b]$, g is also continuous over $[a, b]$. Therefore, g satisfies the criteria of Rolle's theorem. Consequently, there exists a point $c \in (a, b)$ such that $g'(c) = 0$. Since

$$g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a},$$

we see that

$$g'(c) = f'(c) - \frac{f(b) - f(a)}{b - a}.$$

Since $g'(c) = 0$, we conclude that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

□

In the next example, we show how the Mean Value Theorem can be applied to the function $f(x) = \sqrt{x}$ over the interval $[0, 9]$. The method is the same for other functions, although sometimes with more interesting consequences.

Example 4.2.2: Verifying that the Mean Value Theorem Applies

For $f(x)=\sqrt{x}$ over the interval $[0,9]$, show that f satisfies the hypothesis of the Mean Value Theorem, and therefore there exists at least one value $c\in(0,9)$ such that $f'(c)$ is equal to the slope of the line connecting $(0,f(0))$ and $(9,f(9))$. Find these values c guaranteed by the Mean Value Theorem.

Solution

We know that $f(x)=\sqrt{x}$ is continuous over $[0,9]$ and differentiable over $(0,9)$. Therefore, f satisfies the hypotheses of the Mean Value Theorem, and there must exist at least one value $c\in(0,9)$ such that $f'(c)$ is equal to the slope of the line connecting $(0,f(0))$ and $(9,f(9))$ (Figure 4.2.7). To determine which value(s) of c are guaranteed, first calculate the derivative of f. The derivative $f'(x)=\frac{1}{(2\sqrt{x})}$. The slope of the line connecting $(0,f(0))$ and $(9,f(9))$ is given by

$$\frac{f(9)-f(0)}{9-0}=\frac{\sqrt{9}-\sqrt{0}}{9-0}=\frac{3}{9}=\frac{1}{3}.$$

We want to find c such that $f'(c)=\frac{1}{3}$. That is, we want to find c such that

$$\frac{1}{2\sqrt{c}}=\frac{1}{3}.$$

Solving this equation for c, we obtain $c=\frac{9}{4}$. At this point, the slope of the tangent line equals the slope of the line joining the endpoints.

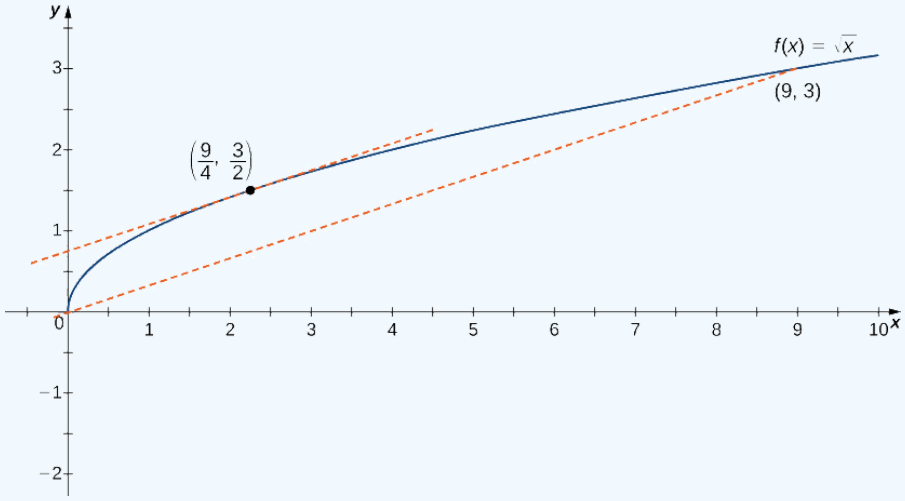


Figure 4.2.7: The slope of the tangent line at $c=9/4$ is the same as the slope of the line segment connecting (0,0) and (9,3).

One application that helps illustrate the Mean Value Theorem involves velocity. For example, suppose we drive a car for 1 h down a straight road with an average velocity of 45 mph. Let $s(t)$ and $v(t)$ denote the position and velocity of the car, respectively, for $0\le t\le 1$ h. Assuming that the position function $s(t)$ is differentiable, we can apply the Mean Value Theorem to conclude that, at some time $c\in(0,1)$, the speed of the car was exactly

$$v(c)=s'(c)=\frac{s(1)-s(0)}{1-0}=45 \text{ mph}.$$

Example 4.2.3: Mean Value Theorem and Velocity

If a rock is dropped from a height of 100 ft, its position t seconds after it is dropped until it hits the ground is given by the function $s(t)=-16t^2+100$.

a. Determine how long it takes before the rock hits the ground.
b. Find the average velocity v_{avg} of the rock for when the rock is released and the rock hits the ground.
c. Find the time t guaranteed by the Mean Value Theorem when the instantaneous velocity of the rock is v_{avg}.

Solution

a. When the rock hits the ground, its position is $s(t)=0$. Solving the equation $-16t^2+100=0$ for t, we find that $t=\pm\frac{5}{2}\,sec$. Since we are only considering $t\geq 0$, the ball will hit the ground $\frac{5}{2}$ sec after it is dropped.

b. The average velocity is given by

$$v_{avg}=\frac{s(5/2)-s(0)}{5/2-0}=\frac{0-100}{5/2}=-40\text{ ft/sec}.$$

c. The instantaneous velocity is given by the derivative of the position function. Therefore, we need to find a time t such that $v(t)=s'(t)=v_{avg}=-40$ ft/sec. Since $s(t)$ is continuous over the interval $[0,5/2]$ and differentiable over the interval $(0,5/2)$, by the Mean Value Theorem, there is guaranteed to be a point $c\in(0,5/2)$ such that

$$s'(c)=\frac{s(5/2)-s(0)}{5/2-0}=-40.$$

Taking the derivative of the position function $s(t)$, we find that $s'(t)=-32t$. Therefore, the equation reduces to $s'(c)=-32c=-40$. Solving this equation for c, we have $c=\frac{5}{4}$. Therefore, $\frac{5}{4}$ sec after the rock is dropped, the instantaneous velocity equals the average velocity of the rock during its free fall: -40 ft/sec.

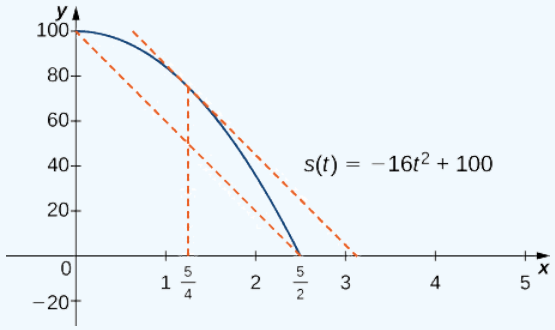


Figure 4.2.8: At time $t=5/4$ sec, the velocity of the rock is equal to its average velocity from the time it is dropped until it hits the ground.

Exercise 4.2.2

Suppose a ball is dropped from a height of 200 ft. Its position at time t is $s(t)=-16t^2+200$. Find the time t when the instantaneous velocity of the ball equals its average velocity.

Hint

First, determine how long it takes for the ball to hit the ground. Then, find the average velocity of the ball from the time it is dropped until it hits the ground.

Answer

$\frac{5}{2\sqrt{2}}$ sec

Corollaries of the Mean Value Theorem

Let's now look at three corollaries of the Mean Value Theorem. These results have important consequences, which we use in upcoming sections.

At this point, we know the derivative of any constant function is zero. The Mean Value Theorem allows us to conclude that the converse is also true. In particular, if $f'(x)=0$ for all x in some interval I, then $f(x)$ is constant over that interval. This result may seem intuitively obvious, but it has important implications that are not obvious, and we discuss them shortly.

Corollary 1: Functions with a Derivative of Zero

Let f be differentiable over an interval I. If $f'(x) = 0$ for all $x \in I$, then $f(x) =$ constant for all $x \in I$.

Proof

Since f is differentiable over I, f must be continuous over I. Suppose $f(x)$ is not constant for all x in I. Then there exist $a, b \in I$, where $a \neq b$ and $f(a) \neq f(b)$. Choose the notation so that $a < b$. Therefore,

$$\frac{f(b) - f(a)}{b - a} \neq 0.$$

Since f is a differentiable function, by the Mean Value Theorem, there exists $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Therefore, there exists $c \in I$ such that $f'(c) \neq 0$, which contradicts the assumption that $f'(x) = 0$ for all $x \in I$.

□

From "Corollary 1: Functions with a Derivative of Zero," it follows that if two functions have the same derivative, they differ by, at most, a constant.

Corollary 2: Constant Difference Theorem

If f and g are differentiable over an interval I and $f'(x) = g'(x)$ for all $x \in I$, then $f(x) = g(x) + C$ for some constant C.

Proof

Let $h(x) = f(x) - g(x)$. Then, $h'(x) = f'(x) - g'(x) = 0$ for all $x \in I$. By Corollary 1, there is a constant C such that $h(x) = C$ for all $x \in I$. Therefore, $f(x) = g(x) + C$ for all $x \in I$.

□

The third corollary of the Mean Value Theorem discusses when a function is increasing and when it is decreasing. Recall that a function f is increasing over I if $f(x_1) < f(x_2)$ whenever $x_1 < x_2$, whereas f is decreasing over I if $f(x_1) > f(x_2)$ whenever $x_1 < x_2$. Using the Mean Value Theorem, we can show that if the derivative of a function is positive, then the function is increasing; if the derivative is negative, then the function is decreasing (Figure 4.2.9). We make use of this fact in the next section, where we show how to use the derivative of a function to locate local maximum and minimum values of the function, and how to determine the shape of the graph.

This fact is important because it means that for a given function f, if there exists a function F such that $F'(x) = f(x)$; then, the only other functions that have a derivative equal to f are $F(x) + C$ for some constant C. We discuss this result in more detail later in the chapter.

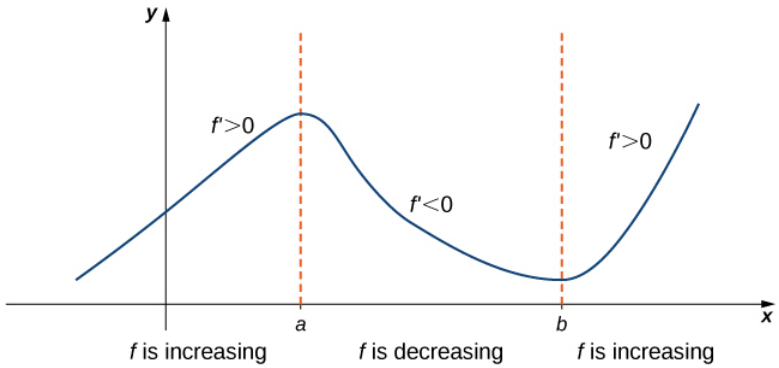


Figure 4.2.9: If a function has a positive derivative over some interval I, then the function increases over that interval I; if the derivative is negative over some interval I, then the function decreases over that interval I.

Corollary 3: Increasing and Decreasing Functions

Let f be continuous over the closed interval $[a, b]$ and differentiable over the open interval (a, b).

i. If $f'(x) > 0$ for all $x \in (a, b)$, then f is an increasing function over $[a, b]$.
ii. If $f'(x) < 0$ for all $x \in (a, b)$, then f is a decreasing function over $[a, b]$.

Proof

We will prove i.; the proof of ii. is similar. Suppose f is not an increasing function on I. Then there exist a and b in I such that $a < b$, but $f(a) \geq f(b)$. Since f is a differentiable function over I, by the Mean Value Theorem there exists $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Since $f(a) \geq f(b)$, we know that $f(b) - f(a) \leq 0$. Also, $a < b$ tells us that $b - a > 0$. We conclude that

$$f'(c) = \frac{f(b) - f(a)}{b - a} \leq 0.$$

However, $f'(x) > 0$ for all $x \in I$. This is a contradiction, and therefore f must be an increasing function over I.

□

Key Concepts

- If f is continuous over $[a, b]$ and differentiable over (a, b) and $f(a) = f(b)$, then there exists a point $c \in (a, b)$ such that $f'(c) = 0$. This is Rolle's theorem.
- If f is continuous over $[a, b]$ and differentiable over (a, b), then there exists a point $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

  This is the Mean Value Theorem.
- If $f'(x) = 0$ over an interval I, then f is constant over I.
- If two differentiable functions f and g satisfy $f'(x) = g'(x)$ over I, then $f(x) = g(x) + C$ for some constant C.
- If $f'(x) > 0$ over an interval I, then f is increasing over I. If $f'(x) < 0$ over I, then f is decreasing over I.

Glossary

mean value theorem

if f is continuous over $[a, b]$ and differentiable over (a, b), then there exists $c \in (a, b)$ such that $f'(c) = \frac{f(b) - f(a)}{b - a}$

rolle's theorem

if f is continuous over $[a, b]$ and differentiable over (a, b), and if $f(a) = f(b)$, then there exists $c \in (a, b)$ such that $f'(c) = 0$

4.2: The Mean Value Theorem is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **4.4: The Mean Value Theorem** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

4.3: The First Derivative Test

Learning Objectives

- Explain how the sign of the first derivative affects the shape of a function's graph.
- State the first derivative test for critical points.
- Find local extrema using the First Derivative Test.

Earlier in this chapter we stated that if a function f has a local extremum at a point c, then c must be a critical point of f. However, a function is not guaranteed to have a local extremum at a critical point. For example, $f(x) = x^3$ has a critical point at $x = 0$ since $f'(x) = 3x^2$ is zero at $x = 0$, but f does not have a local extremum at $x = 0$. Using the results from the previous section, we are now able to determine whether a critical point of a function actually corresponds to a local extreme value. In this section, we also see how the second derivative provides information about the shape of a graph by describing whether the graph of a function curves upward or curves downward.

The First Derivative Test

Corollary 3 of the Mean Value Theorem showed that if the derivative of a function is positive over an interval I then the function is increasing over I. On the other hand, if the derivative of the function is negative over an interval I, then the function is decreasing over I as shown in the following figure.

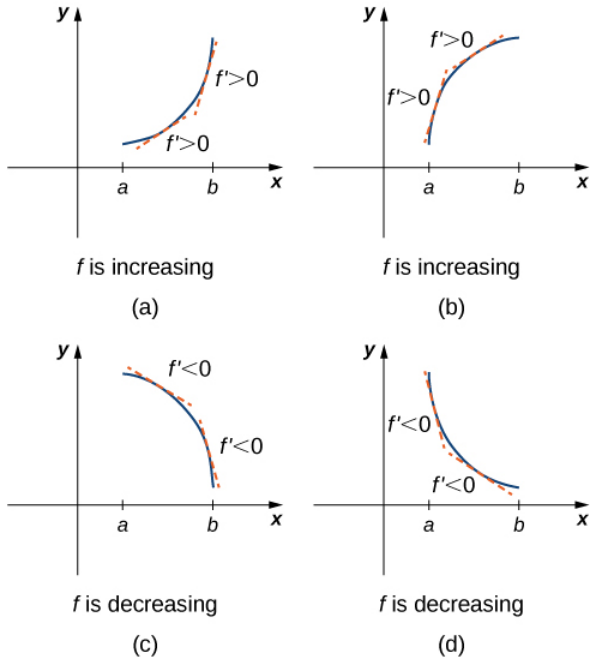


Figure 4.3.1: Both functions are increasing over the interval (a, b). At each point x, the derivative $f'(x) > 0$. Both functions are decreasing over the interval (a, b). At each point x, the derivative $f'(x) < 0$.

A continuous function f has a local maximum at point c if and only if f switches from increasing to decreasing at point c. Similarly, f has a local minimum at c if and only if f switches from decreasing to increasing at c. If f is a continuous function over an interval I containing c and differentiable over I, except possibly at c, the only way f can switch from increasing to decreasing (or vice versa) at point c is if f' changes sign as x increases through c. If f is differentiable at c, the only way that f' can change sign as x increases through c is if $f'(c) = 0$. Therefore, for a function f that is continuous over an interval I containing c and differentiable over I, except possibly at c, the only way f can switch from increasing to decreasing (or vice versa) is if $f'(c) = 0$ or $f'(c)$ is undefined. Consequently, to locate local extrema for a function f, we look for points c in the domain of f such that $f'(c) = 0$ or $f'(c)$ is undefined. Recall that such points are called critical points of f.

Note that f need not have a local extrema at a critical point. The critical points are candidates for local extrema only. In Figure 4.3.2, we show that if a continuous function f has a local extremum, it must occur at a critical point, but a function may not have a

local extremum at a critical point. We show that if f has a local extremum at a critical point, then the sign of f' switches as x increases through that point.

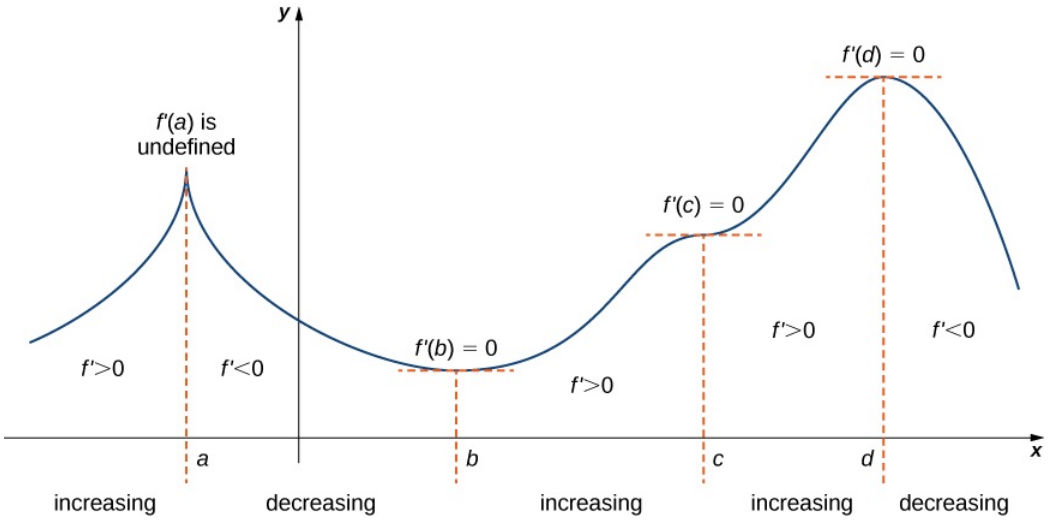


Figure 4.3.2: The function f has four critical points: a, b, c ,and d. The function f has local maxima at a and d, and a local minimum at b. The function f does not have a local extremum at c. The sign of f' changes at all local extrema.

Using Figure 4.3.2, we summarize the main results regarding local extrema.

- If a continuous function f has a local extremum, it must occur at a critical point c.
- The function has a local extremum at the critical point c if and only if the derivative f' switches sign as x increases through c.
- Therefore, to test whether a function has a local extremum at a critical point c, we must determine the sign of $f'(x)$ to the left and right of c.

This result is known as the **first derivative test.**

First Derivative Test

Suppose that f is a continuous function over an interval I containing a critical point c. If f is differentiable over I, except possibly at point c, then $f(c)$ satisfies one of the following descriptions:

i. If f' changes sign from positive when $x < c$ to negative when $x > c$, then $f(c)$ is a local maximum of f.
ii. If f' changes sign from negative when $x < c$ to positive when $x > c$, then $f(c)$ is a local minimum of f.
iii. If f' has the same sign for $x < c$ and $x > c$, then $f(c)$ is neither a local maximum nor a local minimum of f

Now let's look at how to use this strategy to locate all local extrema for particular functions.

Example 4.3.1: Using the First Derivative Test to Find Local Extrema

Use the first derivative test to find the location of all local extrema for $f(x) = x^3 - 3x^2 - 9x - 1$. Use a graphing utility to confirm your results.

Solution

Step 1. The derivative is $f'(x) = 3x^2 - 6x - 9$. To find the critical points, we need to find where $f'(x) = 0$. Factoring the polynomial, we conclude that the critical points must satisfy

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

Therefore, the critical points are $x = 3, -1$. Now divide the interval $(-\infty, \infty)$ into the smaller intervals $(-\infty, -1), (-1, 3)$ and $(3, \infty)$.

Step 2. Since f' is a continuous function, to determine the sign of $f'(x)$ over each subinterval, it suffices to choose a point over each of the intervals $(-\infty, -1), (-1, 3)$ and $(3, \infty)$ and determine the sign of f' at each of these points. For example, let's choose $x = -2$, $x = 0$, and $x = 4$ as test points.

Table: 4.3.1: First Derivative Test for $f(x) = x^3 - 3x^2 - 9x - 1$.

Interval	Test Point	Sign of $f'(x) = 3(x-3)(x+1)$ at Test Point	Conclusion
$(-\infty, -1)$	$x = -2$	(+)(−)(−)=+	f is increasing.
$(-1, 3)$	$x = 0$	(+)(−)(+)=-	f is decreasing.
$(3, \infty)$	$x = 4$	(+)(+)(+)=+	f is increasing.

Step 3. Since f' switches sign from positive to negative as x increases through -1, f has a local maximum at $x = -1$. Since f' switches sign from negative to positive as x increases through 3, f has a local minimum at $x = 3$. These analytical results agree with the following graph.

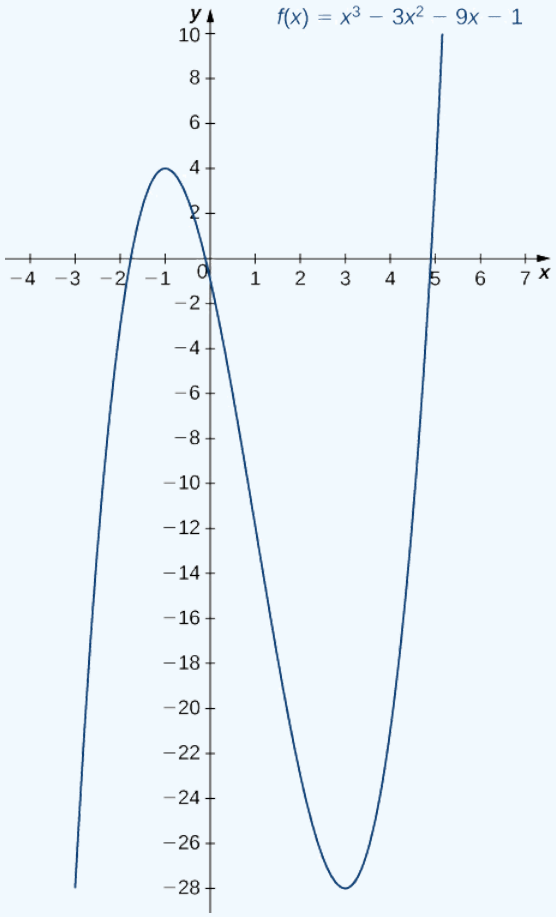


Figure 4.3.3: The function f has a maximum at $x = -1$ and a minimum at $x = 3$

Exercise 4.3.1

Use the first derivative test to locate all local extrema for $f(x) = -x^3 + \frac{3}{2}x^2 + 18x$.

Hint

Find all critical points of f and determine the signs of $f'(x)$ over particular intervals determined by the critical points.

Answer

f has a local minimum at -2 and a local maximum at 3.

Example 4.3.2: Using the First Derivative Test

Use the first derivative test to find the location of all local extrema for $f(x) = 5x^{1/3} - x^{5/3}$. Use a graphing utility to confirm your results.

Solution

Step 1. The derivative is

$$f'(x) = \frac{5}{3}x^{-2/3} - \frac{5}{3}x^{2/3} = \frac{5}{3x^{2/3}} - \frac{5x^{2/3}}{3} = \frac{5 - 5x^{4/3}}{3x^{2/3}} = \frac{5(1 - x^{4/3})}{3x^{2/3}}.$$

The derivative $f'(x) = 0$ when $1 - x^{4/3} = 0$. Therefore, $f'(x) = 0$ at $x = \pm 1$. The derivative $f'(x)$ is undefined at $x = 0$. Therefore, we have three critical points: $x = 0$, $x = 1$, and $x = -1$. Consequently, divide the interval $(-\infty, \infty)$ into the smaller intervals $(-\infty, -1)$, $(-1, 0)$, $(0, 1)$, and $(1, \infty)$.

Step 2: Since f' is continuous over each subinterval, it suffices to choose a test point x in each of the intervals from step 1 and determine the sign of f' at each of these points. The points $x = -2$, $x = -\frac{1}{2}$, $x = \frac{1}{2}$, and $x = 2$ are test points for these intervals.

Table: 4.3.2: First Derivative Test for $f(x) = 5x^{1/3} - x^{5/3}$.

Interval	Test Point	Sign of $f'(x) = \frac{5(1-x^{4/3})}{3x^{2/3}}$ at Test Point	Conclusion
$(-\infty, -1)$	$x = -2$	$\frac{(+)(-)}{+} = -$	f is decreasing.
$(-1, 0)$	$x = -\frac{1}{2}$	$\frac{(+)(+)}{+} = +$	f is increasing.
$(0, 1)$	$x = \frac{1}{2}$	$\frac{(+)(+)}{+} = +$	f is increasing.
$(1, \infty)$	$x = 2$	$\frac{(+)(-)}{+} = -$	f is decreasing.

Step 3: Since f is decreasing over the interval $(-\infty, -1)$ and increasing over the interval $(-1, 0)$, f has a local minimum at $x = -1$. Since f is increasing over the interval $(-1, 0)$ and the interval $(0, 1)$, f does not have a local extremum at $x = 0$. Since f is increasing over the interval $(0, 1)$ and decreasing over the interval $(1, \infty)$, f has a local maximum at $x = 1$. The analytical results agree with the following graph.

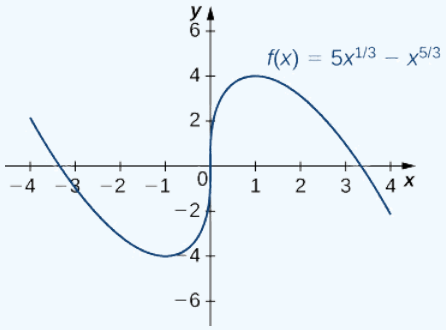


Figure 4.3.4: The function f has a local minimum at $x = -1$ and a local maximum at $x = 1$

Exercise 4.3.2

Use the first derivative test to find all local extrema for $f(x) = \dfrac{3}{x - 1}$.

Hint

The only critical point of f is $x = 1$.

Answer

f has no local extrema because f' does not change sign at $x = 1$.

This page titled 4.3: The First Derivative Test is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **4.5: Derivatives and the Shape of a Graph** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

4.4: The Second Derivative Test and Curve Sketching

Learning Objectives

- Use concavity and inflection points to explain how the sign of the second derivative affects the shape of a function's graph.
- Explain the concavity test for a function over an open interval.
- Explain the relationship between a function and its first and second derivatives.
- State the second derivative test for local extrema.
- Analyze a function and its derivatives to draw its graph.

Concavity and Points of Inflection

We now know how to determine where a function is increasing or decreasing. However, there is another issue to consider regarding the shape of the graph of a function. If the graph curves, does it curve upward or curve downward? This notion is called the **concavity** of the function.

Figure $4.4.5a$ shows a function f with a graph that curves upward. As x increases, the slope of the tangent line increases. Thus, since the derivative increases as x increases, f' is an increasing function. We say this function f is concave up. Figure $4.4.5b$ shows a function f that curves downward. As x increases, the slope of the tangent line decreases. Since the derivative decreases as x increases, f' is a decreasing function. We say this function f is concave down.

Definition: concavity test

Let f be a function that is differentiable over an open interval I. If f' is increasing over I, we say f is **concave up** over I. If f' is decreasing over I, we say f is **concave down** over I.

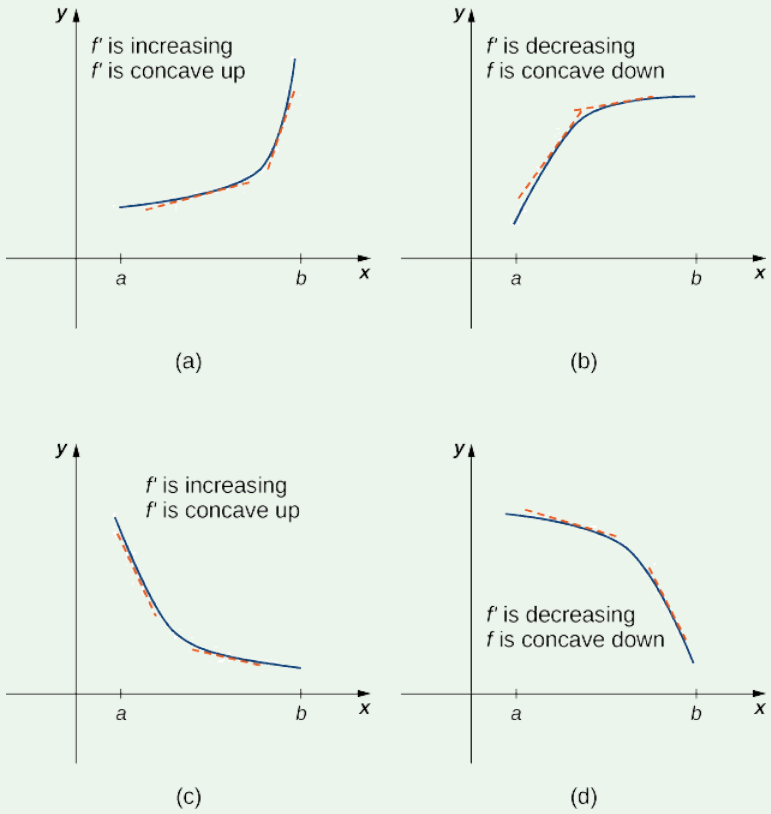


Figure 4.4.5: (a), (c) Since f' is increasing over the interval (a,b), we say f is concave up over (a,b). $(b),(d)$ Since f' is decreasing over the interval (a,b), we say f is concave down over (a,b).

In general, without having the graph of a function f, how can we determine its concavity? By definition, a function f is concave up if f' is increasing. From Corollary 3, we know that if f' is a differentiable function, then f' is increasing if its

derivative $f''(x) > 0$. Therefore, a function f that is twice differentiable is concave up when $f''(x) > 0$. Similarly, a function f is concave down if f' is decreasing. We know that a differentiable function f' is decreasing if its derivative $f''(x) < 0$. Therefore, a twice-differentiable function f is concave down when $f''(x) < 0$. Applying this logic is known as the *concavity test*.

Test for Concavity

Let f be a function that is twice differentiable over an interval I.

i. If $f''(x) > 0$ for all $x \in I$, then f is concave up over I
ii. If $f''(x) < 0$ for all $x \in I$, then f is concave down over I.

We conclude that we can determine the concavity of a function f by looking at the second derivative of f. In addition, we observe that a function f can switch concavity (Figure 4.4.6). However, a continuous function can switch concavity only at a point x if $f''(x) = 0$ or $f''(x)$ is undefined. Consequently, to determine the intervals where a function f is concave up and concave down, we look for those values of x where $f''(x) = 0$ or $f''(x)$ is undefined. When we have determined these points, we divide the domain of f into smaller intervals and determine the sign of f'' over each of these smaller intervals. If f'' changes sign as we pass through a point x, then f changes concavity. It is important to remember that a function f may not change concavity at a point x even if $f''(x) = 0$ or $f''(x)$ is undefined. If, however, f does change concavity at a point a and f is continuous at a, we say the point $(a, f(a))$ is an **inflection point** of f.

Definition: inflection point

If f is continuous at a and f changes concavity at a, the point $(a,\ f(a))$ is an *inflection point* of f.

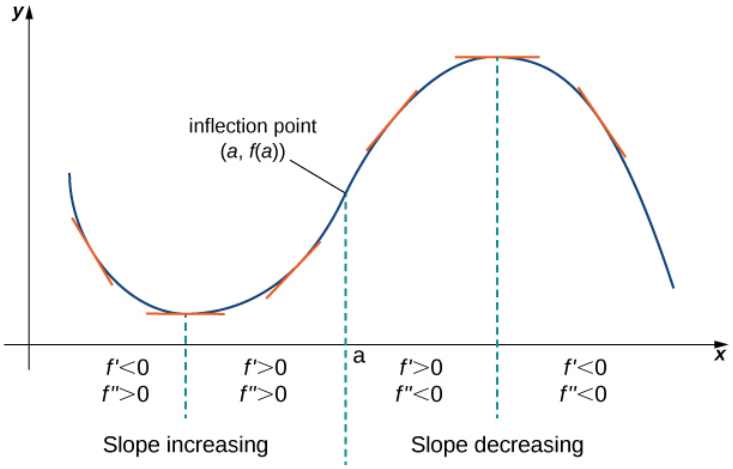


Figure 4.4.6: Since $f''(x) > 0$ for $x < a$, the function f is concave up over the interval $(-\infty, a)$. Since $f''(x) < 0$ for $x > a$, the function f is concave down over the interval (a, ∞). The point $(a, f(a))$ is an inflection point of f.

Example 4.4.3: Testing for Concavity

For the function $f(x) = x^3 - 6x^2 + 9x + 30$, determine all intervals where f is concave up and all intervals where f is concave down. List all inflection points for f. Use a graphing utility to confirm your results.

Solution

To determine concavity, we need to find the second derivative $f''(x)$. The first derivative is $f'(x) = 3x^2 - 12x + 9$, so the second derivative is $f''(x) = 6x - 12$. If the function changes concavity, it occurs either when $f''(x) = 0$ or $f''(x)$ is undefined. Since f'' is defined for all real numbers x, we need only find where $f''(x) = 0$. Solving the equation $6x - 12 = 0$, we see that $x = 2$ is the only place where f could change concavity. We now test points over the intervals $(-\infty, 2)$ and $(2, \infty)$ to determine the concavity of f. The points $x = 0$ and $x = 3$ are test points for these intervals.

Table: 4.4.3: Test for Concavity for $f(x) = x^3 - 6x^2 + 9x + 30$.

Interval	Test Point	Sign of $f''(x) = 6x - 12$ at Test Point	Conclusion
$(-\infty, 2)$	$x = 0$	$-$	f is concave down
$(2, \infty)$	$x = 3$	$+$	f is concave up

We conclude that f is concave down over the interval $(-\infty, 2)$ and concave up over the interval $(2, \infty)$. Since f changes concavity at $x = 2$, the point $(2, f(2)) = (2, 32)$ is an inflection point. Figure 4.4.7 confirms the analytical results.

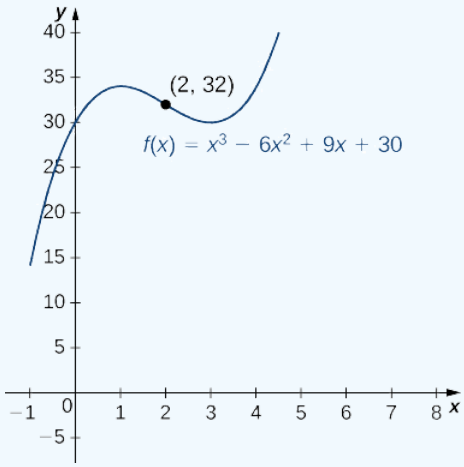


Figure 4.4.7: The given function has a point of inflection at $(2, 32)$ where the graph changes concavity.

Exercise 4.4.3

For $f(x) = -x^3 + \frac{3}{2}x^2 + 18x$, find all intervals where f is concave up and all intervals where f is concave down.

Hint

Find where $f''(x) = 0$

Answer

f is concave up over the interval $(-\infty, \frac{1}{2})$ and concave down over the interval $(\frac{1}{2}, \infty)$

We now summarize, in Table 4.4.4, the information that the first and second derivatives of a function f provide about the graph of f, and illustrate this information in Figure 4.4.8.

Table: 4.4.4: What Derivatives Tell Us about Graphs

Sign of f'	Sign of f''	Is f increasing or decreasing?	Concavity
Positive	Positive	Increasing	Concave up
Positive	Negative	Increasing	Concave down
Negative	Positive	Decreasing	Concave up
Negative	Negative	Decreasing	Concave down

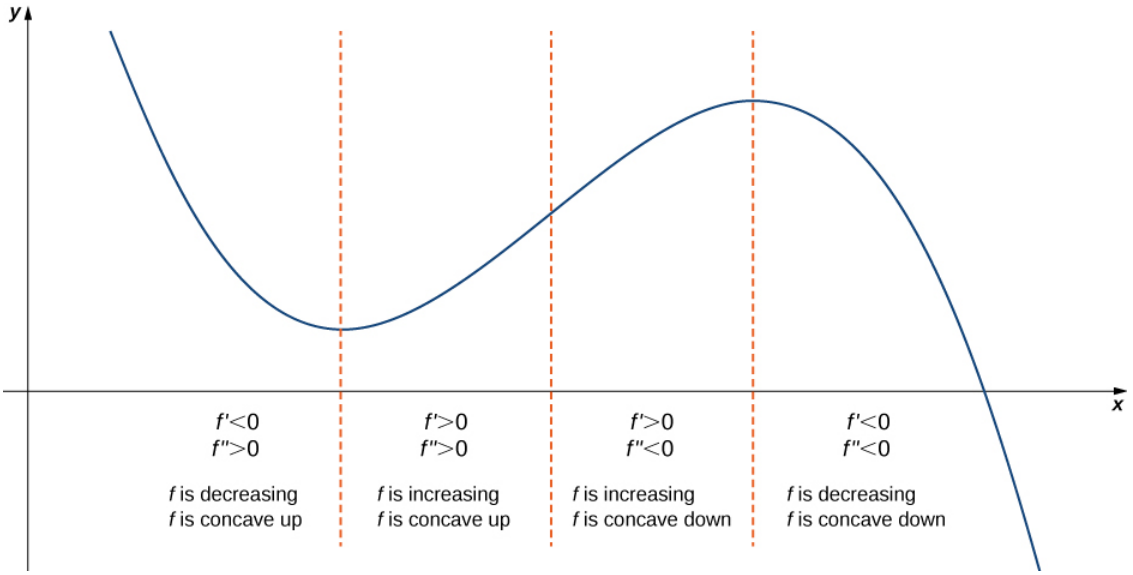


Figure 4.4.8:Consider a twice-differentiable function f over an open interval I. If $f'(x) > 0$ for all $x \in I$, the function is increasing over I. If $f'(x) < 0$ for all $x \in I$, the function is decreasing over I. If $f''(x) > 0$ for all $x \in I$, the function is concave up. If $f''(x) < 0$ for all $x \in I$, the function is concave down on I.

The Second Derivative Test

The first derivative test provides an analytical tool for finding local extrema, but the second derivative can also be used to locate extreme values. Using the second derivative can sometimes be a simpler method than using the first derivative.

We know that if a continuous function has a local extremum, it must occur at a critical point. However, a function need not have a local extremum at a critical point. Here we examine how the **second derivative test** can be used to determine whether a function has a local extremum at a critical point. Let f be a twice-differentiable function such that $f'(a) = 0$ and f'' is continuous over an open interval I containing a. Suppose $f''(a) < 0$. Since f'' is continuous over I, $f''(x) < 0$ for all $x \in I$ (Figure 4.4.9). Then, by Corollary 3, f' is a decreasing function over I. Since $f'(a) = 0$, we conclude that for all $x \in I$, $f'(x) > 0$ if $x < a$ and $f'(x) < 0$ if $x > a$. Therefore, by the first derivative test, f has a local maximum at $x = a$.

On the other hand, suppose there exists a point b such that $f'(b) = 0$ but $f''(b) > 0$. Since f'' is continuous over an open interval I containing b, then $f''(x) > 0$ for all $x \in I$ (Figure 4.4.9). Then, by Corollary 3, f' is an increasing function over I. Since $f'(b) = 0$, we conclude that for all $x \in I$, $f'(x) < 0$ if $x < b$ and $f'(x) > 0$ if $x > b$. Therefore, by the first derivative test, f has a local minimum at $x = b$.

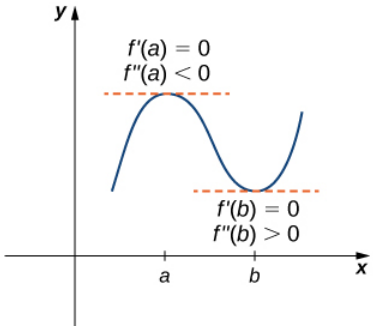


Figure 4.4.9: Consider a twice-differentiable function f such that f'' is continuous. Since $f'(a) = 0$ and $f''(a) < 0$, there is an interval I containing a such that for all x in I, f is increasing if $x < a$ and f is decreasing if $x > a$. As a result, f has a local maximum at $x = a$. Since $f'(b) = 0$ and $f''(b) > 0$, there is an interval I containing b such that for all x in I, f is decreasing if $x < b$ and f is increasing if $x > b$. As a result, f has a local minimum at $x = b$.

Second Derivative Test

Suppose $f'(c) = 0$ and f'' is continuous over an interval containing c.

i. If $f''(c) > 0$, then f has a local minimum at c.
ii. If $f''(c) < 0$, then f has a local maximum at c.
iii. If $f''(c) = 0$, then the test is inconclusive.

Note that for case iii. when $f''(c) = 0$, then f may have a local maximum, local minimum, or neither at c. For example, the functions $f(x) = x^3$, $f(x) = x^4$, and $f(x) = -x^4$ all have critical points at $x = 0$. In each case, the second derivative is zero at $x = 0$. However, the function $f(x) = x^4$ has a local minimum at $x = 0$ whereas the function $f(x) = -x^4$ has a local maximum at $x = 0$, and the function $f(x) = x^3$ does not have a local extremum at $x = 0$.

Let's now look at how to use the second derivative test to determine whether f has a local maximum or local minimum at a critical point c where $f'(c) = 0$.

Example 4.4.4: Using the Second Derivative Test

Use the second derivative to find the location of all local extrema for $f(x) = x^5 - 5x^3$.

Solution

To apply the second derivative test, we first need to find critical points c where $f'(c) = 0$. The derivative is $f'(x) = 5x^4 - 15x^2$. Therefore, $f'(x) = 5x^4 - 15x^2 = 5x^2(x^2 - 3) = 0$ when $x = 0, \pm\sqrt{3}$.

To determine whether f has a local extremum at any of these points, we need to evaluate the sign of f'' at these points. The second derivative is

$$f''(x) = 20x^3 - 30x = 10x(2x^2 - 3).$$

In the following table, we evaluate the second derivative at each of the critical points and use the second derivative test to determine whether f has a local maximum or local minimum at any of these points.

Table: 4.4.5: Second Derivative Test for $f(x) = x^5 - 5x^3$.

x	$f''(x)$	Conclusion
$-\sqrt{3}$	$-30\sqrt{3}$	Local maximum
0	0	Second derivative test is inconclusive
$\sqrt{3}$	$30\sqrt{3}$	Local minimum

By the second derivative test, we conclude that f has a local maximum at $x = -\sqrt{3}$ and f has a local minimum at $x = \sqrt{3}$. The second derivative test is inconclusive at $x = 0$. To determine whether f has a local extrema at $x = 0$, we apply the first derivative test. To evaluate the sign of $f'(x) = 5x^2(x^2 - 3)$ for $x \in (-\sqrt{3}, 0)$ and $x \in (0, \sqrt{3})$, let $x = -1$ and $x = 1$ be the two test points. Since $f'(-1) < 0$ and $f'(1) < 0$, we conclude that f is decreasing on both intervals and, therefore, f does not have a local extrema at $x = 0$ as shown in the following graph.

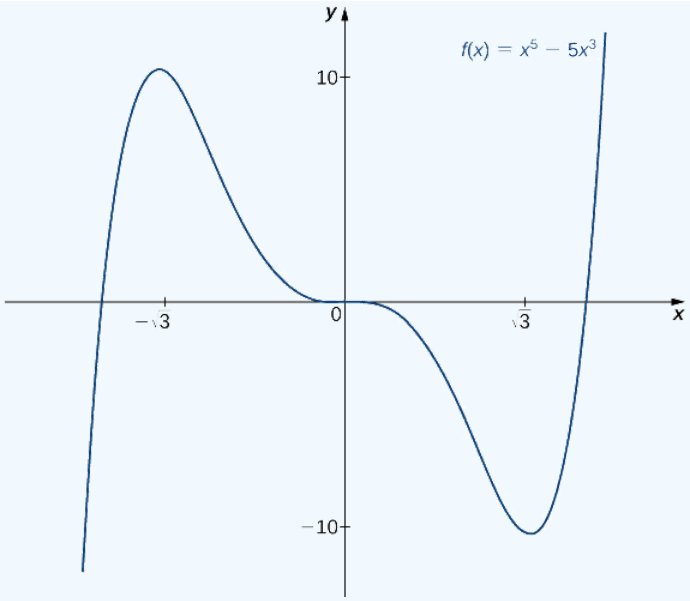


Figure 4.4.10:The function f has a local maximum at $x=-\sqrt{3}$ and a local minimum at $x=\sqrt{3}$

? Exercise 4.4.4

Consider the function $f(x)=x^3-(\frac{3}{2})x^2-18x$. The points $c=3,\ -2$ satisfy $f'(c)=0$. Use the second derivative test to determine whether f has a local maximum or local minimum at those points.

Hint

$f''(x)=6x-3$

Answer

f has a local maximum at -2 and a local minimum at 3.

We have now developed the tools we need to determine where a function is increasing and decreasing, as well as acquired an understanding of the basic shape of the graph. In the next section we discuss what happens to a function as $x\to\pm\infty$. At that point, we have enough tools to provide accurate graphs of a large variety of functions.

Guidelines for Drawing the Graph of a Function

We now have enough analytical tools to draw graphs of a wide variety of algebraic and transcendental functions. Before showing how to graph specific functions, let's look at a general strategy to use when graphing any function.

Problem-Solving Strategy: Drawing the Graph of a Function

Given a function f, use the following steps to sketch a graph of f:

1. Determine the domain of the function.
2. Locate the x- and y-intercepts.
3. Evaluate $\lim\limits_{x\to\infty} f(x)$ and $\lim\limits_{x\to-\infty} f(x)$ to determine the end behavior. If either of these limits is a finite number L, then $y=L$ is a horizontal asymptote. If either of these limits is ∞ or $-\infty$, determine whether f has an oblique asymptote. If f is a rational function such that $f(x)=\dfrac{p(x)}{q(x)}$, where the degree of the numerator is greater than the degree of the denominator, then f can be written as

$$f(x)=\frac{p(x)}{q(x)}=g(x)+\frac{r(x)}{q(x),}$$

where the degree of $r(x)$ is less than the degree of $q(x)$. The values of $f(x)$ approach the values of $g(x)$ as $x \to \pm\infty$. If $g(x)$ is a linear function, it is known as an oblique asymptote.

4. Determine whether f has any vertical asymptotes.
5. Calculate f'. Find all critical points and determine the intervals where f is increasing and where f is decreasing. Determine whether f has any local extrema.
6. Calculate f''. Determine the intervals where f is concave up and where f is concave down. Use this information to determine whether f has any inflection points. The second derivative can also be used as an alternate means to determine or verify that f has a local extremum at a critical point.

Now let's use this strategy to graph several different functions. We start by graphing a polynomial function.

Example 4.4.5: Sketching a Graph of a Polynomial

Sketch a graph of $f(x) = (x-1)^2(x+2)$.

Solution

Step 1: Since f is a polynomial, the domain is the set of all real numbers.

Step 2: When $x = 0$, $f(x) = 2$. Therefore, the y-intercept is $(0, 2)$. To find the x-intercepts, we need to solve the equation $(x-1)^2(x+2) = 0$, which gives us the x-intercepts $(1, 0)$ and $(-2, 0)$

Step 3: We need to evaluate the end behavior of f. As $x \to \infty$, $(x-1)^2 \to \infty$ and $(x+2) \to \infty$. Therefore, $\lim\limits_{x\to\infty} f(x) = \infty$.

As $x \to -\infty$, $(x-1)^2 \to \infty$ and $(x+2) \to -\infty$. Therefore, $\lim\limits_{x\to-\infty} f(x) = -\infty$.

To get even more information about the end behavior of f, we can multiply the factors of f. When doing so, we see that

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

Since the leading term of f is x^3, we conclude that f behaves like $y = x^3$ as $x \to \pm\infty$.

Step 4: Since f is a polynomial function, it does not have any vertical asymptotes.

Step 5: The first derivative of f is

$$f'(x) = 3x^2 - 3.$$

Therefore, f has two critical points: $x = 1, -1$. Divide the interval $(-\infty, \infty)$ into the three smaller intervals: $(-\infty, -1)$, $(-1, 1)$, and $(1, \infty)$. Then, choose test points $x = -2, x = 0$, and $x = 2$ from these intervals and evaluate the sign of $f'(x)$ at each of these test points, as shown in the following table.

Interval	Test point	Sign of Derivative $f'(x) = 3x^2 - 3 = 3(x-1)(x+1)$	Conclusion
$(-\infty, -1)$	$x = -2$	$(+)(-)(-) = +$	f is increasing
$(-1, 1)$	$x = 0$	$(+)(-)(+) = -$	f decreasing
$(1, \infty)$	$x = 2$	$(+)(+)(+) = +$	f is increasing

From the table, we see that f has a local maximum at $x = -1$ and a local minimum at $x = 1$. Evaluating $f(x)$ at those two points, we find that the local maximum value is $f(-1) = 4$ and the local minimum value is $f(1) = 0$.

Step 6: The second derivative of f is

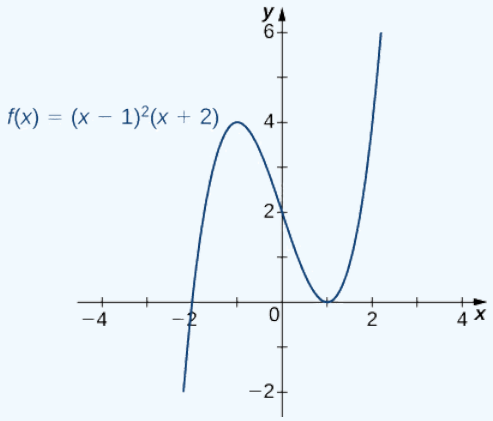
$$f''(x) = 6x.$$

The second derivative is zero at $x = 0$. Therefore, to determine the concavity of f, divide the interval $(-\infty, \infty)$ into the smaller intervals $(-\infty, 0)$ and $(0, \infty)$, and choose test points $x = -1$ and $x = 1$ to determine the concavity of f on each of these smaller intervals as shown in the following table.

Interval	Test Point	Sign of $f''(x) = 6x$	Conclusion
$(-\infty, 0)$	$x = -1$	$-$	f is concave down.
$(0, \infty)$	$x = 1$	$+$	f is concave up.

We note that the information in the preceding table confirms the fact, found in step 5, that f has a local maximum at $x = -1$ and a local minimum at $x = 1$. In addition, the information found in step 5—namely, f has a local maximum at $x = -1$ and a local minimum at $x = 1$, and $f'(x) = 0$ at those points—combined with the fact that f'' changes sign only at $x = 0$ confirms the results found in step 6 on the concavity of f.

Combining this information, we arrive at the graph of $f(x) = (x-1)^2(x+2)$ shown in the following graph.



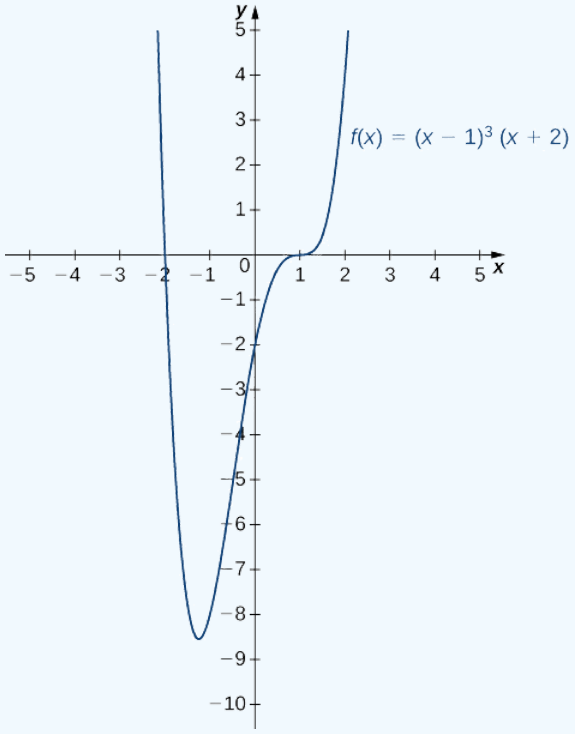
Exercise 4.4.5

Sketch a graph of $f(x) = (x-1)^3(x+2)$.

Hint

f is a fourth-degree polynomial.

Answer



Example 4.4.6: Sketching a Rational Function

Sketch the graph of $f(x) = \dfrac{x^2}{1-x^2}$.

Solution

Step 1: The function f is defined as long as the denominator is not zero. Therefore, the domain is the set of all real numbers x except $x = \pm 1$.

Step 2: Find the intercepts. If $x = 0$, then $f(x) = 0$, so 0 is an intercept. If $y = 0$, then $\dfrac{x^2}{1-x^2} = 0$, which implies $x = 0$. Therefore, $(0, 0)$ is the only intercept.

Step 3: Evaluate the limits at infinity. Since $f$is a rational function, divide the numerator and denominator by the highest power in the denominator: x^2.We obtain

$$\lim_{x\to\pm\infty} \frac{x^2}{1-x^2} = \lim_{x\to\pm\infty} \frac{1}{\frac{1}{x^2}-1} = -1.$$

Therefore, f has a horizontal asymptote of $y = -1$ as $x \to \infty$ and $x \to -\infty$.

Step 4: To determine whether f has any vertical asymptotes, first check to see whether the denominator has any zeroes. We find the denominator is zero when $x = \pm 1$. To determine whether the lines $x = 1$ or $x = -1$ are vertical asymptotes of f, evaluate $\lim\limits_{x\to 1} f(x)$ and $\lim\limits_{x\to -1} f(x)$. By looking at each one-sided limit as $x \to 1$, we see that

$$\lim_{x\to 1^+} \frac{x^2}{1-x^2} = -\infty \text{ and } \lim_{x\to 1^-} \frac{x^2}{1-x^2} = \infty.$$

In addition, by looking at each one-sided limit as $x \to -1$, we find that

$$\lim_{x\to -1^+} \frac{x^2}{1-x^2} = \infty \text{ and } \lim_{x\to -1^-} \frac{x^2}{1-x^2} = -\infty.$$

Step 5: Calculate the first derivative:

$$f'(x)=\frac{(1-x^2)(2x)-x^2(-2x)}{\left(1-x^2\right)^2}=\frac{2x}{\left(1-x^2\right)^2}.$$

Critical points occur at points x where $f'(x)=0$ or $f'(x)$ is undefined. We see that $f'(x)=0$ when $x=0$. The derivative f' is not undefined at any point in the domain of f. However, $x=\pm 1$ are not in the domain of f. Therefore, to determine where f is increasing and where f is decreasing, divide the interval $(-\infty,\infty)$ into four smaller intervals: $(-\infty,-1)$, $(-1,0)$, $(0,1)$, and $(1,\infty)$, and choose a test point in each interval to determine the sign of $f'(x)$ in each of these intervals. The values $x=-2,\ x=-\frac{1}{2},\ x=\frac{1}{2}$, and $x=2$ are good choices for test points as shown in the following table.

Interval	Test point	Sign of $f'(x)=\frac{2x}{(1-x^2)^2}$	Conclusion
$(-\infty,-1)$	$x=-2$	$-/+=-$	f is decreasing.
$(-1,0)$	$x=-1/2$	$-/+=-$	f is decreasing.
$(0,1)$	$x=1/2$	$+/+=+$	f is increasing.
$(1,\infty)$	$x=2$	$+/+=+$	f is increasing.

From this analysis, we conclude that f has a local minimum at $x=0$ but no local maximum.

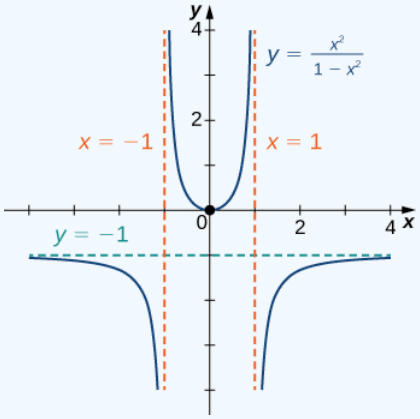
Step 6: Calculate the second derivative:

$$\begin{aligned}f''(x)&=\frac{(1-x^2)^2(2)-2x(2(1-x^2)(-2x))}{(1-x^2)^4}\\&=\frac{(1-x^2)[2(1-x^2)+8x^2]}{\left(1-x^2\right)^4}\\&=\frac{2(1-x^2)+8x^2}{\left(1-x^2\right)^3}\\&=\frac{6x^2+2}{\left(1-x^2\right)^3}.\end{aligned}$$

To determine the intervals where f is concave up and where f is concave down, we first need to find all points x where $f''(x)=0$ or $f''(x)$ is undefined. Since the numerator $6x^2+2\neq 0$ for any x, $f''(x)$ is never zero. Furthermore, f'' is not undefined for any x in the domain of f. However, as discussed earlier, $x=\pm 1$ are not in the domain of f. Therefore, to determine the concavity of f, we divide the interval $(-\infty,\infty)$ into the three smaller intervals $(-\infty,-1)$, $(-1,1)$, and $(1,\infty)$, and choose a test point in each of these intervals to evaluate the sign of $f''(x)$. The values $x=-2,\ x=0$, and $x=2$ are possible test points as shown in the following table.

Interval	Test Point	Sign of $f''(x)=\frac{6x^2+2}{(1-x^2)^3}$	Conclusion
$(-\infty,-1)$	$x=-2$	$+/-=-$	f is concave down.
$(-1,1)$	$x=0$	$+/+=+$	f is concave up
$(1,\infty)$	$x=2$	$+/-=-$	f is concave down.

Combining all this information, we arrive at the graph of f shown below. Note that, although f changes concavity at $x=-1$ and $x=1$, there are no inflection points at either of these places because f is not continuous at $x=-1$ or $x=1$.



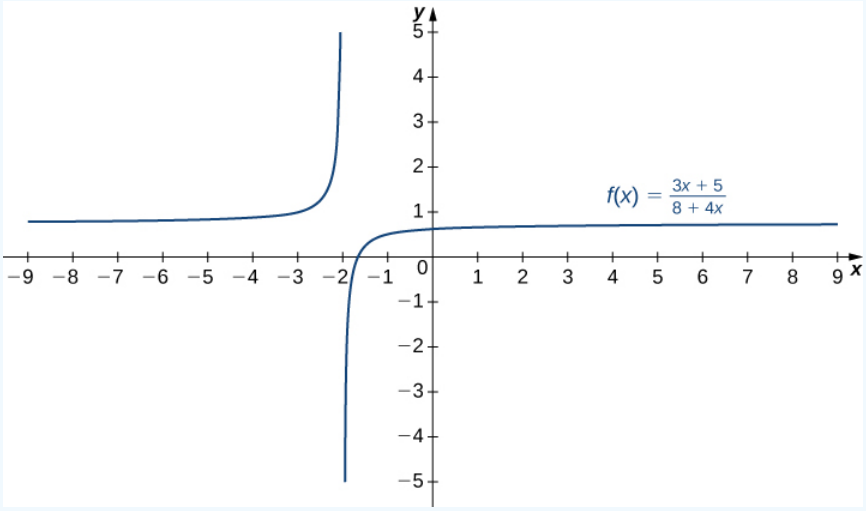
Exercise 4.4.6

Sketch a graph of $f(x)=\dfrac{3x+5}{8+4x}$.

Hint

A line $y=L$ is a horizontal asymptote of f if the limit as $x\to\infty$ or the limit as $x\to-\infty$ of $f(x)$ is L. A line $x=a$ is a vertical asymptote if at least one of the one-sided limits of f as $x\to a$ is ∞ or $-\infty$.

Answer



Example 4.4.7: Sketching a Rational Function with an Oblique Asymptote

Sketch the graph of $f(x)=\dfrac{x^2}{x-1}$

Solution

Step 1: The domain of f is the set of all real numbers x except $x=1$.

Step 2: Find the intercepts. We can see that when $x=0,\ f(x)=0$, so $(0,0)$ is the only intercept.

Step 3: Evaluate the limits at infinity. Since the degree of the numerator is one more than the degree of the denominator, f must have an oblique asymptote. To find the oblique asymptote, use long division of polynomials to write

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

Since $\dfrac{1}{x-1} \to 0$ as $x \to \pm\infty$, $f(x)$ approaches the line $y = x+1$ as $x \to \pm\infty$. The line $y = x+1$ is an oblique asymptote for f.

Step 4: To check for vertical asymptotes, look at where the denominator is zero. Here the denominator is zero at $x = 1$. Looking at both one-sided limits as $x \to 1$, we find

$$\lim_{x\to 1^+} \frac{x^2}{x-1} = \infty \text{ and } \lim_{x\to 1^-} \frac{x^2}{x-1} = -\infty.$$

Therefore, $x = 1$ is a vertical asymptote, and we have determined the behavior of f as x approaches 1 from the right and the left.

Step 5: Calculate the first derivative:

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

We have $f'(x) = 0$ when $x^2 - 2x = x(x-2) = 0$. Therefore, $x = 0$ and $x = 2$ are critical points. Since f is undefined at $x = 1$, we need to divide the interval $(-\infty, \infty)$ into the smaller intervals $(-\infty, 0), (0, 1), (1, 2)$,and $(2, \infty)$, and choose a test point from each interval to evaluate the sign of $f'(x)$ in each of these smaller intervals. For example, let $x = -1, x = \frac{1}{2}, x = \frac{3}{2}$, and $x = 3$ be the test points as shown in the following table.

Interval	Test Point	Sign of $f'(x) = \dfrac{x^2-2x}{(x-1)^2}$	Conclusion
$(-\infty, 0)$	$x = -1$	(−)(−)/+=+	f is increasing.
$(0, 1)$	$x = 1/2$	(+)(−)/+=−	f is decreasing.
$(1, 2)$	$x = 3/2$	(+)(−)/+=−	f is decreasing.
$(2, \infty)$	$x = 3$	(+)(+)/+=+	f is increasing.

From this table, we see that f has a local maximum at $x = 0$ and a local minimum at $x = 2$. The value of f at the local maximum is $f(0) = 0$ and the value of f at the local minimum is $f(2) = 4$. Therefore, $(0, 0)$ and $(2, 4)$ are important points on the graph.

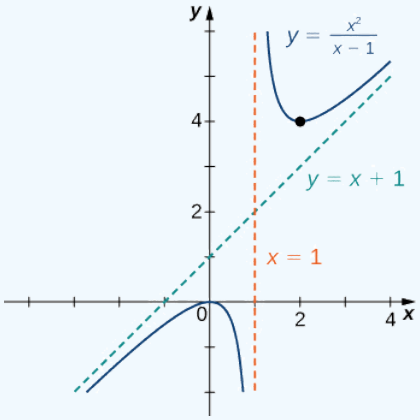
Step 6. Calculate the second derivative:

$$\begin{aligned} f''(x) &= \frac{(x-1)^2(2x-2) - 2(x-1)(x^2-2x)}{(x-1)^4} \\ &= \frac{2(x-1)[(x-1)^2 - (x^2-2x)]}{(x-1)^4} \\ &= \frac{2[x^2-2x+1-x^2+2x]}{(x-1)^3} \\ &= \frac{2}{(x-1)^3}. \end{aligned}$$

We see that $f''(x)$ is never zero or undefined for x in the domain of f. Since f is undefined at $x = 1$, to check concavity we just divide the interval $(-\infty, \infty)$ into the two smaller intervals $(-\infty, 1)$ and $(1, \infty)$, and choose a test point from each interval to evaluate the sign of $f''(x)$ in each of these intervals. The values $x = 0$ and $x = 2$ are possible test points as shown in the following table.

Interval	Test Point	Sign of $f''(x) = \dfrac{2}{(x-1)^3}$	Conclusion
$(-\infty, 1)$	$x = 0$	$+/- = -$	f is concave down.
$(1, \infty)$	$x = 2$	$+/+ = +$	f is concave up

From the information gathered, we arrive at the following graph for f.



Exercise 4.4.7

Find the oblique asymptote for $f(x) = \dfrac{3x^3 - 2x + 1}{2x^2 - 4}$.

Hint

Use long division of polynomials.

Answer

$y = \frac{3}{2}x$

Example 4.4.8: Sketching the Graph of a Function with a Cusp

Sketch a graph of $f(x) = (x-1)^{2/3}$

Solution

Step 1: Since the cube-root function is defined for all real numbers x and $(x-1)^{2/3} = (\sqrt[3]{x-1})^2$, the domain of f is all real numbers.

Step 2: To find the y-intercept, evaluate $f(0)$. Since $f(0) = 1$, the y-intercept is $(0, 1)$. To find the x-intercept, solve $(x-1)^{2/3} = 0$. The solution of this equation is $x = 1$, so the x-intercept is $(1, 0)$.

Step 3: Since $\lim\limits_{x \to \pm\infty} (x-1)^{2/3} = \infty$, the function continues to grow without bound as $x \to \infty$ and $x \to -\infty$.

Step 4: The function has no vertical asymptotes.

Step 5: To determine where f is increasing or decreasing, calculate f'. We find

$$f'(x) = \frac{2}{3}(x-1)^{-1/3} = \frac{2}{3(x-1)^{1/3}}$$

This function is not zero anywhere, but it is undefined when $x = 1$. Therefore, the only critical point is $x = 1$. Divide the interval $(-\infty, \infty)$ into the smaller intervals $(-\infty, 1)$ and $(1, \infty)$, and choose test points in each of these intervals to determine the sign of $f'(x)$ in each of these smaller intervals. Let $x = 0$ and $x = 2$ be the test points as shown in the following table.

Interval	Test Point	Sign of $f'(x) = \frac{2}{3(x-1)^{1/3}}$	Conclusion
$(-\infty, 1)$	$x = 0$	$+/- = -$	f is decreasing
$(1, \infty)$	$x = 2$	$+/+ = +$	f is increasing

We conclude that f has a local minimum at $x=1$. Evaluating f at $x=1$, we find that the value of f at the local minimum is zero. Note that $f'(1)$ is undefined, so to determine the behavior of the function at this critical point, we need to examine $\lim_{x\to 1} f'(x)$. Looking at the one-sided limits, we have

$$\lim_{x\to 1^+} \frac{2}{3(x-1)^{1/3}} = \infty \text{ and } \lim_{x\to 1^-} \frac{2}{3(x-1)^{1/3}} = -\infty.$$

Therefore, f has a cusp at $x=1$.

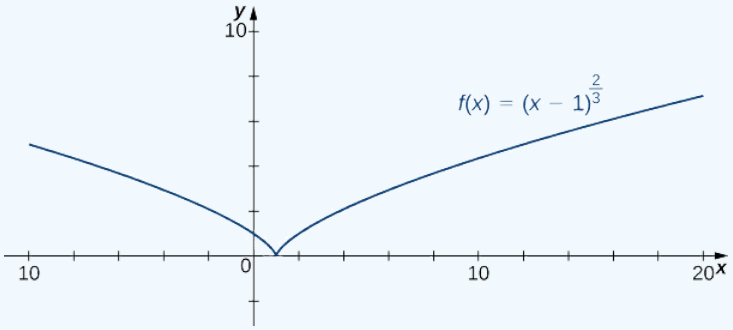
Step 6: To determine concavity, we calculate the second derivative of f :

$$f''(x) = -\frac{2}{9}(x-1)^{-4/3} = \frac{-2}{9(x-1)^{4/3}}.$$

We find that $f''(x)$ is defined for all x, but is undefined when $x=1$. Therefore, divide the interval $(-\infty,\infty)$ into the smaller intervals $(-\infty,1)$ and $(1,\infty)$, and choose test points to evaluate the sign of $f''(x)$ in each of these intervals. As we did earlier, let $x=0$ and $x=2$ be test points as shown in the following table.

Interval	Test Point	Sign of $f''(x) = \frac{-2}{9(x-1)^{4/3}}$	Conclusion
$(-\infty,1)$	$x=0$	$-/+ = -$	f is concave down
$(1,\infty)$	$x=2$	$-/+ = -$	f is concave down

From this table, we conclude that f is concave down everywhere. Combining all of this information, we arrive at the following graph for f.



Exercise 4.4.8

Consider the function $f(x) = 5 - x^{2/3}$. Determine the point on the graph where a cusp is located. Determine the end behavior of f.

Hint

A function f has a cusp at a point a if $f(a)$ exists, $f'(a)$ is undefined, one of the one-sided limits as $x \to a$ of $f'(x)$ is $+\infty$, and the other one-sided limit is $-\infty$.

Answer

The function f has a cusp at $(0,5)$, since $\lim_{x\to 0^-} f'(x) = \infty$ and $\lim_{x\to 0^+} f'(x) = -\infty$. For end behavior, $\lim_{x\to\pm\infty} f(x) = -\infty$.

This page titled 4.4: The Second Derivative Test and Curve Sketching is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin "Jed" Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **4.5: Derivatives and the Shape of a Graph** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.
- **4.6: Limits at Infinity and Asymptotes** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

4.5: Indeterminate Forms and L'Hôpital's Rule

Learning Objectives

- Recognize when to apply L'Hôpital's rule.
- Identify indeterminate forms produced by quotients, products, subtractions, and powers, and apply L'Hôpital's rule in each case.
- Describe the relative growth rates of functions.

In this section, we examine a powerful tool for evaluating limits. This tool, known as **L'Hôpital's rule,** uses derivatives to calculate limits. With this rule, we will be able to evaluate many limits we have not yet been able to determine. Instead of relying on numerical evidence to conjecture that a limit exists, we will be able to show definitively that a limit exists and to determine its exact value.

Applying L'Hôpital's Rule

L'Hôpital's rule can be used to evaluate limits involving the quotient of two functions. Consider

$$\lim_{x\to a}\frac{f(x)}{g(x)}.$$

If $\lim_{x\to a} f(x)=L_1$ and $\lim_{x\to a} g(x)=L_2\neq 0,$ then

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\frac{L_1}{L_2}.$$

However, what happens if $\lim_{x\to a} f(x)=0$ and $\lim_{x\to a} g(x)=0$? We call this one of the **indeterminate forms,** of type $\frac{0}{0}$. This is considered an indeterminate form because we cannot determine the exact behavior of $\frac{f(x)}{g(x)}$ as $x\to a$ without further analysis. We have seen examples of this earlier in the text. For example, consider

$$\lim_{x\to 2}\frac{x^2-4}{x-2}$$

and

$$\lim_{x\to 0}\frac{\sin x}{x}.$$

For the first of these examples, we can evaluate the limit by factoring the numerator and writing

$$\lim_{x\to 2}\frac{x^2-4}{x-2}=\lim_{x\to 2}\frac{(x+2)(x-2)}{x-2}=\lim_{x\to 2}(x+2)=2+2=4.$$

For $\lim_{x\to 0}\frac{\sin x}{x}$ we were able to show, using a geometric argument, that

$$\lim_{x\to 0}\frac{\sin x}{x}=1.$$

Here we use a different technique for evaluating limits such as these. Not only does this technique provide an easier way to evaluate these limits, but also, and more importantly, it provides us with a way to evaluate many other limits that we could not calculate previously.

The idea behind L'Hôpital's rule can be explained using local linear approximations. Consider two differentiable functions f and g such that $\lim_{x\to a} f(x)=0=\lim_{x\to a} g(x)$ and such that $g'(a)\neq 0$ For x near a,we can write

$$f(x)\approx f(a)+f'(a)(x-a)$$

and

$$g(x) \approx g(a) + g'(a)(x-a).$$

Therefore,

$$\frac{f(x)}{g(x)} \approx \frac{f(a) + f'(a)(x-a)}{g(a) + g'(a)(x-a)}.$$

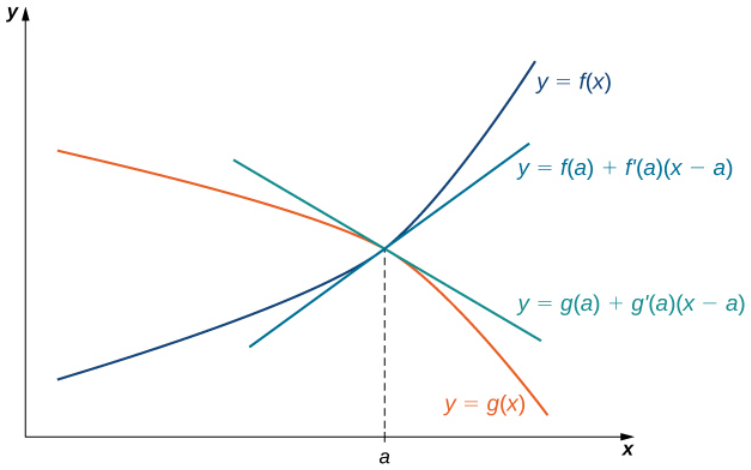


Figure 4.5.1: If $\lim_{x \to a} f(x) = \lim_{x \to a} g(x)$, then the ratio $f(x)/g(x)$ is approximately equal to the ratio of their linear approximations near a.

Since f is differentiable at a, then f is continuous at a, and therefore $f(a) = \lim_{x \to a} f(x) = 0$. Similarly, $g(a) = \lim_{x \to a} g(x) = 0$. If we also assume that f' and g' are continuous at $x = a$, then $f'(a) = \lim_{x \to a} f'(x)$ and $g'(a) = \lim_{x \to a} g'(x)$. Using these ideas, we conclude that

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)(x-a)}{g'(x)(x-a)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}.$$

Note that the assumption that f' and g' are continuous at a and $g'(a) \neq 0$ can be loosened. We state L'Hôpital's rule formally for the indeterminate form $\frac{0}{0}$. Also note that the notation $\frac{0}{0}$ does not mean we are actually dividing zero by zero. Rather, we are using the notation $\frac{0}{0}$ to represent a quotient of limits, each of which is zero.

L'Hôpital's Rule (0/0 Case)

Suppose f and g are differentiable functions over an open interval containing a, except possibly at a. If $\lim_{x \to a} f(x) = 0$ and $\lim_{x \to a} g(x) = 0$, then

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)},$$

assuming the limit on the right exists or is ∞ or $-\infty$. This result also holds if we are considering one-sided limits, or if $a = \infty$ or $a = -\infty$.

Proof

We provide a proof of this theorem in the special case when f, g, f', and g' are all continuous over an open interval containing a. In that case, since $\lim_{x \to a} f(x) = 0 = \lim_{x \to a} g(x)$ and f and g are continuous at a, it follows that $f(a) = 0 = g(a)$. Therefore,

$$\begin{aligned}
\lim_{x\to a}\frac{f(x)}{g(x)} &= \lim_{x\to a}\frac{f(x)-f(a)}{g(x)-g(a)} && \text{Since } f(a)=0=g(a)\\
&= \lim_{x\to a}\frac{\dfrac{f(x)-f(a)}{x-a}}{\dfrac{g(x)-g(a)}{x-a}} && \text{Multiply numerator and denominator by } \frac{1}{x-a}\\
&= \frac{\lim\limits_{x\to a}\dfrac{f(x)-f(a)}{x-a}}{\lim\limits_{x\to a}\dfrac{g(x)-g(a)}{x-a}} && \text{The limit of a quotient is the quotient of the limits.}\\
&= \frac{f'(a)}{g'(a)} && \text{By the definition of the derivative}\\
&= \frac{\lim\limits_{x\to a}f'(x)}{\lim\limits_{x\to a}g'(x)} && \text{By the continuity of } f' \text{ and } g'\\
&= \lim_{x\to a}\frac{f'(x)}{g'(x)}. && \text{The limit of a quotient}
\end{aligned}$$

Note that L'Hôpital's rule states we can calculate the limit of a quotient $\frac{f}{g}$ by considering the limit of the quotient of the derivatives $\frac{f'}{g'}$. It is important to realize that we are not calculating the derivative of the quotient $\frac{f}{g}$.

□

Example 4.5.1: Applying L'Hôpital's Rule (0/0 Case)

Evaluate each of the following limits by applying L'Hôpital's rule.

a. $\lim\limits_{x\to 0}\dfrac{1-\cos x}{x}$

b. $\lim\limits_{x\to 1}\dfrac{\sin(\pi x)}{\ln x}$

c. $\lim\limits_{x\to\infty}\dfrac{e^{1/x}-1}{1/x}$

d. $\lim\limits_{x\to 0}\dfrac{\sin x - x}{x^2}$

Solution

a. Since the numerator $1-\cos x\to 0$ and the denominator $x\to 0$, we can apply L'Hôpital's rule to evaluate this limit. We have

$$\lim_{x\to 0}\frac{1-\cos x}{x}=\lim_{x\to 0}\frac{\frac{d}{dx}(1-\cos x)}{\frac{d}{dx}(x)}=\lim_{x\to 0}\frac{\sin x}{1}=\frac{\lim\limits_{x\to 0}\sin x}{\lim\limits_{x\to 0}1}=\frac{0}{1}=0.$$

b. As $x\to 1$, the numerator $\sin(\pi x)\to 0$ and the denominator $\ln(x)\to 0$. Therefore, we can apply L'Hôpital's rule. We obtain

$$\begin{aligned}
\lim_{x\to 1}\frac{\sin(\pi x)}{\ln x} &= \lim_{x\to 1}\frac{\pi\cos(\pi x)}{1/x}\\
&= \lim_{x\to 1}(\pi x)\cos(\pi x)\\
&= (\pi\cdot 1)(-1)=-\pi.
\end{aligned}$$

c. As $x \to \infty$, the numerator $e^{1/x} - 1 \to 0$ and the denominator $\frac{1}{x} \to 0$. Therefore, we can apply L'Hôpital's rule. We obtain

$$\lim_{x\to\infty} \frac{e^{1/x}-1}{\frac{1}{x}} = \lim_{x\to\infty} \frac{e^{1/x}\left(\frac{-1}{x^2}\right)}{\left(\frac{-1}{x^2}\right)} = \lim_{x\to\infty} e^{1/x} = e^0 = 1.$$

d. As $x \to 0$, both the numerator and denominator approach zero. Therefore, we can apply L'Hôpital's rule. We obtain

$$\lim_{x\to 0} \frac{\sin x - x}{x^2} = \lim_{x\to 0} \frac{\cos x - 1}{2x}.$$

Since the numerator and denominator of this new quotient both approach zero as $x \to 0$, we apply L'Hôpital's rule again. In doing so, we see that

$$\lim_{x\to 0} \frac{\cos x - 1}{2x} = \lim_{x\to 0} \frac{-\sin x}{2} = 0.$$

Therefore, we conclude that

$$\lim_{x\to 0} \frac{\sin x - x}{x^2} = 0.$$

Exercise 4.5.1

Evaluate $\lim_{x\to 0} \frac{x}{\tan x}$.

Hint

$\frac{d}{dx}\left(\tan x\right) = \sec^2 x$

Answer

1

We can also use L'Hôpital's rule to evaluate limits of quotients $\frac{f(x)}{g(x)}$ in which $f(x) \to \pm\infty$ and $g(x) \to \pm\infty$. Limits of this form are classified as indeterminate forms of type ∞/∞. Again, note that we are not actually dividing ∞ by ∞. Since ∞ is not a real number, that is impossible; rather, ∞/∞ is used to represent a quotient of limits, each of which is ∞ or $-\infty$.

L'Hôpital's Rule (∞/∞ Case)

Suppose f and g are differentiable functions over an open interval containing a, except possibly at a. Suppose $\lim_{x\to a} f(x) = \infty$ (or $-\infty$) and $\lim_{x\to a} g(x) = \infty$ (or $-\infty$). Then,

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f\,'(x)}{g'(x)}$$

assuming the limit on the right exists or is ∞ or $-\infty$. This result also holds if the limit is infinite, if $a = \infty$ or $-\infty$, or the limit is one-sided.

Example 4.5.2: Applying L'Hôpital's Rule (∞/∞) Case

Evaluate each of the following limits by applying L'Hôpital's rule.

a. $\lim_{x\to\infty} \frac{3x+5}{2x+1}$

b. $\lim_{x\to 0^+} \frac{\ln x}{\cot x}$

Solution

a. Since $3x+5$ and $2x+1$ are first-degree polynomials with positive leading coefficients, $\lim_{x\to\infty}(3x+5)=\infty$ and $\lim_{x\to\infty}(2x+1)=\infty$. Therefore, we apply L'Hôpital's rule and obtain

$$\lim_{x\to\infty}\frac{3x+5}{2x+1}=\lim_{x\to\infty}\frac{3}{2}=\frac{3}{2}.$$

Note that this limit can also be calculated without invoking L'Hôpital's rule. Earlier in the chapter we showed how to evaluate such a limit by dividing the numerator and denominator by the highest power of x in the denominator. In doing so, we saw that

$$\lim_{x\to\infty}\frac{3x+5}{2x+1}=\lim_{x\to\infty}\frac{3+5/x}{2+1/x}=\frac{3}{2}.$$

L'Hôpital's rule provides us with an alternative means of evaluating this type of limit.

b. Here, $\lim_{x\to0^+}\ln x=-\infty$ and $\lim_{x\to0^+}\cot x=\infty$. Therefore, we can apply L'Hôpital's rule and obtain

$$\lim_{x\to0^+}\frac{\ln x}{\cot x}=\lim_{x\to0^+}\frac{1/x}{-\csc^2 x}=\lim_{x\to0^+}\frac{1}{-x\csc^2 x}.$$

Now as $x\to0^+$, $\csc^2 x\to\infty$. Therefore, the first term in the denominator is approaching zero and the second term is getting really large. In such a case, anything can happen with the product. Therefore, we cannot make any conclusion yet. To evaluate the limit, we use the definition of $\csc x$ to write

$$\lim_{x\to0^+}\frac{1}{-x\csc^2 x}=\lim_{x\to0^+}\frac{\sin^2 x}{-x}.$$

Now $\lim_{x\to0^+}\sin^2 x=0$ and $\lim_{x\to0^+}-x=0$, so we apply L'Hôpital's rule again. We find

$$\lim_{x\to0^+}\frac{\sin^2 x}{-x}=\lim_{x\to0^+}\frac{2\sin x\cos x}{-1}=\frac{0}{-1}=0.$$

We conclude that

$$\lim_{x\to0^+}\frac{\ln x}{\cot x}=0.$$

? Exercise 4.5.2

Evaluate $\lim_{x\to\infty}\frac{\ln x}{5x}$.

Hint

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

Answer

0

As mentioned, L'Hôpital's rule is an extremely useful tool for evaluating limits. It is important to remember, however, that to apply L'Hôpital's rule to a quotient $\frac{f(x)}{g(x)}$, it is essential that the limit of $\frac{f(x)}{g(x)}$ be of the form $\frac{0}{0}$ or ∞/∞. Consider the following example.

Example 4.5.3: When L'Hôpital's Rule Does Not Apply

Consider $\lim\limits_{x\to 1}\dfrac{x^2+5}{3x+4}$.

Show that the limit cannot be evaluated by applying L'Hôpital's rule.

Solution

Because the limits of the numerator and denominator are not both zero and are not both infinite, we cannot apply L'Hôpital's rule. If we try to do so, we get

$$\frac{d}{dx}(x^2+5)=2x$$

and

$$\frac{d}{dx}(3x+4)=3.$$

At which point we would conclude erroneously that

$$\lim_{x\to 1}\frac{x^2+5}{3x+4}=\lim_{x\to 1}\frac{2x}{3}=\frac{2}{3}.$$

However, since $\lim\limits_{x\to 1}(x^2+5)=6$ and $\lim\limits_{x\to 1}(3x+4)=7,$ we actually have

$$\lim_{x\to 1}\frac{x^2+5}{3x+4}=\frac{6}{7}.$$

We can conclude that

$$\lim_{x\to 1}\frac{x^2+5}{3x+4}\neq\lim_{x\to 1}\frac{\dfrac{d}{dx}(x^2+5)}{\dfrac{d}{dx}(3x+4)}.$$

Exercise 4.5.3

Explain why we cannot apply L'Hôpital's rule to evaluate $\lim\limits_{x\to 0^+}\dfrac{\cos x}{x}$. Evaluate $\lim\limits_{x\to 0^+}\dfrac{\cos x}{x}$ by other means.

Hint

Determine the limits of the numerator and denominator separately.

Answer

$\lim\limits_{x\to 0^+}\cos x=1$. Therefore, we cannot apply L'Hôpital's rule. The limit of the quotient is ∞.

Other Indeterminate Forms

L'Hôpital's rule is very useful for evaluating limits involving the indeterminate forms $\dfrac{0}{0}$ and ∞/∞. However, we can also use L'Hôpital's rule to help evaluate limits involving other indeterminate forms that arise when evaluating limits. The expressions $0\cdot\infty, \infty-\infty, 1^\infty, \infty^0$, and 0^0 are all considered indeterminate forms. These expressions are not real numbers. Rather, they represent forms that arise when trying to evaluate certain limits. Next we realize why these are indeterminate forms and then understand how to use L'Hôpital's rule in these cases. The key idea is that we must rewrite the indeterminate forms in such a way that we arrive at the indeterminate form $\dfrac{0}{0}$ or ∞/∞.

Indeterminate Form of Type $0 \cdot \infty$

Suppose we want to evaluate $\lim_{x \to a}(f(x) \cdot g(x))$, where $f(x) \to 0$ and $g(x) \to \infty$ (or $-\infty$) as $x \to a$. Since one term in the product is approaching zero but the other term is becoming arbitrarily large (in magnitude), anything can happen to the product. We use the notation $0 \cdot \infty$ to denote the form that arises in this situation. The expression $0 \cdot \infty$ is considered indeterminate because we cannot determine without further analysis the exact behavior of the product $f(x)g(x)$ as $x \to \infty$. For example, let n be a positive integer and consider

$$f(x) = \frac{1}{(x^n + 1)} \text{ and } g(x) = 3x^2.$$

As $x \to \infty$, $f(x) \to 0$ and $g(x) \to \infty$. However, the limit as $x \to \infty$ of $f(x)g(x) = \dfrac{3x^2}{(x^n + 1)}$ varies, depending on n. If $n = 2$, then $\lim_{x \to \infty} f(x)g(x) = 3$. If $n = 1$, then $\lim_{x \to \infty} f(x)g(x) = \infty$. If $n = 3$, then $\lim_{x \to \infty} f(x)g(x) = 0$. Here we consider another limit involving the indeterminate form $0 \cdot \infty$ and show how to rewrite the function as a quotient to use L'Hôpital's rule.

Example 4.5.4: Indeterminate Form of Type $0 \cdot \infty$

Evaluate $\lim_{x \to 0^+} x \ln x$.

Solution

First, rewrite the function $x \ln x$ as a quotient to apply L'Hôpital's rule. If we write

$$x \ln x = \frac{\ln x}{1/x}$$

we see that $\ln x \to -\infty$ as $x \to 0^+$ and $\dfrac{1}{x} \to \infty$ as $x \to 0^+$. Therefore, we can apply L'Hôpital's rule and obtain

$$\lim_{x \to 0^+} \frac{\ln x}{1/x} = \lim_{x \to 0^+} \frac{\frac{d}{dx}(\ln x)}{\frac{d}{dx}(1/x)} = \lim_{x \to 0^+} \frac{1/x}{-1/x^2} = \lim_{x \to 0^+} (-x) = 0.$$

We conclude that

$$\lim_{x \to 0^+} x \ln x = 0.$$

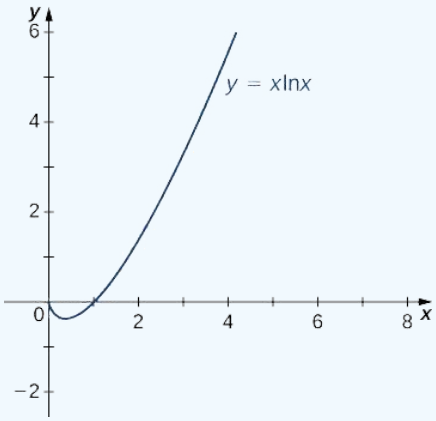


Figure 4.5.2: Finding the limit at $x = 0$ of the function $f(x) = x \ln x$.

Exercise 4.5.4

Evaluate $\lim_{x \to 0} x \cot x$.

Hint

Write $x \cot x = \dfrac{x \cos x}{\sin x}$

Answer

1

Indeterminate Form of Type ∞-∞

Another type of indeterminate form is $\infty - \infty$. Consider the following example. Let n be a positive integer and let $f(x) = 3x^n$ and $g(x) = 3x^2 + 5$. As $x \to \infty$, $f(x) \to \infty$ and $g(x) \to \infty$. We are interested in $\lim\limits_{x \to \infty} (f(x) - g(x))$. Depending on whether $f(x)$ grows faster, $g(x)$ grows faster, or they grow at the same rate, as we see next, anything can happen in this limit. Since $f(x) \to \infty$ and $g(x) \to \infty$, we write $\infty - \infty$ to denote the form of this limit. As with our other indeterminate forms, $\infty - \infty$ has no meaning on its own and we must do more analysis to determine the value of the limit. For example, suppose the exponent n in the function $f(x) = 3x^n$ is $n = 3$, then

$$\lim_{x \to \infty} (f(x) - g(x)) = \lim_{x \to \infty} (3x^3 - 3x^2 - 5) = \infty.$$

On the other hand, if $n = 2$, then

$$\lim_{x \to \infty} (f(x) - g(x)) = \lim_{x \to \infty} (3x^2 - 3x^2 - 5) = -5.$$

However, if $n = 1$, then

$$\lim_{x \to \infty} (f(x) - g(x)) = \lim_{x \to \infty} (3x - 3x^2 - 5) = -\infty.$$

Therefore, the limit cannot be determined by considering only $\infty - \infty$. Next we see how to rewrite an expression involving the indeterminate form $\infty - \infty$ as a fraction to apply L'Hôpital's rule.

Example 4.5.5: Indeterminate Form of Type $\infty - \infty$

Evaluate $\lim\limits_{x \to 0^+} \left(\dfrac{1}{x^2} - \dfrac{1}{\tan x} \right)$.

Solution

By combining the fractions, we can write the function as a quotient. Since the least common denominator is $x^2 \tan x$, we have

$$\frac{1}{x^2} - \frac{1}{\tan x} = \frac{(\tan x) - x^2}{x^2 \tan x}.$$

As $x \to 0^+$, the numerator $\tan x - x^2 \to 0$ and the denominator $x^2 \tan x \to 0$. Therefore, we can apply L'Hôpital's rule. Taking the derivatives of the numerator and the denominator, we have

$$\lim_{x \to 0^+} \frac{(\tan x) - x^2}{x^2 \tan x} = \lim_{x \to 0^+} \frac{(\sec^2 x) - 2x}{x^2 \sec^2 x + 2x \tan x}.$$

As $x \to 0^+$, $(\sec^2 x) - 2x \to 1$ and $x^2 \sec^2 x + 2x \tan x \to 0$. Since the denominator is positive as x approaches zero from the right, we conclude that

$$\lim_{x \to 0^+} \frac{(\sec^2 x) - 2x}{x^2 \sec^2 x + 2x \tan x} = \infty.$$

Therefore,

$$\lim_{x \to 0^+} \left(\frac{1}{x^2} - \frac{1}{\tan x} \right) = \infty.$$

Exercise 4.5.5

Evaluate $\lim_{x\to 0^+}\left(\frac{1}{x}-\frac{1}{\sin x}\right)$.

Hint

Rewrite the difference of fractions as a single fraction.

Answer

0

Indeterminate Forms Involving Exponents

Another type of indeterminate form that arises when evaluating limits involves exponents. The expressions 0^0, ∞^0, and 1^∞ are all indeterminate forms. On their own, these expressions are meaningless because we cannot actually evaluate these expressions as we would evaluate an expression involving real numbers. Rather, these expressions represent forms that arise when finding limits. Now we examine how L'Hôpital's rule can be used to evaluate limits involving these indeterminate forms.

Since L'Hôpital's rule applies to quotients, we use the natural logarithm function and its properties to reduce a problem evaluating a limit involving exponents to a related problem involving a limit of a quotient. For example, suppose we want to evaluate $\lim_{x\to a} f(x)^{g(x)}$ and we arrive at the indeterminate form ∞^0. (The indeterminate forms 0^0 and 1^∞ can be handled similarly.) We proceed as follows. Let

$$y = f(x)^{g(x)}.$$

Then,

$$\ln y = \ln(f(x)^{g(x)}) = g(x)\ln(f(x)).$$

Therefore,

$$\lim_{x\to a}[\ln(y)] = \lim_{x\to a}[g(x)\ln(f(x))].$$

Since $\lim_{x\to a} f(x) = \infty$, we know that $\lim_{x\to a}\ln(f(x)) = \infty$. Therefore, $\lim_{x\to a} g(x)\ln(f(x))$ is of the indeterminate form $0\cdot\infty$, and we can use the techniques discussed earlier to rewrite the expression $g(x)\ln(f(x))$ in a form so that we can apply L'Hôpital's rule. Suppose $\lim_{x\to a} g(x)\ln(f(x)) = L$, where L may be ∞ or $-\infty$. Then

$$\lim_{x\to a}[\ln(y)] = L.$$

Since the natural logarithm function is continuous, we conclude that

$$\ln\left(\lim_{x\to a} y\right) = L,$$

which gives us

$$\lim_{x\to a} y = \lim_{x\to a} f(x)^{g(x)} = e^L.$$

Example 4.5.6: Indeterminate Form of Type ∞^0

Evaluate $\lim_{x\to\infty} x^{1/x}$.

Solution

Let $y = x^{1/x}$.Then,

$$\ln(x^{1/x}) = \frac{1}{x}\ln x = \frac{\ln x}{x}.$$

We need to evaluate $\lim_{x\to\infty}\dfrac{\ln x}{x}$. Applying L'Hôpital's rule, we obtain

$$\lim_{x\to\infty}\ln y=\lim_{x\to\infty}\frac{\ln x}{x}=\lim_{x\to\infty}\frac{1/x}{1}=0.$$

Therefore, $\lim_{x\to\infty}\ln y=0$. Since the natural logarithm function is continuous, we conclude that

$$\ln\left(\lim_{x\to\infty}y\right)=0,$$

which leads to

$$\lim_{x\to\infty}x^{1/x}=\lim_{x\to\infty}y=e^{\ln\left(\lim_{x\to\infty}y\right)}=e^0=1.$$

Hence,

$$\lim_{x\to\infty}x^{1/x}=1.$$

Exercise 4.5.6

Evaluate $\lim_{x\to\infty}x^{1/\ln(x)}$.

Hint

Let $y=x^{1/\ln(x)}$ and apply the natural logarithm to both sides of the equation.

Answer

e

Example 4.5.7: Indeterminate Form of Type 0^0

Evaluate $\lim_{x\to0^+}x^{\sin x}$.

Solution

Let

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

Therefore,

$$\ln y=\ln(x^{\sin x})=\sin x\ln x.$$

We now evaluate $\lim_{x\to0^+}\sin x\ln x$. Since $\lim_{x\to0^+}\sin x=0$ and $\lim_{x\to0^+}\ln x=-\infty$, we have the indeterminate form $0\cdot\infty$. To apply L'Hôpital's rule, we need to rewrite $\sin x\ln x$ as a fraction. We could write

$$\sin x\ln x=\frac{\sin x}{1/\ln x}$$

or

$$\sin x\ln x=\frac{\ln x}{1/\sin x}=\frac{\ln x}{\csc x}.$$

Let's consider the first option. In this case, applying L'Hôpital's rule, we would obtain

$$\lim_{x\to0^+}\sin x\ln x=\lim_{x\to0^+}\frac{\sin x}{1/\ln x}=\lim_{x\to0^+}\frac{\cos x}{-1/(x(\ln x)^2)}=\lim_{x\to0^+}(-x(\ln x)^2\cos x).$$

Unfortunately, we not only have another expression involving the indeterminate form $0 \cdot \infty$, but the new limit is even more complicated to evaluate than the one with which we started. Instead, we try the second option. By writing

$$\sin x \ln x = \frac{\ln x}{1/\sin x} = \frac{\ln x}{\csc x},$$

and applying L'Hôpital's rule, we obtain

$$\lim_{x \to 0^+} \sin x \ln x = \lim_{x \to 0^+} \frac{\ln x}{\csc x} = \lim_{x \to 0^+} \frac{1/x}{-\csc x \cot x} = \lim_{x \to 0^+} \frac{-1}{x \csc x \cot x}.$$

Using the fact that $\csc x = \dfrac{1}{\sin x}$ and $\cot x = \dfrac{\cos x}{\sin x}$, we can rewrite the expression on the right-hand side as

$$\lim_{x \to 0^+} \frac{-\sin^2 x}{x \cos x} = \lim_{x \to 0^+} \left[\frac{\sin x}{x} \cdot (-\tan x) \right] = \left(\lim_{x \to 0^+} \frac{\sin x}{x} \right) \cdot \left(\lim_{x \to 0^+} (-\tan x) \right) = 1 \cdot 0 = 0.$$

We conclude that $\lim\limits_{x \to 0^+} \ln y = 0$. Therefore, $\ln \left(\lim\limits_{x \to 0^+} y \right) = 0$ and we have

$$\lim_{x \to 0^+} y = \lim_{x \to 0^+} x^{\sin x} = e^0 = 1.$$

Hence,

$$\lim_{x \to 0^+} x^{\sin x} = 1.$$

? Exercise 4.5.7

Evaluate $\lim\limits_{x \to 0^+} x^x$.

Hint

Let $y = x^x$ and take the natural logarithm of both sides of the equation.

Answer

1

Growth Rates of Functions

Suppose the functions f and g both approach infinity as $x \to \infty$. Although the values of both functions become arbitrarily large as the values of x become sufficiently large, sometimes one function is growing more quickly than the other. For example, $f(x) = x^2$ and $g(x) = x^3$ both approach infinity as $x \to \infty$. However, as Table 4.5.1 shows, the values of x^3 are growing much faster than the values of x^2.

Table 4.5.1: Comparing the Growth Rates of x^2 and x^3

x	10	100	1000	10,000
$f(x) = x^2$	100	10,000	1,000,000	100,000,000
$g(x) = x^3$	1000	1,000,000	1,000,000,000	1,000,000,000,000

In fact,

$$\lim_{x \to \infty} \frac{x^3}{x^2} = \lim_{x \to \infty} x = \infty.$$

or, equivalently

$$\lim_{x \to \infty} \frac{x^2}{x^3} = \lim_{x \to \infty} \frac{1}{x} = 0.$$

As a result, we say x^3 is growing more rapidly than x^2 as $x \to \infty$. On the other hand, for $f(x) = x^2$ and $g(x) = 3x^2 + 4x + 1$, although the values of $g(x)$ are always greater than the values of $f(x)$ for $x > 0$, each value of $g(x)$ is roughly three times the corresponding value of $f(x)$ as $x \to \infty$, as shown in Table 4.5.2. In fact,

$$\lim_{x \to \infty} \frac{x^2}{3x^2 + 4x + 1} = \frac{1}{3}.$$

Table 4.5.2: Comparing the Growth Rates of x^2 and $3x^2 + 4x + 1$

x	10	100	1000	10,000
$f(x) = x^2$	100	10,000	1,000,000	100,000,000
$g(x) = 3x^2 + 4x + 1$	341	30,401	3,004,001	300,040,001

In this case, we say that x^2 and $3x^2 + 4x + 1$ are growing at the same rate as $x \to \infty$.

More generally, suppose f and g are two functions that approach infinity as $x \to \infty$. We say g grows more rapidly than f as $x \to \infty$ if

$$\lim_{x \to \infty} \frac{g(x)}{f(x)} = \infty \quad \text{or, equivalently,} \quad \lim_{x \to \infty} \frac{f(x)}{g(x)} = 0.$$

On the other hand, if there exists a constant $M \neq 0$ such that

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = M,$$

we say f and g grow at the same rate as $x \to \infty$.

Next we see how to use L'Hôpital's rule to compare the growth rates of power, exponential, and logarithmic functions.

Example 4.5.8: Comparing the Growth Rates of $\ln(x)$, x^2, and e^x

For each of the following pairs of functions, use L'Hôpital's rule to evaluate

$$\lim_{x \to \infty} \frac{f(x)}{g(x)}.$$

a. $f(x) = x^2$ and $g(x) = e^x$
b. $f(x) = \ln(x)$ and $g(x) = x^2$

Solution

a. Since $\lim_{x \to \infty} x^2 = \infty$ and $\lim_{x \to \infty} e^x = \infty$, we can use L'Hôpital's rule to evaluate $\lim_{x \to \infty} \left[\frac{x^2}{e^x}\right]$. We obtain

$$\lim_{x \to \infty} \frac{x^2}{e^x} = \lim_{x \to \infty} \frac{2x}{e^x}.$$

Since $\lim_{x \to \infty} 2x = \infty$ and $\lim_{x \to \infty} e^x = \infty$, we can apply L'Hôpital's rule again. Since

$$\lim_{x \to \infty} \frac{2x}{e^x} = \lim_{x \to \infty} \frac{2}{e^x} = 0,$$

we conclude that

$$\lim_{x \to \infty} \frac{x^2}{e^x} = 0.$$

Therefore, e^x grows more rapidly than x^2 as $x \to \infty$ (See Figure 4.5.3 and Table 4.5.3)

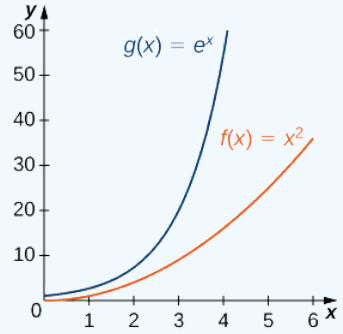


Figure 4.5.3: An exponential function grows at a faster rate than a power function.

Table 4.5.3: Growth rates of a power function and an exponential function.

x	5	10	15	20
x^2	25	100	225	400
e^x	148	22,026	3,269,017	485,165,195

b. Since $\lim_{x\to\infty} \ln x = \infty$ and $\lim_{x\to\infty} x^2 = \infty$, we can use L'Hôpital's rule to evaluate $\lim_{x\to\infty} \dfrac{\ln x}{x^2}$. We obtain

$$\lim_{x\to\infty} \frac{\ln x}{x^2} = \lim_{x\to\infty} \frac{1/x}{2x} = \lim_{x\to\infty} \frac{1}{2x^2} = 0.$$

Thus, x^2 grows more rapidly than $\ln x$ as $x \to \infty$ (see Figure 4.5.4 and Table 4.5.4).

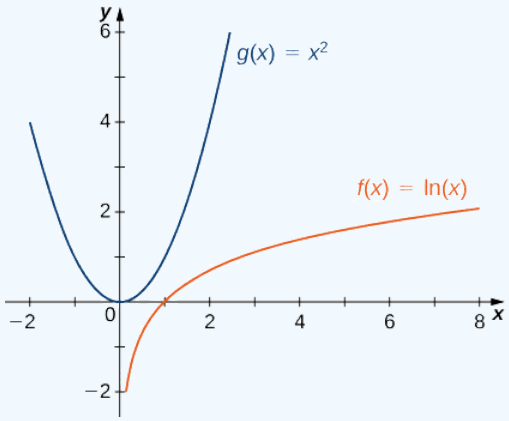


Figure 4.5.4: A power function grows at a faster rate than a logarithmic function.

Table 4.5.4: Growth rates of a power function and a logarithmic function

x	10	100	1000	10,000
$\ln(x)$	2.303	4.605	6.908	9.210
x^2	100	10,000	1,000,000	100,000,000

? Exercise 4.5.8

Compare the growth rates of x^{100} and 2^x.

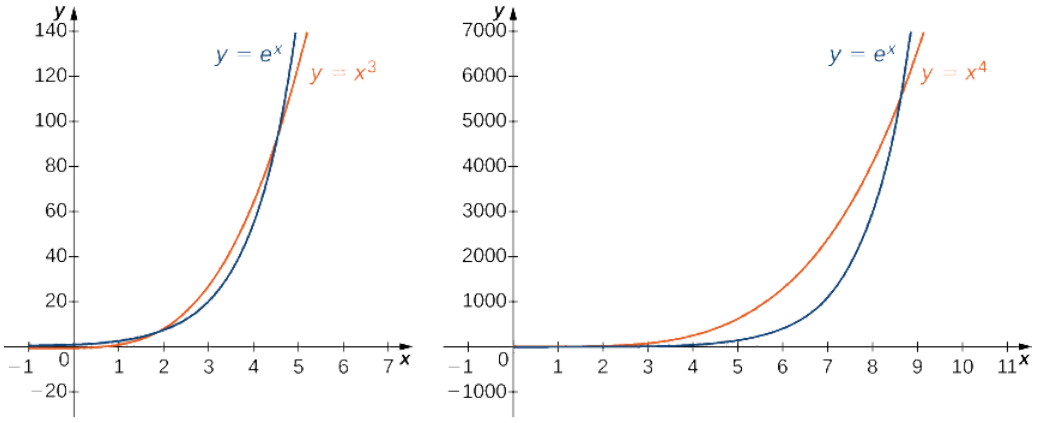
Hint

Apply L'Hôpital's rule to $x^{100}/2^x$.

Answer

The function 2^x grows faster than x^{100}.

Using the same ideas as in Example 4.5.8a, it is not difficult to show that e^x grows more rapidly than x^p for any $p > 0$. In Figure 4.5.5 and Table 4.5.5, we compare e^x with x^3 and x^4 as $x \to \infty$.



(a) (b)

Figure 4.5.5: The exponential function e^x grows faster than x^p for any $p > 0$. (a) A comparison of e^x with x^3. (b) A comparison of e^x with x^4.

Table 4.5.5: An exponential function grows at a faster rate than any power function

x	5	10	15	20
x^3	125	1000	3375	8000
x^4	625	10,000	50,625	160,000
e^x	148	22,026	3,269,017	485,165,195

Similarly, it is not difficult to show that x^p grows more rapidly than $\ln x$ for any $p > 0$. In Figure 4.5.6 and Table 4.5.6, we compare $\ln x$ with $\sqrt[3]{x}$ and $\sqrt{x}$.

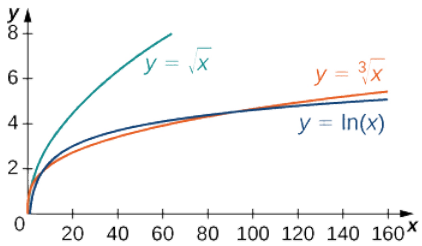


Figure 4.5.6: The function $y = \ln(x)$ grows more slowly than x^p for any $p > 0$ as $x \to \infty$.

Table 4.5.6: A logarithmic function grows at a slower rate than any root function

x	10	100	1000	10,000
$\ln(x)$	2.303	4.605	6.908	9.210
$\sqrt[3]{x}$	2.154	4.642	10	21.544
$\sqrt{x}$	3.162	10	31.623	100

Key Concepts

- L'Hôpital's rule can be used to evaluate the limit of a quotient when the indeterminate form $\frac{0}{0}$ or ∞/∞ arises.
- L'Hôpital's rule can also be applied to other indeterminate forms if they can be rewritten in terms of a limit involving a quotient that has the indeterminate form $\frac{0}{0}$ or ∞/∞.

- The exponential function e^x grows faster than any power function $x^p, p > 0$.
- The logarithmic function $\ln x$ grows more slowly than any power function $x^p, p > 0$.

Glossary

indeterminate forms

When evaluating a limit, the forms $\frac{0}{0}, \infty/\infty, 0 \cdot \infty, \infty - \infty, 0^0, \infty^0$, and 1^∞ are considered indeterminate because further analysis is required to determine whether the limit exists and, if so, what its value is.

L'Hôpital's rule

If f and g are differentiable functions over an interval a, except possibly at a, and $\lim_{x \to a} f(x) = 0 = \lim_{x \to a} g(x)$ or $\lim_{x \to a} f(x)$ and $\lim_{x \to a} g(x)$ are infinite, then $\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}$, assuming the limit on the right exists or is ∞ or $-\infty$.

4.5: Indeterminate Forms and L'Hôpital's Rule is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **4.8: L'Hôpital's Rule** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

4.6: Applied Optimization Problems

Learning Objectives

- Set up and solve optimization problems in several applied fields.

One common application of calculus is calculating the minimum or maximum value of a function. For example, companies often want to minimize production costs or maximize revenue. In manufacturing, it is often desirable to minimize the amount of material used to package a product with a certain volume. In this section, we show how to set up these types of minimization and maximization problems and solve them by using the tools developed in this chapter.

Solving Optimization Problems over a Closed, Bounded Interval

The basic idea of the optimization problems that follow is the same. We have a particular quantity that we are interested in maximizing or minimizing. However, we also have some auxiliary condition that needs to be satisfied. For example, in Example 4.6.1, we are interested in maximizing the area of a rectangular garden. Certainly, if we keep making the side lengths of the garden larger, the area will continue to become larger. However, what if we have some restriction on how much fencing we can use for the perimeter? In this case, we cannot make the garden as large as we like. Let's look at how we can maximize the area of a rectangle subject to some constraint on the perimeter.

Example 4.6.1: Maximizing the Area of a Garden

A rectangular garden is to be constructed using a rock wall as one side of the garden and wire fencing for the other three sides (Figure 4.6.1). Given 100 ft of wire fencing, determine the dimensions that would create a garden of maximum area. What is the maximum area?

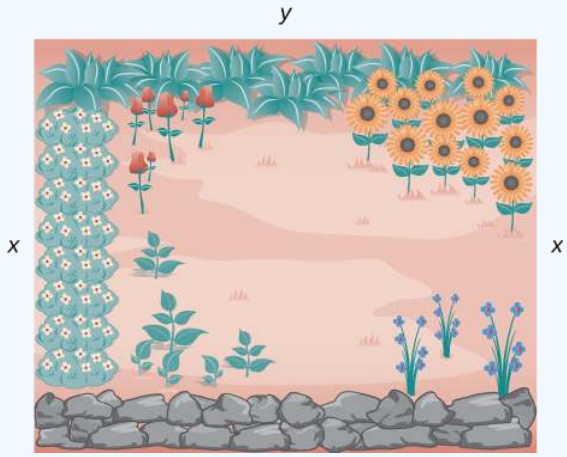


Figure 4.6.1: We want to determine the measurements x and y that will create a garden with a maximum area using 100 ft of fencing.

Solution

Let x denote the length of the side of the garden perpendicular to the rock wall and y denote the length of the side parallel to the rock wall. Then the area of the garden is

$$A = x \cdot y.$$

We want to find the maximum possible area subject to the constraint that the total fencing is 100 ft. From Figure 4.6.1, the total amount of fencing used will be $2x + y$. Therefore, the constraint equation is

$$2x + y = 100.$$

Solving this equation for y, we have $y = 100 - 2x$. Thus, we can write the area as

$$A(x) = x \cdot (100 - 2x) = 100x - 2x^2.$$

Before trying to maximize the area function $A(x) = 100x - 2x^2$, we need to determine the domain under consideration. To construct a rectangular garden, we certainly need the lengths of both sides to be positive. Therefore, we need $x > 0$ and $y > 0$. Since $y = 100 - 2x$, if $y > 0$, then $x < 50$. Therefore, we are trying to determine the maximum value of $A(x)$ for x over the open interval $(0, 50)$. We do not know that a function necessarily has a maximum value over an open interval. However, we do know that a continuous function has an absolute maximum (and absolute minimum) over a closed interval. Therefore, let's consider the function $A(x) = 100x - 2x^2$ over the closed interval $[0, 50]$. If the maximum value occurs at an interior point, then we have found the value x in the open interval $(0, 50)$ that maximizes the area of the garden.

Therefore, we consider the following problem:

Maximize $A(x) = 100x - 2x^2$ over the interval $[0, 50]$.

As mentioned earlier, since A is a continuous function on a closed, bounded interval, by the extreme value theorem, it has a maximum and a minimum. These extreme values occur either at endpoints or critical points. At the endpoints, $A(x) = 0$. Since the area is positive for all x in the open interval $(0, 50)$, the maximum must occur at a critical point. Differentiating the function $A(x)$, we obtain

$$A'(x) = 100 - 4x.$$

Therefore, the only critical point is $x = 25$ (Figure 4.6.2). We conclude that the maximum area must occur when $x = 25$.

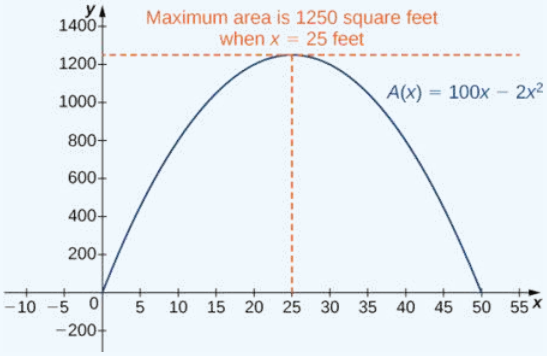


Figure 4.6.2: To maximize the area of the garden, we need to find the maximum value of the function $A(x) = 100x - 2x^2$.

Then we have $y = 100 - 2x = 100 - 2(25) = 50$. To maximize the area of the garden, let $x = 25$ ft and $y = 50$ ft. The area of this garden is 1250 ft^2.

Exercise 4.6.1

Determine the maximum area if we want to make the same rectangular garden as in Figure 4.6.2, but we have 200 ft of fencing.

Hint

We need to maximize the function $A(x) = 200x - 2x^2$ over the interval $[0, 100]$.

Answer

The maximum area is 5000 ft^2.

Now let's look at a general strategy for solving optimization problems similar to Example 4.6.1.

Problem-Solving Strategy: Solving Optimization Problems

1. Introduce all variables. If applicable, draw a figure and label all variables.
2. Determine which quantity is to be maximized or minimized, and for what range of values of the other variables (if this can be determined at this time).
3. Write a formula for the quantity to be maximized or minimized in terms of the variables. This formula may involve more than one variable.

4. Write any equations relating the independent variables in the formula from step 3. Use these equations to write the quantity to be maximized or minimized as a function of one variable.
5. Identify the domain of consideration for the function in step 4 based on the physical problem to be solved.
6. Locate the maximum or minimum value of the function from step 4. This step typically involves looking for critical points and evaluating a function at endpoints.

Now let's apply this strategy to maximize the volume of an open-top box given a constraint on the amount of material to be used.

Example 4.6.2: Maximizing the Volume of a Box

An open-top box is to be made from a 24 in. by 36 in. piece of cardboard by removing a square from each corner of the box and folding up the flaps on each side. What size square should be cut out of each corner to get a box with the maximum volume?

Solution

Step 1: Let x be the side length of the square to be removed from each corner (Figure $4.6.3$). Then, the remaining four flaps can be folded up to form an open-top box. Let V be the volume of the resulting box.

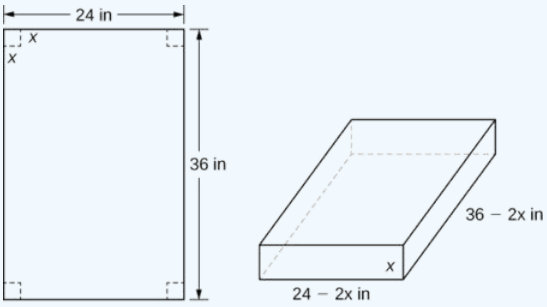


Figure $4.6.3$: A square with side length x inches is removed from each corner of the piece of cardboard. The remaining flaps are folded to form an open-top box.

Step 2: We are trying to maximize the volume of a box. Therefore, the problem is to maximize V.

Step 3: As mentioned in step 2, are trying to maximize the volume of a box. The volume of a box is

$$V = L \cdot W \cdot H,$$

where L, W, and H are the length, width, and height, respectively.

Step 4: From Figure $4.6.3$, we see that the height of the box is x inches, the length is $36 - 2x$ inches, and the width is $24 - 2x$ inches. Therefore, the volume of the box is

$$\begin{aligned} V(x) &= (36 - 2x)(24 - 2x)x \\ &= 4x^3 - 120x^2 + 864x \end{aligned}.$$

Step 5: To determine the domain of consideration, let's examine Figure $4.6.3$. Certainly, we need $x > 0$. Furthermore, the side length of the square cannot be greater than or equal to half the length of the shorter side, 24 in.; otherwise, one of the flaps would be completely cut off. Therefore, we are trying to determine whether there is a maximum volume of the box for x over the open interval $(0, 12)$. Since V is a continuous function over the closed interval $[0, 12]$, we know V will have an absolute maximum over the closed interval. Therefore, we consider V over the closed interval $[0, 12]$ and check whether the absolute maximum occurs at an interior point.

Step 6: Since $V(x)$ is a continuous function over the closed, bounded interval $[0, 12]$, V must have an absolute maximum (and an absolute minimum). Since $V(x) = 0$ at the endpoints and $V(x) > 0$ for $0 < x < 12$, the maximum must occur at a critical point. The derivative is

$$V'(x) = 12x^2 - 240x + 864.$$

To find the critical points, we need to solve the equation

$$12x^2 - 240x + 864 = 0.$$

Dividing both sides of this equation by 12, the problem simplifies to solving the equation

$$x^2 - 20x + 72 = 0.$$

Using the quadratic formula, we find that the critical points are

$$\begin{aligned} x &= \frac{20 \pm \sqrt{(-20)^2 - 4(1)(72)}}{2} \\ &= \frac{20 \pm \sqrt{112}}{2} \\ &= \frac{20 \pm 4\sqrt{7}}{2} \\ &= 10 \pm 2\sqrt{7} \end{aligned} .$$

Since $10 + 2\sqrt{7}$ is not in the domain of consideration, the only critical point we need to consider is $10 - 2\sqrt{7}$. Therefore, the volume is maximized if we let $x = 10 - 2\sqrt{7}$ in. The maximum volume is

$$V(10 - 2\sqrt{7}) = 640 + 448\sqrt{7} \approx 1825 \text{ in}^3.$$

as shown in the following graph.

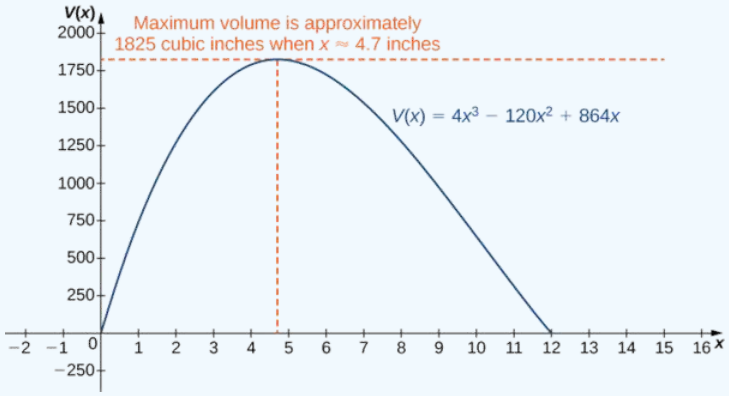


Figure 4.6.4: Maximizing the volume of the box leads to finding the maximum value of a cubic polynomial.

Exercise 4.6.2

Suppose the dimensions of the cardboard in Example 4.6.2 are 20 in. by 30 in. Let x be the side length of each square and write the volume of the open-top box as a function of x. Determine the domain of consideration for x.

Hint

The volume of the box is $L \cdot W \cdot H$.

Answer

$V(x) = x(20 - 2x)(30 - 2x)$. The domain is $[0, 10]$.

Example 4.6.3: Minimizing Travel Time

An island is 2 mi due north of its closest point along a straight shoreline. A visitor is staying at a cabin on the shore that is 6 mi west of that point. The visitor is planning to go from the cabin to the island. Suppose the visitor runs at a rate of 8 mph and swims at a rate of 3 mph. How far should the visitor run before swimming to minimize the time it takes to reach the island?

Solution

Step 1: Let x be the distance running and let y be the distance swimming (Figure 4.6.5). Let T be the time it takes to get from the cabin to the island.

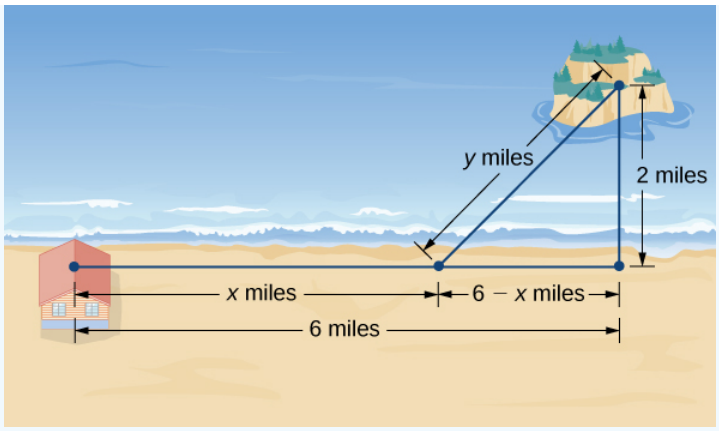


Figure 4.6.5: How can we choose x and y to minimize the travel time from the cabin to the island?

Step 2: The problem is to minimize T.

Step 3: To find the time spent traveling from the cabin to the island, add the time spent running and the time spent swimming. Since Distance = Rate × Time $(D = R \times T)$, the time spent running is

$$T_{running} = \frac{D_{running}}{R_{running}} = \frac{x}{8},$$

and the time spent swimming is

$$T_{swimming} = \frac{D_{swimming}}{R_{swimming}} = \frac{y}{3}.$$

Therefore, the total time spent traveling is

$$T = \frac{x}{8} + \frac{y}{3}.$$

Step 4: From Figure 4.6.5, the line segment of y miles forms the hypotenuse of a right triangle with legs of length 2 mi and $6 - x$ mi. Therefore, by the Pythagorean theorem, $2^2 + (6-x)^2 = y^2$, and we obtain $y = \sqrt{(6-x)^2 + 4}$. Thus, the total time spent traveling is given by the function

$$T(x) = \frac{x}{8} + \frac{\sqrt{(6-x)^2 + 4}}{3}.$$

Step 5: From Figure 4.6.5, we see that $0 \le x \le 6$. Therefore, $[0, 6]$ is the domain of consideration.

Step 6: Since $T(x)$ is a continuous function over a closed, bounded interval, it has a maximum and a minimum. Let's begin by looking for any critical points of T over the interval $[0, 6]$. The derivative is

$$\begin{aligned} T'(x) &= \frac{1}{8} - \frac{1}{2}\frac{[(6-x)^2 + 4]^{-1/2}}{3} \cdot 2(6-x) \\ &= \frac{1}{8} - \frac{(6-x)}{3\sqrt{(6-x)^2 + 4}} \end{aligned}$$

If $T'(x) = 0$,, then

$$\frac{1}{8} = \frac{6-x}{3\sqrt{(6-x)^2 + 4}} \tag{4.6.1}$$

Therefore,

$$3\sqrt{(6-x)^2 + 4} = 8(6-x). \tag{4.6.2}$$

Squaring both sides of this equation, we see that if x satisfies this equation, then x must satisfy

$$9[(6-x)^2 + 4] = 64(6-x)^2,$$

which implies

$$55(6-x)^2 = 36.$$

We conclude that if x is a critical point, then x satisfies

$$(x-6)^2 = \frac{36}{55}.$$

[Note that since we are squaring, $(x-6)^2 = (6-x)^2$.]

Therefore, the possibilities for critical points are

$$x = 6 \pm \frac{6}{\sqrt{55}}.$$

Since $x = 6 + 6/\sqrt{55}$ is not in the domain, it is not a possibility for a critical point. On the other hand, $x = 6 - 6/\sqrt{55}$ is in the domain. Since we squared both sides of Equation 4.6.2 to arrive at the possible critical points, it remains to verify that $x = 6 - 6/\sqrt{55}$ satisfies Equation 4.6.1. Since $x = 6 - 6/\sqrt{55}$ does satisfy that equation, we conclude that $x = 6 - 6/\sqrt{55}$ is a critical point, and it is the only one. To justify that the time is minimized for this value of x, we just need to check the values of $T(x)$ at the endpoints $x = 0$ and $x = 6$, and compare them with the value of $T(x)$ at the critical point $x = 6 - 6/\sqrt{55}$. We find that $T(0) \approx 2.108$ h and $T(6) \approx 1.417$ h, whereas

$$T(6 - 6/\sqrt{55}) \approx 1.368 \text{ h}.$$

Therefore, we conclude that T has a local minimum at $x \approx 5.19$ mi.

? Exercise 4.6.3

Suppose the island is 1 mi from shore, and the distance from the cabin to the point on the shore closest to the island is 15 mi. Suppose a visitor swims at the rate of 2.5 mph and runs at a rate of 6 mph. Let x denote the distance the visitor will run before swimming, and find a function for the time it takes the visitor to get from the cabin to the island.

Hint

The time $T = T_{running} + T_{swimming}$.

Answer

$$T(x) = \frac{x}{6} + \frac{\sqrt{(15-x)^2 + 1}}{2.5}$$

In business, companies are interested in **maximizing revenue.** In the following example, we consider a scenario in which a company has collected data on how many cars it is able to lease, depending on the price it charges its customers to rent a car. Let's use these data to determine the price the company should charge to maximize the amount of money it brings in.

✓ Example 4.6.4: Maximizing Revenue

Owners of a car rental company have determined that if they charge customers p dollars per day to rent a car, where $50 \le p \le 200$, the number of cars n they rent per day can be modeled by the linear function $n(p) = 1000 - 5p$. If they charge $\$50$ per day or less, they will rent all their cars. If they charge $\$200$ per day or more, they will not rent any cars. Assuming the owners plan to charge customers between $\$50$ per day and $\$200$ per day to rent a car, how much should they charge to maximize their revenue?

Solution

Step 1: Let p be the price charged per car per day and let n be the number of cars rented per day. Let R be the revenue per day.

Step 2: The problem is to maximize R.

Step 3: The revenue (per day) is equal to the number of cars rented per day times the price charged per car per day—that is, $R = n \times p$.

Step 4: Since the number of cars rented per day is modeled by the linear function $n(p) = 1000 - 5p$, the revenue R can be represented by the function

$$\begin{aligned} R(p) &= n \times p \\ &= (1000 - 5p)p \\ &= -5p^2 + 1000p. \end{aligned}$$

Step 5: Since the owners plan to charge between $\$50$ per car per day and $\$200$ per car per day, the problem is to find the maximum revenue $R(p)$ for p in the closed interval $[50, 200]$.

Step 6: Since R is a continuous function over the closed, bounded interval $[50, 200]$, it has an absolute maximum (and an absolute minimum) in that interval. To find the maximum value, look for critical points. The derivative is $R'(p) = -10p + 1000$. Therefore, the critical point is $p = 100$. When $p = 100, R(100) = \$50,000$. When $p = 50, R(p) = \$37,500$. When $p = 200, R(p) = \$0$.

Therefore, the absolute maximum occurs at $p = \$100$. The car rental company should charge $\$100$ per day per car to maximize revenue as shown in the following figure.

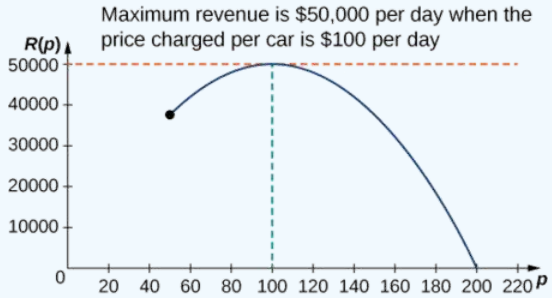


Figure 4.6.6: To maximize revenue, a car rental company has to balance the price of a rental against the number of cars people will rent at that price.

Exercise 4.6.4

A car rental company charges its customers p dollars per day, where $60 \leq p \leq 150$. It has found that the number of cars rented per day can be modeled by the linear function $n(p) = 750 - 5p$. How much should the company charge each customer to maximize revenue?

Hint

$R(p) = n \times p$, where n is the number of cars rented and p is the price charged per car.

Answer

The company should charge $\$75$ per car per day.

Example 4.6.5: Maximizing the Area of an Inscribed Rectangle

A rectangle is to be inscribed in the ellipse

$$\frac{x^2}{4} + y^2 = 1.$$

What should the dimensions of the rectangle be to maximize its area? What is the maximum area?

Solution

Step 1: For a rectangle to be inscribed in the ellipse, the sides of the rectangle must be parallel to the axes. Let L be the length of the rectangle and W be its width. Let A be the area of the rectangle.

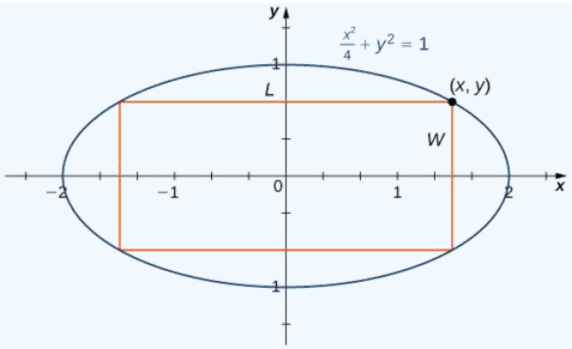


Figure 4.6.7: We want to maximize the area of a rectangle inscribed in an ellipse.

Step 2: The problem is to maximize A.

Step 3: The area of the rectangle is $A = LW$.

Step 4: Let (x, y) be the corner of the rectangle that lies in the first quadrant, as shown in Figure 4.6.7. We can write length $L = 2x$ and width $W = 2y$. Since $\frac{x^2}{4} + y^2 = 1$ and $y > 0$, we have $y = \sqrt{1 - \frac{x^2}{4}}$. Therefore, the area is

$$A = LW = (2x)(2y) = 4x\sqrt{1 - \frac{x^2}{4}} = 2x\sqrt{4 - x^2}$$

Step 5: From Figure 4.6.7, we see that to inscribe a rectangle in the ellipse, the x-coordinate of the corner in the first quadrant must satisfy $0 < x < 2$. Therefore, the problem reduces to looking for the maximum value of $A(x)$ over the open interval $(0, 2)$. Since $A(x)$ will have an absolute maximum (and absolute minimum) over the closed interval $[0, 2]$, we consider $A(x) = 2x\sqrt{4 - x^2}$ over the interval $[0, 2]$. If the absolute maximum occurs at an interior point, then we have found an absolute maximum in the open interval.

Step 6: As mentioned earlier, $A(x)$ is a continuous function over the closed, bounded interval $[0, 2]$. Therefore, it has an absolute maximum (and absolute minimum). At the endpoints $x = 0$ and $x = 2$, $A(x) = 0$. For $0 < x < 2$, $A(x) > 0$.

Therefore, the maximum must occur at a critical point. Taking the derivative of $A(x)$, we obtain

$$\begin{aligned} A'(x) &= 2\sqrt{4 - x^2} + 2x \cdot \frac{1}{2\sqrt{4 - x^2}}(-2x) \\ &= 2\sqrt{4 - x^2} - \frac{2x^2}{\sqrt{4 - x^2}} \\ &= \frac{8 - 4x^2}{\sqrt{4 - x^2}}. \end{aligned}$$

To find critical points, we need to find where $A'(x) = 0$. We can see that if x is a solution of

$$\frac{8 - 4x^2}{\sqrt{4 - x^2}} = 0, \tag{4.6.3}$$

then x must satisfy

$$8 - 4x^2 = 0.$$

Therefore, $x^2 = 2$. Thus, $x = \pm\sqrt{2}$ are the possible solutions of Equation 4.6.3. Since we are considering x over the interval $[0, 2]$, $x = \sqrt{2}$ is a possibility for a critical point, but $x = -\sqrt{2}$ is not. Therefore, we check whether $\sqrt{2}$ is a solution of Equation 4.6.3. Since $x = \sqrt{2}$ is a solution of Equation 4.6.3, we conclude that $\sqrt{2}$ is the only critical point of $A(x)$ in the interval $[0, 2]$.

Therefore, $A(x)$ must have an absolute maximum at the critical point $x = \sqrt{2}$. To determine the dimensions of the rectangle, we need to find the length L and the width W. If $x = \sqrt{2}$ then

$$y = \sqrt{1 - \frac{(\sqrt{2})^2}{4}} = \sqrt{1 - \frac{1}{2}} = \frac{1}{\sqrt{2}}.$$

Therefore, the dimensions of the rectangle are $L = 2x = 2\sqrt{2}$ and $W = 2y = \dfrac{2}{\sqrt{2}} = \sqrt{2}$. The area of this rectangle is $A = LW = (2\sqrt{2})(\sqrt{2}) = 4$.

Exercise 4.6.5

Modify the area function A if the rectangle is to be inscribed in the unit circle $x^2 + y^2 = 1$. What is the domain of consideration?

Hint

If (x, y) is the vertex of the square that lies in the first quadrant, then the area of the square is $A = (2x)(2y) = 4xy$.

Answer

$A(x) = 4x\sqrt{1 - x^2}$. The domain of consideration is $[0, 1]$.

Solving Optimization Problems when the Interval Is Not Closed or Is Unbounded

In the previous examples, we considered functions on closed, bounded domains. Consequently, by the extreme value theorem, we were guaranteed that the functions had absolute extrema. Let's now consider functions for which the domain is neither closed nor bounded.

Many functions still have at least one absolute extrema, even if the domain is not closed or the domain is unbounded. For example, the function $f(x) = x^2 + 4$ over $(-\infty, \infty)$ has an absolute minimum of 4 at $x = 0$. Therefore, we can still consider functions over unbounded domains or open intervals and determine whether they have any absolute extrema. In the next example, we try to minimize a function over an unbounded domain. We will see that, although the domain of consideration is $(0, \infty)$, the function has an absolute minimum.

In the following example, we look at constructing a box of least surface area with a prescribed volume. It is not difficult to show that for a closed-top box, by symmetry, among all boxes with a specified volume, a cube will have the smallest surface area. Consequently, we consider the modified problem of determining which open-topped box with a specified volume has the smallest surface area.

Example 4.6.6: Minimizing Surface Area

A rectangular box with a square base, an open top, and a volume of $216\ \text{in}^3$ is to be constructed. What should the dimensions of the box be to minimize the surface area of the box? What is the minimum surface area?

Solution

Step 1: Draw a rectangular box and introduce the variable x to represent the length of each side of the square base; let y represent the height of the box. Let S denote the surface area of the open-top box.

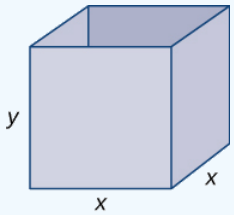


Figure 4.6.8: We want to minimize the surface area of a square-based box with a given volume.

Step 2: We need to minimize the surface area. Therefore, we need to minimize S.

Step 3: Since the box has an open top, we need only determine the area of the four vertical sides and the base. The area of each of the four vertical sides is $x \cdot y$. The area of the base is x^2. Therefore, the surface area of the box is

$$S = 4xy + x^2.$$

Step 4: Since the volume of this box is x^2y and the volume is given as $216\,\text{in}^3$, the constraint equation is

$$x^2y = 216.$$

Solving the constraint equation for y, we have $y = \dfrac{216}{x^2}$. Therefore, we can write the surface area as a function of x only:

$$S(x) = 4x\left(\frac{216}{x^2}\right) + x^2.$$

Therefore, $S(x) = \dfrac{864}{x} + x^2$.

Step 5: Since we are requiring that $x^2y = 216$, we cannot have $x = 0$. Therefore, we need $x > 0$. On the other hand, x is allowed to have any positive value. Note that as x becomes large, the height of the box y becomes correspondingly small so that $x^2y = 216$. Similarly, as x becomes small, the height of the box becomes correspondingly large. We conclude that the domain is the open, unbounded interval $(0, \infty)$. Note that, unlike the previous examples, we cannot reduce our problem to looking for an absolute maximum or absolute minimum over a closed, bounded interval. However, in the next step, we discover why this function must have an absolute minimum over the interval $(0, \infty)$.

Step 6: Note that as $x \to 0^+$, $S(x) \to \infty$. Also, as $x \to \infty$, $S(x) \to \infty$. Since S is a continuous function that approaches infinity at the ends, it must have an absolute minimum at some $x \in (0, \infty)$. This minimum must occur at a critical point of S. The derivative is

$$S'(x) = -\frac{864}{x^2} + 2x.$$

Therefore, $S'(x) = 0$ when $2x = \dfrac{864}{x^2}$. Solving this equation for x, we obtain $x^3 = 432$, so $x = \sqrt[3]{432} = 6\sqrt[3]{2}$. Since this is the only critical point of S, the absolute minimum must occur at $x = 6\sqrt[3]{2}$ (see Figure 4.6.9).

When $x = 6\sqrt[3]{2}$, $y = \dfrac{216}{(6\sqrt[3]{2})^2} = 3\sqrt[3]{2}\,\text{in}$. Therefore, the dimensions of the box should be $x = 6\sqrt[3]{2}\,\text{in}$. and $y = 3\sqrt[3]{2}\,\text{in}$. With these dimensions, the surface area is

$$S(6\sqrt[3]{2}) = \frac{864}{6\sqrt[3]{2}} + (6\sqrt[3]{2})^2 = 108\sqrt[3]{4}\,\text{in}^2$$

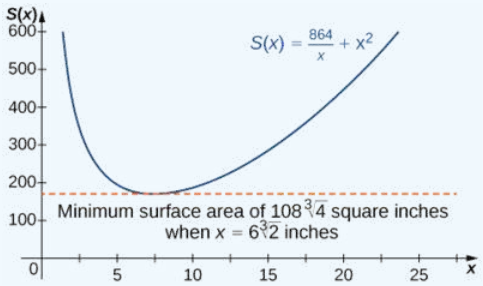


Figure 4.6.9: We can use a graph to determine the dimensions of a box of given the volume and the minimum surface area.

Exercise 4.6.6

Consider the same open-top box, which is to have volume $216\,\text{in}^3$. Suppose the cost of the material for the base is $20¢/\text{in}^2$ and the cost of the material for the sides is $30¢/\text{in}^2$ and we are trying to minimize the cost of this box. Write the cost as a function of the side lengths of the base. (Let x be the side length of the base and y be the height of the box.)

Hint

If the cost of one of the sides is $30¢/\text{in}^2$,the cost of that side is $0.30xy$dollars.

Answer

$$c(x) = \frac{259.2}{x} + 0.2x^2 \text{ dollars}$$

Key Concepts

- To solve an optimization problem, begin by drawing a picture and introducing variables.
- Find an equation relating the variables.
- Find a function of one variable to describe the quantity that is to be minimized or maximized.
- Look for critical points to locate local extrema.

Glossary

optimization problems

problems that are solved by finding the maximum or minimum value of a function

---

4.6: Applied Optimization Problems is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **4.7: Applied Optimization Problems** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

4.7: Newton's Method

Learning Objectives

- Describe the steps of Newton's method.
- Explain what an iterative process means.
- Recognize when Newton's method does not work.
- Apply iterative processes to various situations.

In many areas of pure and applied mathematics, we are interested in finding solutions to an equation of the form $f(x)=0$. For most functions, however, it is difficult—if not impossible—to calculate their zeroes explicitly. In this section, we take a look at a technique that provides a very efficient way of approximating the **zeroes of functions**. This technique makes use of tangent line approximations and is behind the method used often by calculators and computers to find zeroes.

Describing Newton's Method

Consider the task of finding the solutions of $f(x)=0$. If f is the first-degree polynomial $f(x)=ax+b$, then the solution of $f(x)=0$ is given by the formula $x=-\frac{b}{a}$. If f is the second-degree polynomial $f(x)=ax^2+bx+c$, the solutions of $f(x)=0$ can be found by using the quadratic formula. However, for polynomials of degree **3** or more, finding roots of f becomes more complicated. Although formulas exist for third- and fourth-degree polynomials, they are quite complicated. Also, if f is a polynomial of degree **5** or greater, it is known that no such formulas exist. For example, consider the function

$$f(x)=x^5+8x^4+4x^3-2x-7.$$

No formula exists that allows us to find the solutions of $f(x)=0$. Similar difficulties exist for nonpolynomial functions. For example, consider the task of finding solutions of $\tan(x)-x=0$. No simple formula exists for the solutions of this equation. In cases such as these, we can use Newton's method to approximate the roots.

Newton's method makes use of the following idea to approximate the solutions of $f(x)=0$. By sketching a graph of f, we can estimate a root of $f(x)=0$. Let's call this estimate x_0. We then draw the tangent line to f at x_0. If $f\,'(x_0)\neq 0$, this tangent line intersects the x-axis at some point $(x_1,0)$. Now let x_1 be the next approximation to the actual root. Typically, x_1 is closer than x_0 to an actual root. Next we draw the tangent line to f at x_1. If $f\,'(x_1)\neq 0$, this tangent line also intersects the x-axis, producing another approximation, x_2. We continue in this way, deriving a list of approximations: x_0, x_1, x_2, ... Typically, the numbers x_0, x_1, x_2, ... quickly approach an actual root x^*, as shown in the following figure.

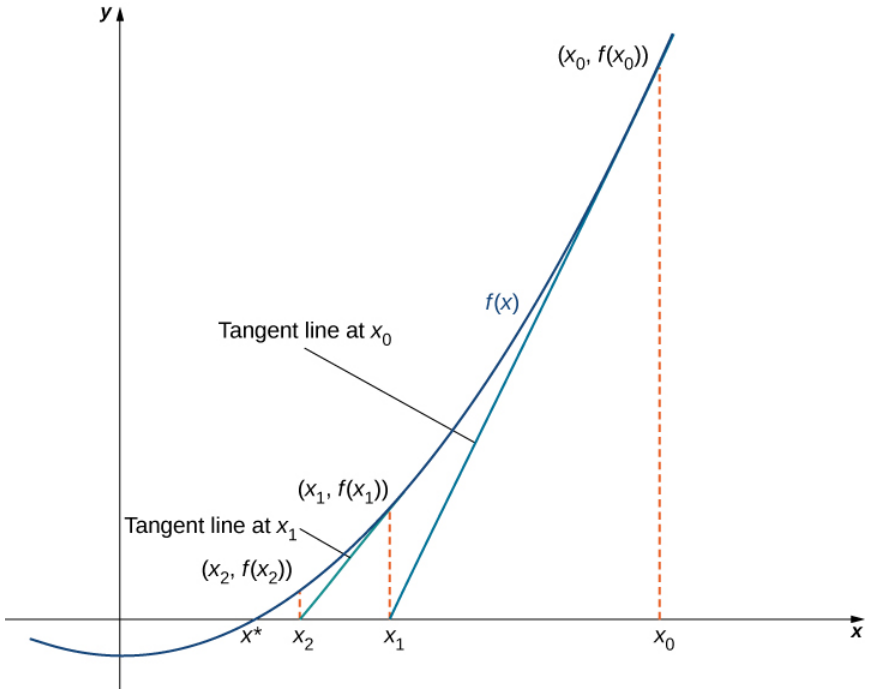


Figure 4.7.1:The approximations $x_0, x_1, x_2, \ldots$ approach the actual root x^*. The approximations are derived by looking at tangent lines to the graph of f.

Now let's look at how to calculate the approximations $x_0, x_1, x_2, \ldots$. If x_0 is our first approximation, the approximation x_1 is defined by letting $(x_1, 0)$ be the x-intercept of the tangent line to f at x_0. The equation of this tangent line is given by

$$y = f(x_0) + f'(x_0)(x - x_0).$$

Therefore, x_1 must satisfy

$$f(x_0) + f'(x_0)(x_1 - x_0) = 0.$$

Solving this equation for x_1, we conclude that

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Similarly, the point $(x_2, 0)$ is the x-intercept of the tangent line to f at x_1. Therefore, x_2 satisfies the equation

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

In general, for $n > 0$, x_n satisfies

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}. \tag{4.7.1}$$

Next we see how to make use of this technique to approximate the root of the polynomial $f(x) = x^3 - 3x + 1$.

Example 4.7.1: Finding a Root of a Polynomial

Use Newton's method to approximate a root of $f(x) = x^3 - 3x + 1$ in the interval $[1, 2]$. Let $x_0 = 2$ and find x_1, x_2, x_3, x_4, and x_5.

Solution

From Figure 4.7.2, we see that f has one root over the interval $[1, 2]$. Therefore $x_0 = 2$ seems like a reasonable first approximation. To find the next approximation, we use Equation 4.7.1. Since $f(x) = x^3 - 3x + 1$, the derivative is $f'(x) = 3x^2 - 3$. Using Equation 4.7.1 with $n = 1$ (and a calculator that displays 10 digits), we obtain

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 2 - \frac{f(2)}{f'(2)} = 2 - \frac{3}{9} \approx 1.666666667.$$

To find the next approximation, x_2, we use Equation 4.7.1 with $n = 2$ and the value of x_1 stored on the calculator. We find that

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \approx 1.548611111.$$

Continuing in this way, we obtain the following results:

- $x_1 \approx 1.666666667$
- $x_2 \approx 1.548611111$
- $x_3 \approx 1.532390162$
- $x_4 \approx 1.532088989$
- $x_5 \approx 1.532088886$
- $x_6 \approx 1.532088886.$

We note that we obtained the same value for x_5 and x_6. Therefore, any subsequent application of Newton's method will most likely give the same value for x_n.

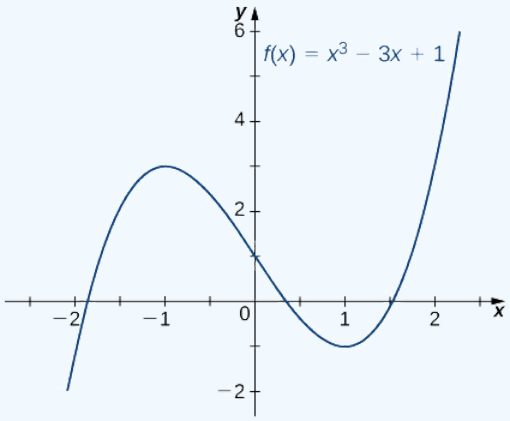


Figure 4.7.2: The function $f(x) = x^3 - 3x + 1$ has one root over the interval $[1, 2]$.

? Exercise 4.7.1

Letting $x_0 = 0$, let's use Newton's method to approximate the root of $f(x) = x^3 - 3x + 1$ over the interval $[0, 1]$ by calculating x_1 and x_2.

Hint

Use Equation 4.7.1.

Answer

$x_1 \approx 0.33333333$
$x_2 \approx 0.347222222$

Newton's method can also be used to approximate square roots. Here we show how to approximate $\sqrt{2}$. This method can be modified to approximate the square root of any positive number.

✔ Example 4.7.2: Finding a Square Root

Use Newton's method to approximate $\sqrt{2}$ (Figure 4.7.3). Let $f(x) = x^2 - 2$, let $x_0 = 2$, and calculate x_1, x_2, x_3, x_4, x_5. (We note that since $f(x) = x^2 - 2$ has a zero at $\sqrt{2}$, the initial value $x_0 = 2$ is a reasonable choice to approximate $\sqrt{2}$).

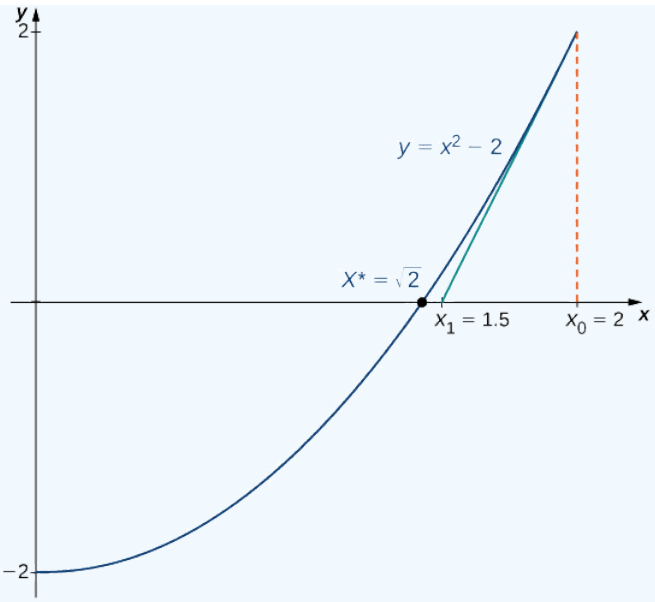


Figure 4.7.3: We can use Newton's method to find $\sqrt{2}$.

Solution

For $f(x)=x^2-2,\ f'(x)=2x$. From Equation 4.7.1, we know that

$$\begin{aligned} x_n &= x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \\ &= x_{n-1} - \frac{x_{n-1}^2 - 2}{2x_{n-1}} \\ &= \frac{1}{2}x_{n-1} + \frac{1}{x_{n-1}} \\ &= \frac{1}{2}\left(x_{n-1} + \frac{2}{x_{n-1}}\right). \end{aligned}$$

Therefore,

$$x_1 = \tfrac{1}{2}\left(x_0 + \tfrac{2}{x_0}\right) = \tfrac{1}{2}\left(2 + \tfrac{2}{2}\right) = 1.5$$

$$x_2 = \tfrac{1}{2}\left(x_1 + \tfrac{2}{x_1}\right) = \tfrac{1}{2}\left(1.5 + \tfrac{2}{1.5}\right) \approx 1.416666667.$$

Continuing in this way, we find that

$$\begin{aligned} x_1 &= 1.5 \\ x_2 &\approx 1.416666667 \\ x_3 &\approx 1.414215686 \\ x_4 &\approx 1.414213562 \\ x_5 &\approx 1.414213562. \end{aligned}$$

Since we obtained the same value for x_4 and x_5, it is unlikely that the value x_n will change on any subsequent application of Newton's method. We conclude that $\sqrt{2} \approx 1.414213562$.

? Exercise 4.7.2

Use Newton's method to approximate $\sqrt{3}$ by letting $f(x)=x^2-3$ and $x_0=3$. Find x_1 and x_2.

Hint

For $f(x) = x^2 - 3$, Equation 4.7.1reduces to $x_n = \frac{x_{n-1}}{2} + \frac{3}{2x_{n-1}}$.

Answer

$x_1 = 2$
$x_2 = 1.75$

When using Newton's method, each approximation after the initial guess is defined in terms of the previous approximation by using the same formula. In particular, by defining the function $F(x) = x - \left[\frac{f(x)}{f'(x)}\right]$, we can rewrite Equation 4.7.1 as $x_n = F(x_{n-1})$. This type of process, where each x_n is defined in terms of x_{n-1} by repeating the same function, is an example of an iterative process. Shortly, we examine other iterative processes. First, let's look at the reasons why Newton's method could fail to find a root.

Failures of Newton's Method

Typically, Newton's method is used to find roots fairly quickly. However, things can go wrong. Some reasons why Newton's method might fail include the following:

1. At one of the approximations x_n, the derivative f' is zero at x_n, but $f(x_n) \neq 0$. As a result, the tangent line of f at x_n does not intersect the x-axis. Therefore, we cannot continue the iterative process.
2. The approximations $x_0, x_1, x_2, \ldots$ may approach a different root. If the function f has more than one root, it is possible that our approximations do not approach the one for which we are looking, but approach a different root (see Figure 4.7.4). This event most often occurs when we do not choose the approximation x_0 close enough to the desired root.
3. The approximations may fail to approach a root entirely. In Example 4.7.3, we provide an example of a function and an initial guess x_0 such that the successive approximations never approach a root because the successive approximations continue to alternate back and forth between two values.

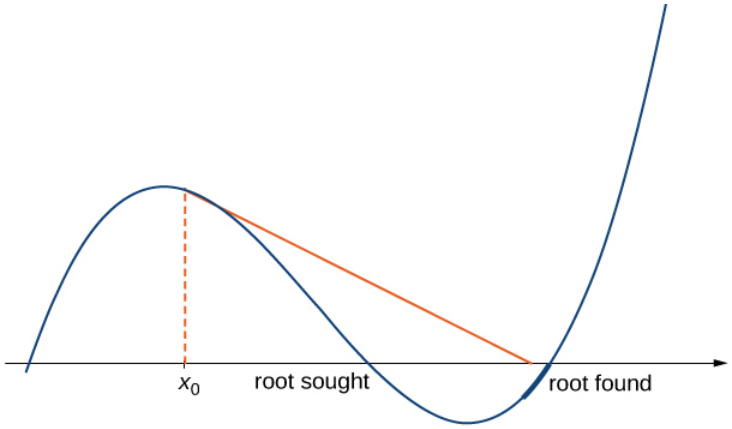


Figure 4.7.4: If the initial guess x_0 is too far from the root sought, it may lead to approximations that approach a different root.

Example 4.7.3: When Newton's Method Fails

Consider the function $f(x) = x^3 - 2x + 2$. Let $x_0 = 0$. Show that the sequence $x_1, x_2, \ldots$ fails to approach a root of f.

Solution

For $f(x) = x^3 - 2x + 2$, the derivative is $f'(x) = 3x^2 - 2$ .Therefore,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 0 - \frac{f(0)}{f'(0)} = -\frac{2}{-2} = 1.$$

In the next step,

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 1 - \frac{f(1)}{f'(1)} = 1 - \frac{1}{1} = 0.$$

Consequently, the numbers $x_0, x_1, x_2, \ldots$ continue to bounce back and forth between 0 and 1 and never get closer to the root of f which is over the interval $[-2, -1]$ (Figure 4.7.5). Fortunately, if we choose an initial approximation x_0 closer to the actual root, we can avoid this situation.

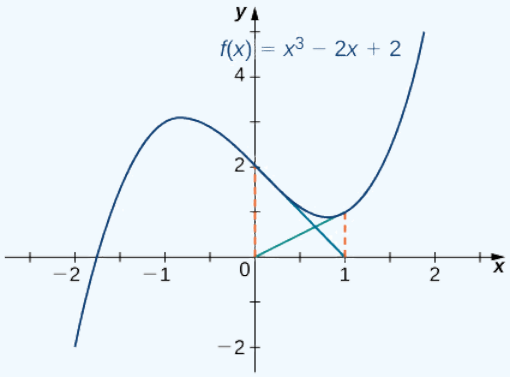


Figure 4.7.5: The approximations continue to alternate between 0 and 1 and never approach the root of f.

? Exercise 4.7.3

For $f(x) = x^3 - 2x + 2$, let $x_0 = -1.5$ and find x_1 and x_2.

Hint

Use Equation 4.7.1.

Answer

$x_1 \approx -1.842105263$
$x_2 \approx -1.772826920$

From Example 4.7.3, we see that Newton's method does not always work. However, when it does work, the sequence of approximations approaches the root very quickly. Discussions of how quickly the sequence of approximations approach a root found using Newton's method are included in texts on numerical analysis.

Other Iterative Processes

As mentioned earlier, Newton's method is a type of iterative process. We now look at an example of a different type of iterative process.

Consider a function F and an initial number x_0. Define the subsequent numbers x_n by the formula $x_n = F(x_{n-1})$. This process is an iterative process that creates a list of numbers $x_0, x_1, x_2, \ldots, x_n, \ldots$. This list of numbers may approach a finite number x^* as n gets larger, or it may not. In Example 4.7.4, we see an example of a function F and an initial guess x_0 such that the resulting list of numbers approaches a finite value.

✓ Example 4.7.4: Finding a Limit for an Iterative Process

Let $F(x) = \frac{1}{2}x + 4$ and let $x_0 = 0$. For all $n \geq 1$, let $x_n = F(x_{n-1})$. Find the values x_1, x_2, x_3, x_4, x_5. Make a conjecture about what happens to this list of numbers $x_1, x_2, x_3, \ldots, x_n, \ldots$ as $n \to \infty$. If the list of numbers $x_1, x_2, x_3, \ldots$ approaches a finite number x^*, then x^* satisfies $x^* = F(x^*)$, and x^* is called a fixed point of F.

Solution

If $x_0 = 0$, then

- $x_1 = \frac{1}{2}(0) + 4 = 4$
- $x_2 = \frac{1}{2}(4) + 4 = 6$
- $x_3 = \frac{1}{2}(6) + 4 = 7$
- $x_4 = \frac{1}{2}(7) + 4 = 7.5$

- $x_5 = \frac{1}{2}(7.5) + 4 = 7.75$
- $x_6 = \frac{1}{2}(7.75) + 4 = 7.875$
- $x_7 = \frac{1}{2}(7.875) + 4 = 7.9375$
- $x_8 = \frac{1}{2}(7.9375) + 4 = 7.96875$
- $x_9 = \frac{1}{2}(7.96875) + 4 = 7.984375.$

From this list, we conjecture that the values x_n approach 8.

Figure 4.7.6 provides a graphical argument that the values approach 8 as $n \to \infty$. Starting at the point (x_0, x_0), we draw a vertical line to the point $(x_0, F(x_0))$. The next number in our list is $x_1 = F(x_0)$. We use x_1 to calculate x_2. Therefore, we draw a horizontal line connecting (x_0, x_1) to the point (x_1, x_1) on the line $y = x$, and then draw a vertical line connecting (x_1, x_1) to the point $(x_1, F(x_1))$. The output $F(x_1)$ becomes x_2. Continuing in this way, we could create an infinite number of line segments. These line segments are trapped between the lines $F(x) = \frac{x}{2} + 4$ and $y = x$. The line segments get closer to the intersection point of these two lines, which occurs when $x = F(x)$. Solving the equation $x = \frac{x}{2} + 4$, we conclude they intersect at $x = 8$. Therefore, our graphical evidence agrees with our numerical evidence that the list of numbers $x_0, x_1, x_2, \ldots$ approaches $x^* = 8$ as $n \to \infty$.

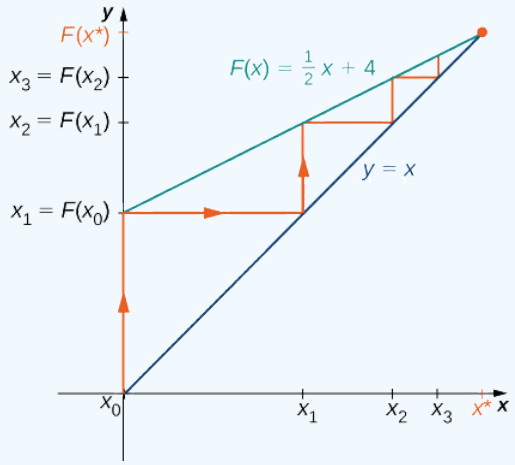


Figure 4.7.6: This iterative process approaches the value $x^* = 8$.

Exercise 4.7.4

Consider the function $F(x) = \frac{1}{3}x + 6$. Let $x_0 = 0$ and let $x_n = F(x_{n-1})$ for $n \geq 2$. Find x_1, x_2, x_3, x_4, x_5. Make a conjecture about what happens to the list of numbers $x_1, x_2, x_3, \ldots x_n, \ldots$ as $n \to \infty$.

Hint

Consider the point where the lines $y = x$ and $y = F(x)$ intersect.

Answer

$x_1 = 6,\ x_2 = 8,\ x_3 = \frac{26}{3},\ x_4 = \frac{80}{9},\ x_5 = \frac{242}{27};\ x^* = 9$

Iterative Processes and Chaos

Iterative processes can yield some very interesting behavior. In this section, we have seen several examples of iterative processes that converge to a fixed point. We also saw in Example 4.7.3 that the iterative process bounced back and forth between two values. We call this kind of behavior a 2-cycle. Iterative processes can converge to cycles with various periodicities, such as 2−cycles, 4−cycles (where the iterative process repeats a sequence of four values), 8-cycles, and so on.

Some iterative processes yield what mathematicians call chaos. In this case, the iterative process jumps from value to value in a seemingly random fashion and never converges or settles into a cycle. Although a complete exploration of chaos is beyond the scope of this text, in this project we look at one of the key properties of a chaotic iterative process: sensitive dependence on

initial conditions. This property refers to the concept that small changes in initial conditions can generate drastically different behavior in the iterative process.

Probably the best-known example of chaos is the **Mandelbrot set** (see Figure 4.7.7), named after Benoit Mandelbrot (1924–2010), who investigated its properties and helped popularize the field of chaos theory. The Mandelbrot set is usually generated by computer and shows fascinating details on enlargement, including self-replication of the set. Several colorized versions of the set have been shown in museums and can be found online and in popular books on the subject.

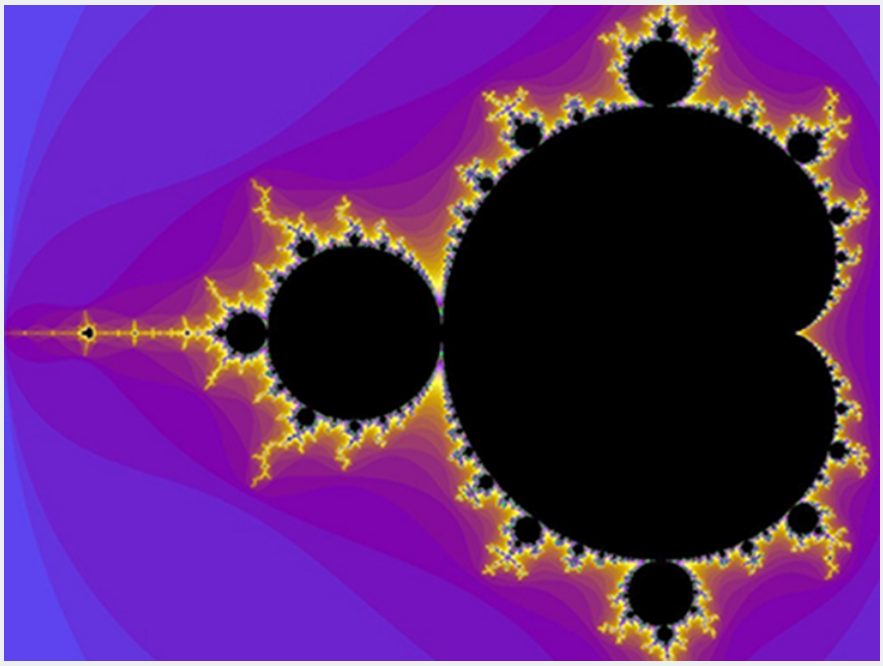


Figure 4.7.7: The Mandelbrot set is a well-known example of a set of points generated by the iterative chaotic behavior of a relatively simple function.

In this project we use the logistic map

$$f(x) = rx(1-x)$$

$$\text{where } x \in [0,1] \text{ and } r > 0$$

as the function in our iterative process. The logistic map is a deceptively simple function; but, depending on the value of r, the resulting iterative process displays some very interesting behavior. It can lead to fixed points, cycles, and even chaos.

To visualize the long-term behavior of the iterative process associated with the logistic map, we will use a tool called a cobweb diagram. As we did with the iterative process we examined earlier in this section, we first draw a vertical line from the point $(x_0, 0)$ to the point $(x_0, f(x_0)) = (x_0, x_1)$. We then draw a horizontal line from that point to the point (x_1, x_1), then draw a vertical line to $(x_1, f(x_1)) = (x_1, x_2)$, and continue the process until the long-term behavior of the system becomes apparent. Figure 4.7.8 shows the long-term behavior of the logistic map when $r = 3.55$ and $x_0 = 0.2$. (The first 100 iterations are not plotted.) The long-term behavior of this iterative process is an 8-cycle.

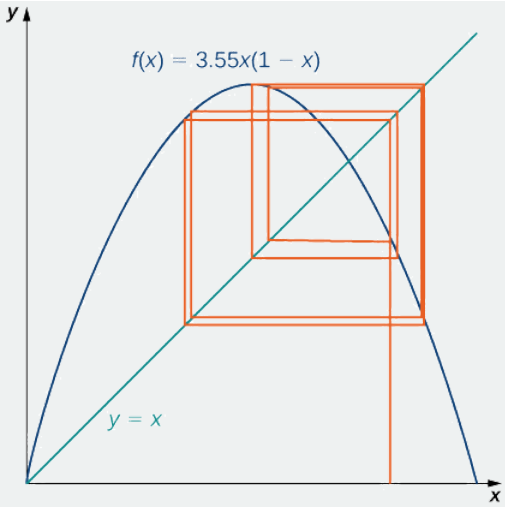


Figure 4.7.8: A cobweb diagram for $f(x) = 3.55x(1-x)$ is presented here. The sequence of values results in an 8-cycle.

1. Let $r = 0.5$ and choose $x_0 = 0.2$. Either by hand or by using a computer, calculate the first 10 values in the sequence. Does the sequence appear to converge? If so, to what value? Does it result in a cycle? If so, what kind of cycle (for example, $2-$cycle, $4-$cycle.)?
2. What happens when $r = 2$?
3. For $r = 3.2$ and $r = 3.5$, calculate the first 100 sequence values. Generate a cobweb diagram for each iterative process. (Several free applets are available online that generate cobweb diagrams for the logistic map.) What is the long-term behavior in each of these cases?
4. Now let $r = 4$. Calculate the first 100 sequence values and generate a cobweb diagram. What is the long-term behavior in this case?
5. Repeat the process for $r = 4$, but let $x_0 = 0.201$. How does this behavior compare with the behavior for $x_0 = 0.2$?

Key Concepts

- Newton's method approximates roots of $f(x) = 0$ by starting with an initial approximation x_0, then uses tangent lines to the graph of f to create a sequence of approximations $x_1, x_2, x_3, \ldots$.
- Typically, Newton's method is an efficient method for finding a particular root. In certain cases, Newton's method fails to work because the list of numbers $x_0, x_1, x_2, \ldots$ does not approach a finite value or it approaches a value other than the root sought.
- Any process in which a list of numbers $x_0, x_1, x_2, \ldots$ is generated by defining an initial number x_0 and defining the subsequent numbers by the equation $x_n = F(x_{n-1})$ for some function F is an iterative process. Newton's method is an example of an iterative process, where the function $F(x) = x - \left[\frac{f(x)}{f'(x)}\right]$ for a given function f.

Glossary

iterative process

process in which a list of numbers $x_0, x_1, x_2, x_3 \ldots$ is generated by starting with a number x_0 and defining $x_n = F(x_{n-1})$ for $n \geq 1$

Newton's method

method for approximating roots of $f(x) = 0$; using an initial guess x_0; each subsequent approximation is defined by the equation $x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$

4.7: Newton's Method is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **4.9: Newton's Method** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

CHAPTER OVERVIEW

5: Integration

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

5.1: Antiderivatives
5.2: Approximating Areas with Limits of Finite Sums
5.3: The Definite Integral
5.4: The Fundamental Theorem of Calculus
5.5: Substitution with Indefinite Integrals
5.6: Substitution with Definite Integrals

5: Integration is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

5.1: Antiderivatives

Learning Objectives

- Find the general antiderivative of a given function.
- Explain the terms and notation used for an indefinite integral.
- State the power rule for integrals.
- Use antidifferentiation to solve simple initial-value problems.

At this point, we have seen how to calculate derivatives of many functions and have been introduced to a variety of their applications. We now ask a question that turns this process around: Given a function f, how do we find a function with the derivative f and why would we be interested in such a function?

We answer the first part of this question by defining antiderivatives. The antiderivative of a function f is a function with a derivative f. Why are we interested in antiderivatives? The need for antiderivatives arises in many situations, and we look at various examples throughout the remainder of the text. Here we examine one specific example that involves rectilinear motion. In our examination in Derivatives of rectilinear motion, we showed that given a position function $s(t)$ of an object, then its velocity function $v(t)$ is the derivative of $s(t)$—that is, $v(t) = s'(t)$. Furthermore, the acceleration $a(t)$ is the derivative of the velocity $v(t)$—that is, $a(t) = v'(t) = s''(t)$. Now suppose we are given an acceleration function a, but not the velocity function v or the position function s. Since $a(t) = v'(t)$, determining the velocity function requires us to find an antiderivative of the acceleration function. Then, since $v(t) = s'(t)$, determining the position function requires us to find an antiderivative of the velocity function. Rectilinear motion is just one case in which the need for antiderivatives arises. We will see many more examples throughout the remainder of the text. For now, let's look at the terminology and notation for antiderivatives, and determine the antiderivatives for several types of functions. We examine various techniques for finding antiderivatives of more complicated functions later in the text (Introduction to Techniques of Integration).

The Reverse of Differentiation

At this point, we know how to find derivatives of various functions. We now ask the opposite question. Given a function f, how can we find a function with derivative f? If we can find a function F with derivative f, we call F an antiderivative of f.

Definition: Antiderivative

A function F is an **antiderivative** of the function f if

$$F'(x) = f(x)$$

for all x in the domain of f.

Consider the function $f(x) = 2x$. Knowing the power rule of differentiation, we conclude that $F(x) = x^2$ is an antiderivative of f since $F'(x) = 2x$.

Are there any other antiderivatives of f?

Yes; since the derivative of any constant C is zero, $x^2 + C$ is also an antiderivative of $2x$. Therefore, $x^2 + 5$ and $x^2 - \sqrt{2}$ are also antiderivatives.

Are there any others that are not of the form $x^2 + C$ for some constant C?

The answer is no. From Corollary 2 of the Mean Value Theorem, we know that if F and G are differentiable functions such that $F'(x) = G'(x)$, then $F(x) - G(x) = C$ for some constant C. This fact leads to the following important theorem.

Theorem 5.1.1: General Form of an Antiderivative

Let F be an antiderivative of f over an interval I. Then,

I. for each constant C, the function $F(x) + C$ is also an antiderivative of f over I;
II. if G is an antiderivative of f over I, there is a constant C for which $G(x) = F(x) + C$ over I.

In other words, the most general form of the antiderivative of f over I is $F(x) + C$.

We use this fact and our knowledge of derivatives to find all the antiderivatives for several functions.

Example 5.1.1: Finding Antiderivatives

For each of the following functions, find all antiderivatives.

a. $f(x) = 3x^2$
b. $f(x) = \dfrac{1}{x}$
c. $f(x) = \cos x$
d. $f(x) = e^x$

Solution:

a. Because

$$\frac{d}{dx}\left(x^3\right) = 3x^2$$

then $F(x) = x^3$ is an antiderivative of $3x^2$. Therefore, every antiderivative of $3x^2$ is of the form $x^3 + C$ for some constant C, and every function of the form $x^3 + C$ is an antiderivative of $3x^2$.

b. Let $f(x) = \ln|x|$.

For $x > 0,\ f(x) = \ln|x| = \ln(x)$ and

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

For $x < 0,\ f(x) = \ln|x| = \ln(-x)$ and

$$\frac{d}{dx}\left(\ln(-x)\right) = -\frac{1}{-x} = \frac{1}{x}.$$

Therefore,

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

Thus, $F(x) = \ln|x|$ is an antiderivative of $\dfrac{1}{x}$. Therefore, every antiderivative of $\dfrac{1}{x}$ is of the form $\ln|x| + C$ for some constant C and every function of the form $\ln|x| + C$ is an antiderivative of $\dfrac{1}{x}$.

c. We have

$$\frac{d}{dx}\left(\sin x\right) = \cos x,$$

so $F(x) = \sin x$ is an antiderivative of $\cos x$. Therefore, every antiderivative of $\cos x$ is of the form $\sin x + C$ for some constant C and every function of the form $\sin x + C$ is an antiderivative of $\cos x$.

d. Since

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

then $F(x) = e^x$ is an antiderivative of e^x. Therefore, every antiderivative of e^x is of the form $e^x + C$ for some constant C and every function of the form $e^x + C$ is an antiderivative of e^x.

Exercise 5.1.1

Find all antiderivatives of $f(x) = \sin x$.

Hint

What function has a derivative of $\sin x$?

Answer

$F(x) = -\cos x + C$

Indefinite Integrals

We now look at the formal notation used to represent antiderivatives and examine some of their properties. These properties allow us to find antiderivatives of more complicated functions. Given a function f, we use the notation $f'(x)$ or $\frac{df}{dx}$ to denote the derivative of f. Here we introduce notation for antiderivatives. If F is an antiderivative of f, we say that $F(x)+C$ is the most general antiderivative of f and write

$$\int f(x)\,dx = F(x)+C.$$

The symbol $\int$ is called an integral sign, and $\int f(x)\,dx$ is called the indefinite integral of f.

Definition: Indefinite Integrals

Given a function f, the indefinite integral of f, denoted

$$\int f(x)\,dx,$$

is the most general antiderivative of f. If F is an antiderivative of f, then

$$\int f(x)\,dx = F(x)+C.$$

The expression $f(x)$ is called the integrand and the variable x is the variable of integration.

Given the terminology introduced in this definition, the act of finding the antiderivatives of a function f is usually referred to as integrating f.

For a function f and an antiderivative F, the functions $F(x)+C$, where C is any real number, is often referred to as the family of antiderivatives of f. For example, since x^2 is an antiderivative of $2x$ and any antiderivative of $2x$ is of the form x^2+C, we write

$$\int 2x\,dx = x^2+C.$$

The collection of all functions of the form x^2+C, where C is any real number, is known as the family of antiderivatives of $2x$. Figure 5.1.1 shows a graph of this family of antiderivatives.

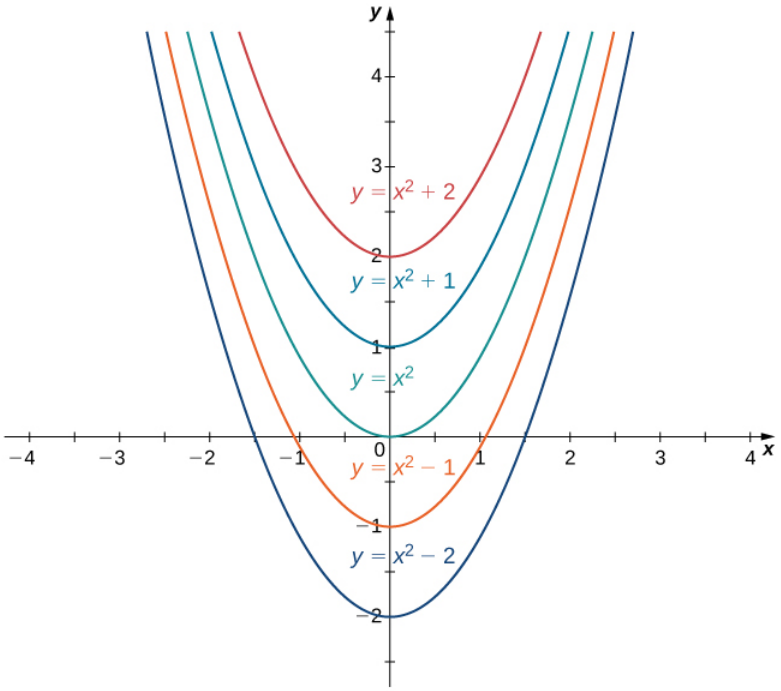


Figure 5.1.1: The family of antiderivatives of $2x$ consists of all functions of the form $x^2 + C$, where C is any real number.

For some functions, evaluating indefinite integrals follows directly from properties of derivatives. For example, for $n \neq -1$,

$$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C,$$

which comes directly from

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

This fact is known as the power rule for integrals.

Power Rule for Integrals

For $n \neq -1$,

$$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C.$$

Evaluating indefinite integrals for some other functions is also a straightforward calculation. The following table lists the indefinite integrals for several common functions. A more complete list appears in Appendix B.

Table 5.1.1: Integration Formulas

Differentiation Formula	Indefinite Integral
$\frac{d}{dx}\left(k\right) = 0$	$\int k \, dx = \int kx^0 \, dx = kx + C$
$\frac{d}{dx}\left(x^n\right) = nx^{n-1}$	$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C$ for $n \neq -1$
$\frac{d}{dx}\left(\ln\lvert x\rvert\right) = \frac{1}{x}$	$\int \frac{1}{x} \, dx = \ln\lvert x\rvert + C$
$\frac{d}{dx}\left(e^x\right) = e^x$	$\int e^x \, dx = e^x + C$
$\frac{d}{dx}\left(\sin x\right) = \cos x$	$\int \cos x \, dx = \sin x + C$

Differentiation Formula	Indefinite Integral
$\frac{d}{dx}\left(\cos x\right) = -\sin x$	$\int \sin x\,dx = -\cos x + C$
$\frac{d}{dx}\left(\tan x\right) = \sec^2 x$	$\int \sec^2 x\,dx = \tan x + C$
$\frac{d}{dx}\left(\csc x\right) = -\csc x \cot x$	$\int \csc x \cot x\,dx = -\csc x + C$
$\frac{d}{dx}\left(\sec x\right) = \sec x \tan x$	$\int \sec x \tan x\,dx = \sec x + C$
$\frac{d}{dx}\left(\cot x\right) = -\csc^2 x$	$\int \csc^2 x\,dx = -\cot x + C$
$\frac{d}{dx}\left(\sin^{-1} x\right) = \frac{1}{\sqrt{1-x^2}}$	$\int \frac{1}{\sqrt{1-x^2}} = \sin^{-1} x + C$
$\frac{d}{dx}\left(\tan^{-1} x\right) = \frac{1}{1+x^2}$	$\int \frac{1}{1+x^2}\,dx = \tan^{-1} x + C$
$\frac{d}{dx}\left(\sec^{-1} \lvert x\rvert\right) = \frac{1}{x\sqrt{x^2-1}}$	$\int \frac{1}{x\sqrt{x^2-1}}\,dx = \sec^{-1} \lvert x\rvert + C$

From the definition of indefinite integral of f, we know

$$\int f(x)\,dx = F(x) + C$$

if and only if F is an antiderivative of f.

Therefore, when claiming that

$$\int f(x)\,dx = F(x) + C$$

it is important to check whether this statement is correct by verifying that $F'(x) = f(x)$.

Example 5.1.2: Verifying an Indefinite Integral

Each of the following statements is of the form $\int f(x)\,dx = F(x) + C$. Verify that each statement is correct by showing that $F'(x) = f(x)$.

a. $\int \left(x + e^x\right) dx = \frac{x^2}{2} + e^x + C$

b. $\int xe^x\,dx = xe^x - e^x + C$

Solution:

a. Since

$$\frac{d}{dx}\left(\frac{x^2}{2} + e^x + C\right) = x + e^x\ ,$$

the statement

$$\int \left(x + e^x\right) dx = \frac{x^2}{2} + e^x + C$$

is correct.

Note that we are verifying an indefinite integral for a sum. Furthermore, $\frac{x^2}{2}$ and e^x are antiderivatives of x and e^x, respectively, and the sum of the antiderivatives is an antiderivative of the sum. We discuss this fact again later in this section.

b. Using the product rule, we see that

$$\frac{d}{dx}(xe^x - e^x + C) = e^x + xe^x - e^x = xe^x.$$

Therefore, the statement

$$\int xe^x \, dx = xe^x - e^x + C$$

is correct.

Note that we are verifying an indefinite integral for a product. The antiderivative $xe^x - e^x$ is not a product of the antiderivatives. Furthermore, the product of antiderivatives, $x^2e^x/2$ is not an antiderivative of xe^x since

$$\frac{d}{dx}\left(\frac{x^2e^x}{2}\right) = xe^x + \frac{x^2e^x}{2} \neq xe^x \, .$$

In general, the product of antiderivatives is not an antiderivative of a product.

? Exercise 5.1.2

Verify that $\int x\cos x \, dx = x\sin x + \cos x + C$.

Hint

Calculate $\frac{d}{dx}\Big(x\sin x + \cos x + C\Big)$.

Answer

$$\frac{d}{dx}\Big(x\sin x + \cos x + C\Big) = \sin x + x\cos x - \sin x = x\cos x$$

In Table 5.1.1, we listed the indefinite integrals for many elementary functions. Let's now turn our attention to evaluating indefinite integrals for more complicated functions. For example, consider finding an antiderivative of a sum $f + g$. In Example 5.1.2a we showed that an antiderivative of the sum $x + e^x$ is given by the sum $\frac{x^2}{2} + e^x$ —that is, an antiderivative of a sum is given by a sum of antiderivatives. This result was not specific to this example. In general, if F and G are antiderivatives of any functions f and g, respectively, then

$$\frac{d}{dx}\big(F(x) + G(x)\big) = F'(x) + G'(x) = f(x) + g(x).$$

Therefore, $F(x) + G(x)$ is an antiderivative of $f(x) + g(x)$ and we have

$$\int \big(f(x) + g(x)\big) \, dx = F(x) + G(x) + C.$$

Similarly,

$$\int \big(f(x) - g(x)\big) \, dx = F(x) - G(x) + C.$$

In addition, consider the task of finding an antiderivative of $kf(x)$, where k is any real number. Since

$$\frac{d}{dx}\Big(kF(x)\Big) = k\frac{d}{dx}\Big(F(x)\Big) = kF'(x)$$

for any real number k, we conclude that

$$\int kf(x)\,dx = kF(x) + C.$$

These properties are summarized next.

Properties of Indefinite Integrals

Let F and G be antiderivatives of f and g, respectively, and let k be any real number.

Sums and Differences

$$\int \big(f(x) \pm g(x)\big)\,dx = F(x) \pm G(x) + C$$

Constant Multiples

$$\int kf(x)\,dx = kF(x) + C$$

From this theorem, we can evaluate any integral involving a sum, difference, or constant multiple of functions with antiderivatives that are known. Evaluating integrals involving products, quotients, or compositions is more complicated. (See Example $5.1.2b$ for an example involving an antiderivative of a product.) We look at and address integrals involving these more complicated functions in Introduction to Integration. In the next example, we examine how to use this theorem to calculate the indefinite integrals of several functions.

Example 5.1.3: Evaluating Indefinite Integrals

Evaluate each of the following indefinite integrals:

a. $\int \big(5x^3 - 7x^2 + 3x + 4\big)\,dx$

b. $\int \dfrac{x^2 + 4\sqrt[3]{x}}{x}\,dx$

c. $\int \dfrac{4}{1 + x^2}\,dx$

d. $\int \tan x \cos x\,dx$

Solution:

a. Using Properties of Indefinite Integrals, we can integrate each of the four terms in the integrand separately. We obtain

$$\int \big(5x^3 - 7x^2 + 3x + 4\big)\,dx = \int 5x^3\,dx - \int 7x^2\,dx + \int 3x\,dx + \int 4\,dx.$$

From the second part of Properties of Indefinite Integrals, each coefficient can be written in front of the integral sign, which gives

$$\int 5x^3\,dx - \int 7x^2\,dx + \int 3x\,dx + \int 4\,dx = 5\int x^3\,dx - 7\int x^2\,dx + 3\int x\,dx + 4\int 1\,dx.$$

Using the power rule for integrals, we conclude that

$$\int \big(5x^3 - 7x^2 + 3x + 4\big)\,dx = \frac{5}{4}x^4 - \frac{7}{3}x^3 + \frac{3}{2}x^2 + 4x + C.$$

b. Rewrite the integrand as

$$\frac{x^2 + 4\sqrt[3]{x}}{x} = \frac{x^2}{x} + \frac{4\sqrt[3]{x}}{x}.$$

Then, to evaluate the integral, integrate each of these terms separately. Using the power rule, we have

$$\int \left(x+\frac{4}{x^{2/3}}\right)\,dx = \int x\,dx + 4\int x^{-2/3}\,dx$$
$$= \frac{1}{2}x^2 + 4\frac{1}{\left(\frac{-2}{3}\right)+1}x^{(-2/3)+1} + C$$
$$= \frac{1}{2}x^2 + 12x^{1/3} + C.$$

c. Using Properties of Indefinite Integrals, write the integral as

$$4\int \frac{1}{1+x^2}\,dx.$$

Then, use the fact that $\tan^{-1}(x)$ is an antiderivative of $\dfrac{1}{1+x^2}$ to conclude that

$$\int \frac{4}{1+x^2}\,dx = 4\tan^{-1}(x) + C.$$

d. Rewrite the integrand as

$$\tan x \cos x = \frac{\sin x}{\cos x}\cdot \cos x = \sin x.$$

Therefore,

$$\int \tan x \cos x\,dx = \int \sin x\,dx = -\cos x + C.$$

? Exercise 5.1.3

Evaluate $\int \left(4x^3 - 5x^2 + x - 7\right)\,dx$.

Hint

Integrate each term in the integrand separately, making use of the power rule.

Answer

$$\int \left(4x^3 - 5x^2 + x - 7\right)\,dx = \quad x^4 - \frac{5}{3}x^3 + \frac{1}{2}x^2 - 7x + C$$

Initial-Value Problems

We look at techniques for integrating a large variety of functions involving products, quotients, and compositions later in the text. Here we turn to one common use for antiderivatives that arises often in many applications: solving differential equations.

A differential equation is an equation that relates an unknown function and one or more of its derivatives. The equation

$$\frac{dy}{dx} = f(x) \tag{5.1.1}$$

is a simple example of a differential equation. Solving this equation means finding a function y with a derivative f. Therefore, the solutions of Equation 5.1.1 are the antiderivatives of f. If F is one antiderivative of f, every function of the form $y = F(x) + C$ is a solution of that differential equation. For example, the solutions of

$$\frac{dy}{dx} = 6x^2$$

are given by

$$y = \int 6x^2\,dx = 2x^3 + C.$$

Sometimes we are interested in determining whether a particular solution curve passes through a certain point (x_0, y_0) —that is, $y(x_0) = y_0$. The problem of finding a function y that satisfies a differential equation

$$\frac{dy}{dx} = f(x)$$

with the additional condition

$$y(x_0) = y_0$$

is an example of an initial-value problem. The condition $y(x_0) = y_0$ is known as an initial condition. For example, looking for a function y that satisfies the differential equation

$$\frac{dy}{dx} = 6x^2$$

and the initial condition

$$y(1) = 5$$

is an example of an initial-value problem. Since the solutions of the differential equation are $y = 2x^3 + C,$ to find a function y that also satisfies the initial condition, we need to find C such that $y(1) = 2(1)^3 + C = 5$. From this equation, we see that $C = 3,$ and we conclude that $y = 2x^3 + 3$ is the solution of this initial-value problem as shown in the following graph.

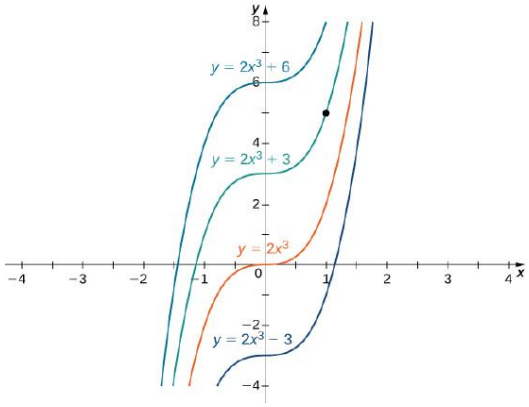


Figure 5.1.2: Some of the solution curves of the differential equation $\frac{dy}{dx} = 6x^2$ are displayed. The function $y = 2x^3 + 3$ satisfies the differential equation and the initial condition $y(1) = 5.$

Example 5.1.4: Solving an Initial-Value Problem

Solve the initial-value problem

$$\frac{dy}{dx} = \sin x, \quad y(0) = 5.$$

Solution

First we need to solve the differential equation. If $\frac{dy}{dx} = \sin x$, then

$$y = \int \sin(x)\, dx = -\cos x + C.$$

Next we need to look for a solution y that satisfies the initial condition. The initial condition $y(0) = 5$ means we need a constant C such that $-\cos x + C = 5.$ Therefore,

$$C = 5 + \cos(0) = 6.$$

The solution of the initial-value problem is $y = -\cos x + 6.$

Exercise 5.1.4

Solve the initial value problem $\dfrac{dy}{dx} = 3x^{-2}, \quad y(1) = 2.$

Hint

Find all antiderivatives of $f(x) = 3x^{-2}.$

Answer

$$y = -\frac{3}{x} + 5$$

Initial-value problems arise in many applications. Next we consider a problem in which a driver applies the brakes in a car. We are interested in how long it takes for the car to stop. Recall that the velocity function $v(t)$ is the derivative of a position function $s(t)$, and the acceleration $a(t)$ is the derivative of the velocity function. In earlier examples in the text, we could calculate the velocity from the position and then compute the acceleration from the velocity. In the next example we work the other way around. Given an acceleration function, we calculate the velocity function. We then use the velocity function to determine the position function.

Example 5.1.5:

A car is traveling at the rate of 88 ft/sec (60 mph) when the brakes are applied. The car begins decelerating at a constant rate of 15 ft/sec^2.

a. How many seconds elapse before the car stops?
b. How far does the car travel during that time?

Solution

a. First we introduce variables for this problem. Let t be the time (in seconds) after the brakes are first applied. Let $a(t)$ be the acceleration of the car (in feet per seconds squared) at time t. Let $v(t)$ be the velocity of the car (in feet per second) at time t. Let $s(t)$ be the car's position (in feet) beyond the point where the brakes are applied at time t.

The car is traveling at a rate of 88 ft/sec. Therefore, the initial velocity is $v(0) = 88$ ft/sec. Since the car is decelerating, the acceleration is

$$a(t) = -15 \text{ ft/sec}^2.$$

The acceleration is the derivative of the velocity,

$$v'(t) = -15.$$

Therefore, we have an initial-value problem to solve:

$$v'(t) = -15, \quad v(0) = 88.$$

Integrating, we find that

$$v(t) = -15t + C.$$

Since $v(0) = 88, C = 88.$ Thus, the velocity function is

$$v(t) = -15t + 88.$$

To find how long it takes for the car to stop, we need to find the time t such that the velocity is zero. Solving $-15t + 88 = 0$, we obtain $t = \dfrac{88}{15}$ sec.

b. To find how far the car travels during this time, we need to find the position of the car after $\dfrac{88}{15}$ sec. We know the velocity $v(t)$ is the derivative of the position $s(t)$. Consider the initial position to be $s(0) = 0$. Therefore, we need to solve the initial-value problem

$$s'(t) = -15t + 88, \quad s(0) = 0.$$

Integrating, we have

$$s(t) = -\frac{15}{2}t^2 + 88t + C.$$

Since $s(0) = 0$, the constant is $C = 0$. Therefore, the position function is

$$s(t) = -\frac{15}{2}t^2 + 88t.$$

After $t = \frac{88}{15}$ sec, the position is $s\left(\frac{88}{15}\right) \approx 258.133$ ft.

Exercise 5.1.5

Suppose the car is traveling at the rate of 44 ft/sec. How long does it take for the car to stop? How far will the car travel?

Hint

$v(t) = -15t + 44.$

Answer

$2.93\text{sec}, 64.5\text{ft}$

Key Concepts

- If F is an antiderivative of f, then every antiderivative of f is of the form $F(x) + C$ for some constant C.
- Solving the initial-value problem

$$\frac{dy}{dx} = f(x), \quad y(x_0) = y_0$$

requires us first to find the set of antiderivatives of f and then to look for the particular antiderivative that also satisfies the initial condition.

Glossary

antiderivative

a function F such that $F'(x) = f(x)$ for all x in the domain of f is an antiderivative of f

indefinite integral

the most general antiderivative of $f(x)$ is the indefinite integral of f; we use the notation $\int f(x)\,dx$ to denote the indefinite integral of f

initial value problem

a problem that requires finding a function y that satisfies the differential equation $\frac{dy}{dx} = f(x)$ together with the initial condition $y(x_0) = y_0$

This page titled 5.1: Antiderivatives is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin "Jed" Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **4.10: Antiderivatives** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

5.2: Approximating Areas with Limits of Finite Sums

Learning Objectives

- Use sigma (summation) notation to calculate sums and powers of integers.
- Use the sum of rectangular areas to approximate the area under a curve.
- Use Riemann sums to approximate area.

Archimedes was fascinated with calculating the areas of various shapes—in other words, the amount of space enclosed by the shape. He used a process that has come to be known as the **method of exhaustion**, which used smaller and smaller shapes, the areas of which could be calculated exactly, to fill an irregular region and thereby obtain closer and closer approximations to the total area. In this process, an area bounded by curves is filled with rectangles, triangles, and shapes with exact area formulas. These areas are then summed to approximate the area of the curved region.

In this section, we develop techniques to approximate the area between a curve, defined by a function $f(x)$, and the x-axis on a closed interval $[a, b]$. Like Archimedes, we first approximate the area under the curve using shapes of known area (namely, rectangles). By using smaller and smaller rectangles, we get closer and closer approximations to the area. Taking a limit allows us to calculate the exact area under the curve.

Let's start by introducing some notation to make the calculations easier. We then consider the case when $f(x)$ is continuous and nonnegative. Later in the chapter, we relax some of these restrictions and develop techniques that apply in more general cases.

Sigma (Summation) Notation

As mentioned, we will use shapes of known area to approximate the area of an irregular region bounded by curves. This process often requires adding up long strings of numbers. To make it easier to write down these lengthy sums, we look at some new notation here, called **sigma notation** (also known as **summation notation**). The Greek capital letter Σ, sigma, is used to express long sums of values in a compact form. For example, if we want to add all the integers from 1 to 20 without sigma notation, we have to write

$$1+2+3+4+5+6+7+8+9+10+11+12+13+14+15+16+17+18+19+20.$$

We could probably skip writing a couple of terms and write

$$1+2+3+4+\cdots+19+20,$$

which is better, but still cumbersome. With sigma notation, we write this sum as

$$\sum_{i=1}^{20} i$$

which is much more compact. Typically, sigma notation is presented in the form

$$\sum_{i=1}^{n} a_i$$

where a_i describes the terms to be added, and the i is called the $index$. Each term is evaluated, then we sum all the values, beginning with the value when $i = 1$ and ending with the value when $i = n$. For example, an expression like $\sum_{i=2}^{7} s_i$ is interpreted as $s_2 + s_3 + s_4 + s_5 + s_6 + s_7$. Note that the index is used only to keep track of the terms to be added; it does not factor into the calculation of the sum itself. The index is therefore called a **dummy variable**. We can use any letter we like for the index. Typically, mathematicians use i, j, k, m, and n for indices.

Let's try a couple of examples of using sigma notation.

Example 5.2.1: Using Sigma Notation

a. Write in sigma notation and evaluate the sum of terms 3^i for $i = 1, 2, 3, 4, 5$.
b. Write the sum in sigma notation:

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25}.$$

Solution

a. Write

$$\sum_{i=1}^{5} 3^i = 3 + 3^2 + 3^3 + 3^4 + 3^5 = 363.$$

b. The denominator of each term is a perfect square. Using sigma notation, this sum can be written as $\sum_{i=1}^{5} \frac{1}{i^2}$.

Exercise 5.2.1

Write in sigma notation and evaluate the sum of terms 2^i for $i = 3, 4, 5, 6$.

Hint

Use the solving steps in Example 5.2.1as a guide.

Answer

$$\sum_{i=3}^{6} 2^i = 2^3 + 2^4 + 2^5 + 2^6 = 120$$

The properties associated with the summation process are given in the following rule.

Rule: Properties of Sigma Notation

Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ represent two sequences of terms and let c be a constant. The following properties hold for all positive integers n and for integers m, with $1 \le m \le n$.

i. $\sum_{i=1}^{n} c = nc$
ii. $\sum_{i=1}^{n} ca_i = c\sum_{i=1}^{n} a_i$
iii. $\sum_{i=1}^{n} (a_i + b_i) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i$
iv. $\sum_{i=1}^{n} (a_i - b_i) = \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} b_i$
v. $\sum_{i=1}^{n} a_i = \sum_{i=1}^{m} a_i + \sum_{i=m+1}^{n} a_i$

Proof

We prove properties (ii.) and (iii.) here, and leave proof of the other properties to the Exercises.

(ii.) We have

$$\sum_{i=1}^{n} ca_i = ca_1 + ca_2 + ca_3 + \cdots + ca_n = c(a_1 + a_2 + a_3 + \cdots + a_n) = c\sum_{i=1}^{n} a_i.$$

(iii.) We have

$$\sum_{i=1}^{n}(a_i+b_i) = (a_1+b_1)+(a_2+b_2)+(a_3+b_3)+\cdots+(a_n+b_n) \tag{5.2.1}$$

$$= (a_1+a_2+a_3+\cdots+a_n)+(b_1+b_2+b_3+\cdots+b_n) \tag{5.2.2}$$

$$= \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i. \tag{5.2.3}$$

□

A few more formulas for frequently found functions simplify the summation process further. These are shown in the next rule, for **sums and powers of integers**, and we use them in the next set of examples.

Rule: Sums and Powers of Integers

1. The sum of n integers is given by

$$\sum_{i=1}^{n} i = 1+2+\cdots+n = \frac{n(n+1)}{2}. \tag{5.2.4}$$

2. The sum of consecutive integers squared is given by

$$\sum_{i=1}^{n} i^2 = 1^2+2^2+\cdots+n^2 = \frac{n(n+1)(2n+1)}{6}. \tag{5.2.5}$$

3. The sum of consecutive integers cubed is given by

$$\sum_{i=1}^{n} i^3 = 1^3+2^3+\cdots+n^3 = \frac{n^2(n+1)^2}{4}. \tag{5.2.6}$$

Example 5.2.2: Evaluation Using Sigma Notation

Write using sigma notation and evaluate:

a. The sum of the terms $(i-3)^2$ for $i = 1, 2, \ldots, 200$.
b. The sum of the terms (i^3-i^2) for $i = 1, 2, 3, 4, 5, 6$

Solution

a. Multiplying out $(i-3)^2$, we can break the expression into three terms.

$$\begin{aligned}
\sum_{i=1}^{200}(i-3)^2 &= \sum_{i=1}^{200}(i^2-6i+9) \\
&= \sum_{i=1}^{200} i^2 - \sum_{i=1}^{200} 6i + \sum_{i=1}^{200} 9 \\
&= \sum_{i=1}^{200} i^2 - 6\sum_{i=1}^{200} i + \sum_{i=1}^{200} 9 \\
&= \frac{200(200+1)(400+1)}{6} - 6\left[\frac{200(200+1)}{2}\right] + 9(200) \\
&= 2,686,700 - 120,600 + 1800 \\
&= 2,567,900
\end{aligned}$$

b. Use sigma notation property iv. and the rules for the sum of squared terms and the sum of cubed terms.

$$\begin{aligned}\sum_{i=1}^{6}(i^3-i^2) &= \sum_{i=1}^{6} i^3 - \sum_{i=1}^{6} i^2 \\ &= \frac{6^2(6+1)^2}{4} - \frac{6(6+1)(2(6)+1)}{6} \\ &= \frac{1764}{4} - \frac{546}{6} \\ &= 350\end{aligned}$$

Exercise 5.2.2

Find the sum of the values of $4+3i$ for $i = 1, 2, \ldots, 100$.

Hint

Use the properties of sigma notation to solve the problem.

Answer

$15,550$

Example 5.2.3: Finding the Sum of the Function Values

Find the sum of the values of $f(x) = x^3$ over the integers $1, 2, 3, \ldots, 10$.

Solution

Using Equation 5.2.6, we have

$$\sum_{i=0}^{10} i^3 = \frac{(10)^2(10+1)^2}{4} = \frac{100(121)}{4} = 3025$$

Exercise 5.2.3

Evaluate the sum indicated by the notation $\sum_{k=1}^{20}(2k+1)$.

Hint

Use the rule on sum and powers of integers (Equations 5.2.4–5.2.6).

Answer

440

Approximating Area

Now that we have the necessary notation, we return to the problem at hand: approximating the area under a curve. Let $f(x)$ be a continuous, nonnegative function defined on the closed interval $[a, b]$. We want to approximate the area A bounded by $f(x)$ above, the x-axis below, the line $x = a$ on the left, and the line $x = b$ on the right (Figure 5.2.1).

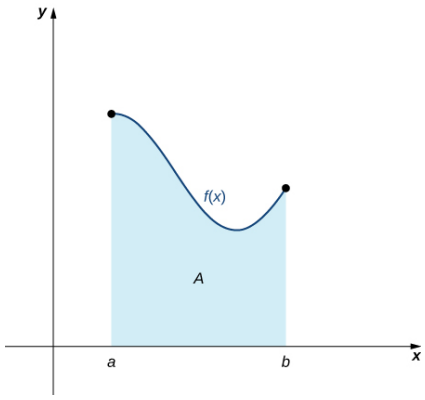


Figure 5.2.1: An area (shaded region) bounded by the curve $f(x)$ at top, the x-axis at bottom, the line $x = a$ to the left, and the line $x = b$ at right.

How do we approximate the area under this curve? The approach is a geometric one. By dividing a region into many small shapes that have known area formulas, we can sum these areas and obtain a reasonable estimate of the true area. We begin by dividing the interval $[a, b]$ into n subintervals of equal width, $\dfrac{b-a}{n}$. We do this by selecting equally spaced points $x_0, x_1, x_2, \ldots, x_n$ with $x_0 = a, x_n = b,$ and

$$x_i - x_{i-1} = \frac{b-a}{n}$$

for $i = 1, 2, 3, \ldots, n$.

We denote the width of each subinterval with the notation Δx, so $\Delta x = \frac{b-a}{n}$ and

$$x_i = x_0 + i\Delta x$$

for $i = 1, 2, 3, \ldots, n$. This notion of dividing an interval $[a, b]$ into subintervals by selecting points from within the interval is used quite often in approximating the area under a curve, so let's define some relevant terminology.

Definition: Partitions

A set of points $P = x_i$ for $i = 0, 1, 2, \ldots, n$ with $a = x_0 < x_1 < x_2 < \ldots < x_n = b$, which divides the interval $[a, b]$ into subintervals of the form $[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]$ is called a **partition** of $[a, b]$. If the subintervals all have the same width, the set of points forms a **regular partition** (or uniform partition) of the interval $[a, b]$.

We can use this regular partition as the basis of a method for estimating the area under the curve. We next examine two methods: the left-endpoint approximation and the right-endpoint approximation.

Rule: Left-Endpoint Approximation

On each subinterval $[x_{i-1}, x_i]$ (for $i = 1, 2, 3, \ldots, n$), construct a rectangle with width Δx and height equal to $f(x_{i-1})$, which is the function value at the left endpoint of the subinterval. Then the area of this rectangle is $f(x_{i-1})\Delta x$. Adding the areas of all these rectangles, we get an approximate value for A (Figure 5.2.2). We use the notation L_n to denote that this is a **left-endpoint approximation** of A using n subintervals.

$$A \approx L_n = f(x_0)\Delta x + f(x_1)\Delta x + \cdots + f(x_{n-1})\Delta x = \sum_{i=1}^{n} f(x_{i-1})\Delta x$$

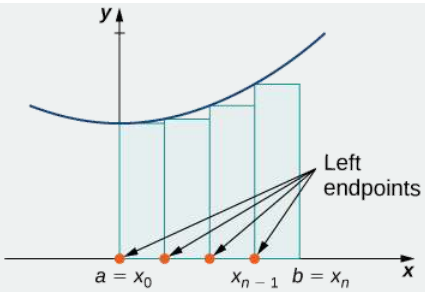


Figure 5.2.2: In the left-endpoint approximation of area under a curve, the height of each rectangle is determined by the function value at the left of each subinterval.

The second method for approximating area under a curve is the right-endpoint approximation. It is almost the same as the left-endpoint approximation, but now the heights of the rectangles are determined by the function values at the right of each subinterval.

Rule: Right-Endpoint Approximation

Construct a rectangle on each subinterval $[x_{i-1}, x_i]$, only this time the height of the rectangle is determined by the function value $f(x_i)$ at the right endpoint of the subinterval. Then, the area of each rectangle is $f(x_i)\,\Delta x$ and the approximation for A is given by

$$A \approx R_n = f(x_1)\Delta x + f(x_2)\Delta x + \cdots + f(x_n)\Delta x = \sum_{i=1}^{n} f(x_i)\Delta x.$$

The notation R_n indicates this is a **right-endpoint approximation** for A (Figure 5.2.3).

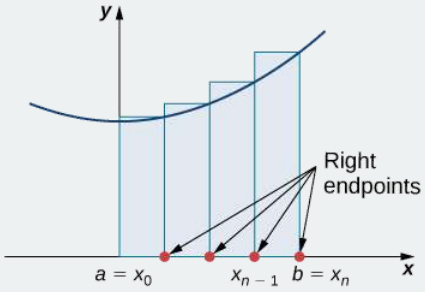


Figure 5.2.3: In the right-endpoint approximation of area under a curve, the height of each rectangle is determined by the function value at the right of each subinterval. Note that the right-endpoint approximation differs from the left-endpoint approximation in Figure 5.2.2.

The graphs in Figure 5.2.4 represent the curve $f(x) = \dfrac{x^2}{2}$. In Figure 5.2.4*b* we divide the region represented by the interval $[0, 3]$ into six subintervals, each of width 0.5. Thus, $\Delta x = 0.5$. We then form six rectangles by drawing vertical lines perpendicular to x_{i-1}, the left endpoint of each subinterval. We determine the height of each rectangle by calculating $f(x_{i-1})$ for $i = 1, 2, 3, 4, 5, 6$. The intervals are $[0, 0.5], [0.5, 1], [1, 1.5], [1.5, 2], [2, 2.5], [2.5, 3]$. We find the area of each rectangle by multiplying the height by the width. Then, the sum of the rectangular areas approximates the area between $f(x)$ and the x-axis. When the left endpoints are used to calculate height, we have a left-endpoint approximation. Thus,

$$\begin{aligned} A \approx L_6 &= \sum_{i=1}^{6} f(x_{i-1})\Delta x = f(x_0)\Delta x + f(x_1)\Delta x + f(x_2)\Delta x + f(x_3)\Delta x + f(x_4)\Delta x + f(x_5)\Delta x \\ &= f(0)0.5 + f(0.5)0.5 + f(1)0.5 + f(1.5)0.5 + f(2)0.5 + f(2.5)0.5 \\ &= (0)0.5 + (0.125)0.5 + (0.5)0.5 + (1.125)0.5 + (2)0.5 + (3.125)0.5 \\ &= 0 + 0.0625 + 0.25 + 0.5625 + 1 + 1.5625 \\ &= 3.4375 \text{ units}^2 \end{aligned}$$

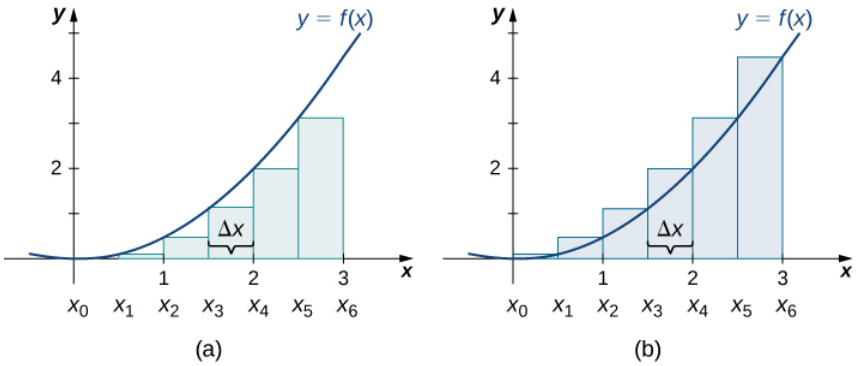


Figure 5.2.4: Methods of approximating the area under a curve by using (a) the left endpoints and (b) the right endpoints.

In Figure 5.2.4b, we draw vertical lines perpendicular to x_i such that x_i is the right endpoint of each subinterval, and calculate $f(x_i)$ for $i = 1, 2, 3, 4, 5, 6$. We multiply each $f(x_i)$ by Δx to find the rectangular areas, and then add them. This is a right-endpoint approximation of the area under $f(x)$. Thus,

$$\begin{aligned} A \approx R_6 &= \sum_{i=1}^{6} f(x_i)\Delta x = f(x_1)\Delta x + f(x_2)\Delta x + f(x_3)\Delta x + f(x_4)\Delta x + f(x_5)\Delta x + f(x_6)\Delta x \\ &= f(0.5)0.5 + f(1)0.5 + f(1.5)0.5 + f(2)0.5 + f(2.5)0.5 + f(3)0.5 \\ &= (0.125)0.5 + (0.5)0.5 + (1.125)0.5 + (2)0.5 + (3.125)0.5 + (4.5)0.5 \\ &= 0.0625 + 0.25 + 0.5625 + 1 + 1.5625 + 2.25 \\ &= 5.6875 \text{ units}^2. \end{aligned}$$

Example 5.2.4: Approximating the Area Under a Curve

Use both left-endpoint and right-endpoint approximations to approximate the area under the curve of $f(x) = x^2$ on the interval $[0, 2]$; use $n = 4$.

Solution

First, divide the interval $[0, 2]$ into n equal subintervals. Using $n = 4$, $\Delta x = \dfrac{(2-0)}{4} = 0.5$. This is the width of each rectangle. The intervals $[0, 0.5], [0.5, 1], [1, 1.5], [1.5, 2]$ are shown in Figure 5.2.5. Using a left-endpoint approximation, the heights are $f(0) = 0$, $f(0.5) = 0.25$, $f(1) = 1$, and $f(1.5) = 2.25$. Then,

$$\begin{aligned} L_4 &= f(x_0)\Delta x + f(x_1)\Delta x + f(x_2)\Delta x + f(x_3)\Delta x \\ &= 0(0.5) + 0.25(0.5) + 1(0.5) + 2.25(0.5) \\ &= 1.75 \text{ units}^2 \end{aligned}$$

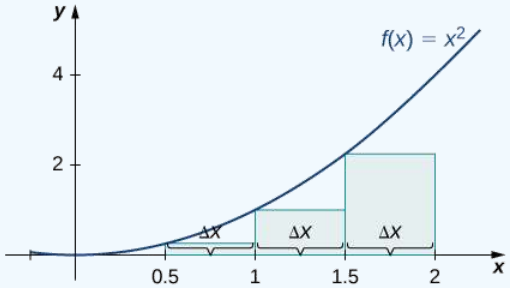


Figure 5.2.5: The graph shows the left-endpoint approximation of the area under $f(x) = x^2$ from 0 to 2.

The right-endpoint approximation is shown in Figure 5.2.6. The intervals are the same, $\Delta x = 0.5$, but now use the right endpoint to calculate the height of the rectangles. We have

$$\begin{aligned} R_4 &= f(x_1)\Delta x + f(x_2)\Delta x + f(x_3)\Delta x + f(x_4)\Delta x \\ &= 0.25(0.5) + 1(0.5) + 2.25(0.5) + 4(0.5) \\ &= 3.75 \text{ units}^2 \end{aligned}$$

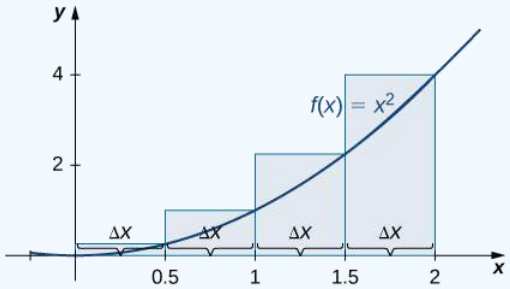


Figure 5.2.6: The graph shows the right-endpoint approximation of the area under $f(x) = x^2$ from 0 to 2.

The left-endpoint approximation is 1.75 units^2; the right-endpoint approximation is 3.75 units^2.

Exercise 5.2.4

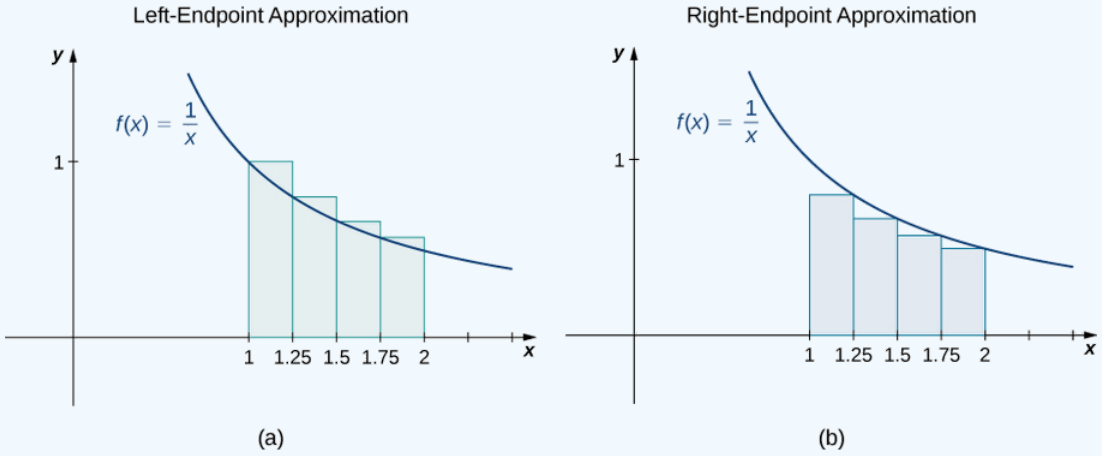
Sketch left-endpoint and right-endpoint approximations for $f(x) = \dfrac{1}{x}$ on $[1, 2]$; use $n = 4$. Approximate the area using both methods.

Hint

Follow the solving strategy in Example 5.2.4step-by-step.

Answer

The left-endpoint approximation is 0.7595 units^2. The right-endpoint approximation is 0.6345 units^2. See the below Media.



Looking at Figure 5.2.4 and the graphs in Example 5.2.4, we can see that when we use a small number of intervals, neither the left-endpoint approximation nor the right-endpoint approximation is a particularly accurate estimate of the area under the curve. However, it seems logical that if we increase the number of points in our partition, our estimate of A will improve. We will have more rectangles, but each rectangle will be thinner, so we will be able to fit the rectangles to the curve more precisely.

We can demonstrate the improved approximation obtained through smaller intervals with an example. Let's explore the idea of increasing n, first in a left-endpoint approximation with four rectangles, then eight rectangles, and finally 32 rectangles. Then, let's do the same thing in a right-endpoint approximation, using the same sets of intervals, of the same curved region. Figure 5.2.7 shows the area of the region under the curve $f(x) = (x-1)^3 + 4$ on the interval $[0, 2]$ using a left-endpoint approximation where $n = 4$. The width of each rectangle is

$$\Delta x = \frac{2-0}{4} = \frac{1}{2}.$$

The area is approximated by the summed areas of the rectangles, or

$$L_4 = f(0)(0.5) + f(0.5)(0.5) + f(1)(0.5) + f(1.5)0.5 = 7.5 \text{ units}^2$$

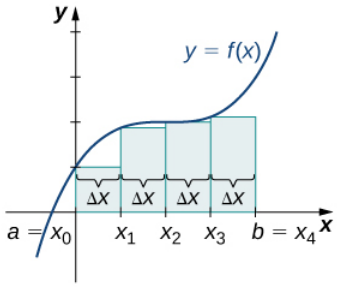


Figure 5.2.7: With a left-endpoint approximation and dividing the region from a to b into four equal intervals, the area under the curve is approximately equal to the sum of the areas of the rectangles.

Figure 5.2.8 shows the same curve divided into eight subintervals. Comparing the graph with four rectangles in Figure 5.2.7 with this graph with eight rectangles, we can see there appears to be less white space under the curve when $n = 8$. This white space is area under the curve we are unable to include using our approximation. The area of the rectangles is

$$L_8 = f(0)(0.25) + f(0.25)(0.25) + f(0.5)(0.25) + f(0.75)(0.25) + f(1)(0.25) + f(1.25)(0.25) + f(1.5)(0.25) + f(1.75)(0.25) = 7.75 \text{ units}^2$$

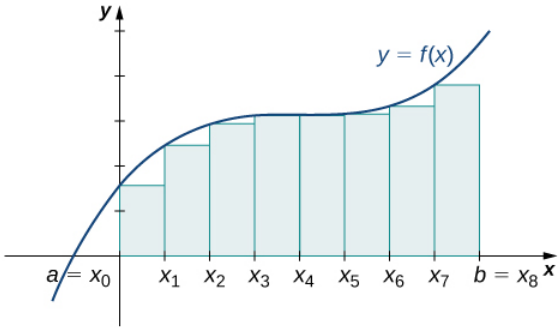


Figure 5.2.8: The region under the curve is divided into $n = 8$ rectangular areas of equal width for a left-endpoint approximation.

The graph in Figure 5.2.9 shows the same function with 32 rectangles inscribed under the curve. There appears to be little white space left. The area occupied by the rectangles is

$$L_{32} = f(0)(0.0625) + f(0.0625)(0.0625) + f(0.125)(0.0625) + \cdots + f(1.9375)(0.0625) = 7.9375 \text{ units}^2.$$

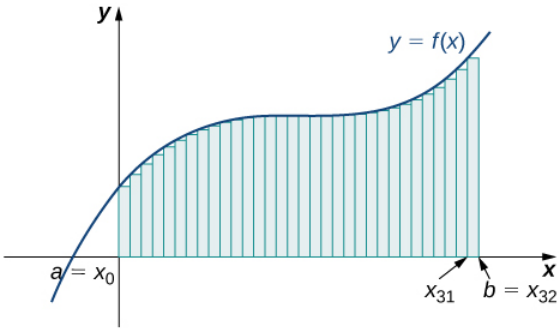


Figure 5.2.9: Here, 32 rectangles are inscribed under the curve for a left-endpoint approximation.

We can carry out a similar process for the right-endpoint approximation method. A right-endpoint approximation of the same curve, using four rectangles (Figure 5.2.10), yields an area

$$R_4 = f(0.5)(0.5) + f(1)(0.5) + f(1.5)(0.5) + f(2)(0.5) = 8.5 \text{ units}^2.$$

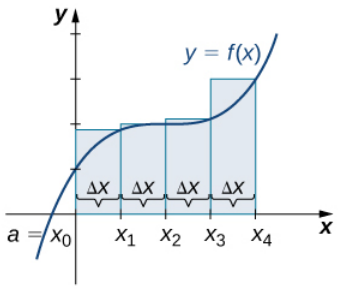


Figure 5.2.10: Now we divide the area under the curve into four equal subintervals for a right-endpoint approximation.

Dividing the region over the interval $[0, 2]$ into eight rectangles results in $\Delta x = \dfrac{2-0}{8} = 0.25$. The graph is shown in Figure 5.2.11 The area is

$$R_8 = f(0.25)(0.25) + f(0.5)(0.25) + f(0.75)(0.25) + f(1)(0.25) + f(1.25)(0.25) + f(1.5)(0.25) + f(1.75)(0.25) + f(2)(0.25) = 8.25 \text{ units}^2$$

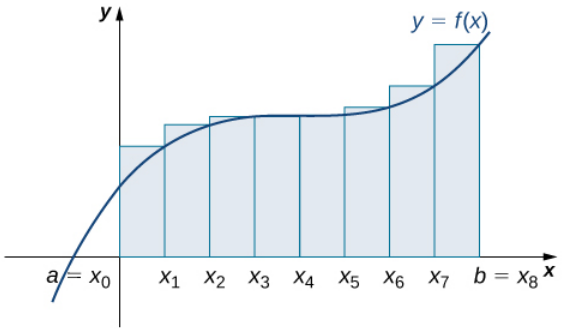


Figure 5.2.11: Here we use right-endpoint approximation for a region divided into eight equal subintervals.

Last, the right-endpoint approximation with $n = 32$ is close to the actual area (Figure 5.2.12). The area is approximately

$$R_{32} = f(0.0625)(0.0625) + f(0.125)(0.0625) + f(0.1875)(0.0625) + \cdots + f(2)(0.0625) = 8.0625 \text{ units}^2$$

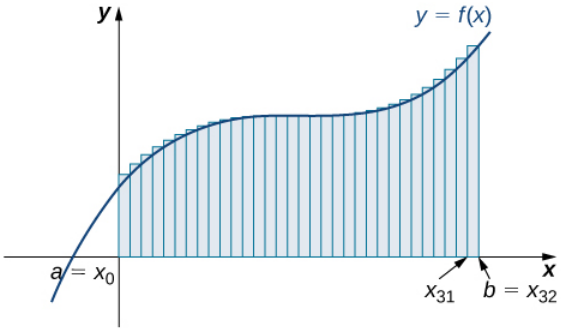


Figure 5.2.12: The region is divided into 32 equal subintervals for a right-endpoint approximation.

Based on these figures and calculations, it appears we are on the right track; the rectangles appear to approximate the area under the curve better as n gets larger. Furthermore, as n increases, both the left-endpoint and right-endpoint approximations appear to approach an area of 8 square units. Table 5.2.15 shows a numerical comparison of the left- and right-endpoint methods. The idea that the approximations of the area under the curve get better and better as n gets larger and larger is very important, and we now explore this idea in more detail.

Table 5.2.15: Converging Values of Left- and Right-Endpoint Approximations as n Increases

Value of n	Approximate Area L_n	Approximate Area R_n
$n = 4$	7.5	8.5
$n = 8$	7.75	8.25

Value of n	Approximate Area L_n	Approximate Area R_n
$n = 32$	7.94	8.06

Forming Riemann Sums

So far we have been using rectangles to approximate the area under a curve. The heights of these rectangles have been determined by evaluating the function at either the right or left endpoints of the subinterval $[x_{i-1}, x_i]$. In reality, there is no reason to restrict evaluation of the function to one of these two points only. We could evaluate the function at any point x_i^* in the subinterval $[x_{i-1}, x_i]$, and use $f(x_i^*)$ as the height of our rectangle. This gives us an estimate for the area of the form

$$A \approx \sum_{i=1}^{n} f(x_i^*)\,\Delta x.$$

A sum of this form is called a Riemann sum, named for the 19th-century mathematician Bernhard Riemann, who developed the idea.

Definition: Riemann sum

Let $f(x)$ be defined on a closed interval $[a, b]$ and let P be any partition of $[a, b]$. Let Δx_i be the width of each subinterval $[x_{i-1}, x_i]$ and for each i, let x_i^* be any point in $[x_{i-1},\ x_i]$. A Riemann sum is defined for $f(x)$ as

$$\sum_{i=1}^{n} f(x_i^*)\,\Delta x_i.$$

At this point, we'll choose a regular partition P, as we have in our examples above. This forces all Δx_i to be equal to $\Delta x = \dfrac{b-a}{n}$ for any natural number of intervals n.

Recall that with the left- and right-endpoint approximations, the estimates seem to get better and better as n get larger and larger. The same thing happens with Riemann sums. Riemann sums give better approximations for larger values of n. We are now ready to define the area under a curve in terms of Riemann sums.

Definition: Area Under the Curve

Let $f(x)$ be a continuous, nonnegative function on an interval $[a, b]$, and let $\sum_{i=1}^{n} f(x_i^*)\,\Delta x$ be a Riemann sum for $f(x)$ with a regular partition P. Then, the **area under the curve** $y = f(x)$ on $[a, b]$ is given by

$$A = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\,\Delta x.$$

See a graphical demonstration of the construction of a Riemann sum.

Some subtleties here are worth discussing. First, note that taking the limit of a sum is a little different from taking the limit of a function $f(x)$ as x goes to infinity. Limits of sums are discussed in detail in the chapter on Sequences and Series; however, for now we can assume that the computational techniques we used to compute limits of functions can also be used to calculate limits of sums.

Second, we must consider what to do if the expression converges to different limits for different choices of x_i^*. Fortunately, this does not happen. Although the proof is beyond the scope of this text, it can be shown that if $f(x)$ is continuous on the closed interval $[a, b]$, then $\lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x$ exists and is unique (in other words, it does not depend on the choice of x_i^*).

We look at some examples shortly. But, before we do, let's take a moment and talk about some specific choices for x_i^*. Although any choice for x_i^* gives us an estimate of the area under the curve, we don't necessarily know whether that estimate is too high

(overestimate) or too low (underestimate). If it is important to know whether our estimate is high or low, we can select our value for x_i^* to guarantee one result or the other.

If we want an overestimate, for example, we can choose x_i^* such that for $i = 1, 2, 3, \ldots, n, f(x_i^*) \geq f(x)$ for all $x \in [x_i - 1, x_i]$. In other words, we choose x_i^* so that for $i = 1, 2, 3, \ldots, n, f(x_i^*)$ is the maximum function value on the interval $[x_{i-1}, x_i]$. If we select x_i^* in this way, then the Riemann sum $\sum_{i=1}^{n} f(x_i^*)\Delta x$ is called an **upper sum**. Similarly, if we want an underestimate, we can choose x_i^* so that for $i = 1, 2, 3, \ldots, n, f(x_i^*)$ is the minimum function value on the interval $[x_{i-1}, x_i]$. In this case, the associated Riemann sum is called a **lower sum**. Note that if $f(x)$ is either increasing or decreasing throughout the interval $[a, b]$, then the maximum and minimum values of the function occur at the endpoints of the subintervals, so the upper and lower sums are just the same as the left- and right-endpoint approximations.

Example 5.2.5: Finding Lower and Upper Sums

Find a lower sum for $f(x) = 10 - x^2$ on $[1, 2]$; let $n = 4$ subintervals.

Solution

With $n = 4$ over the interval $[1, 2]$, $\Delta x = \frac{1}{4}$. We can list the intervals as $[1, 1.25]$, $[1.25, 1.5]$, $[1.5, 1.75]$, and $[1.75, 2]$. Because the function is decreasing over the interval $[1, 2]$, Figure shows that a lower sum is obtained by using the right endpoints.

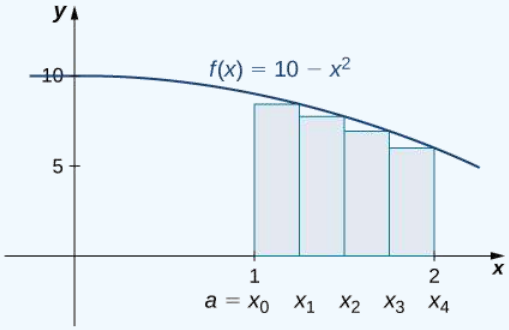


Figure 5.2.13: The graph of $f(x) = 10 - x^2$ is set up for a right-endpoint approximation of the area bounded by the curve and the x-axis on $[1, 2]$, and it shows a lower sum.

The Riemann sum is

$$\begin{aligned}\sum_{k=1}^{4}(10 - x^2)(0.25) &= 0.25[10 - (1.25)^2 + 10 - (1.5)^2 + 10 - (1.75)^2 + 10 - (2)^2] \\ &= 0.25[8.4375 + 7.75 + 6.9375 + 6] \\ &= 7.28 \text{ units}^2.\end{aligned}$$

The area of 7.28 units^2 is a lower sum and an underestimate.

Exercise 5.2.5

a. Find an upper sum for $f(x) = 10 - x^2$ on $[1, 2]$; let $n = 4$.
b. Sketch the approximation.

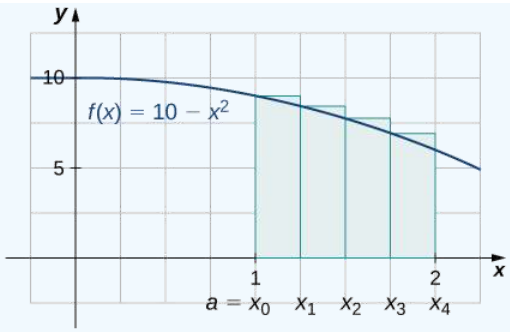
Hint

$f(x)$ is decreasing on $[1, 2]$, so the maximum function values occur at the left endpoints of the subintervals.

Answer

a. Upper sum=8.0313 units^2.

b.



Example 5.2.6: Finding Lower and Upper Sums for $f(x) = \sin x$

Find a lower sum for $f(x) = \sin x$ over the interval $[a, b] = \left[0, \frac{\pi}{2}\right]$; let $n = 6$.

Solution

Let's first look at the graph in Figure 5.2.14 to get a better idea of the area of interest.

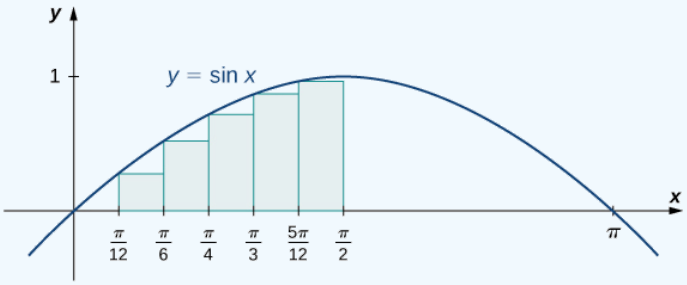


Figure 5.2.14: The graph of $y = \sin x$ is divided into six regions: $\Delta x = \dfrac{\pi/2}{6} = \dfrac{\pi}{12}$.

The intervals are $\left[0, \frac{\pi}{12}\right]$, $\left[\frac{\pi}{12}, \frac{\pi}{6}\right]$, $\left[\frac{\pi}{6}, \frac{\pi}{4}\right]$, $\left[\frac{\pi}{4}, \frac{\pi}{3}\right]$, $\left[\frac{\pi}{3}, \frac{5\pi}{12}\right]$, and $\left[\frac{5\pi}{12}, \frac{\pi}{2}\right]$. Note that $f(x) = \sin x$ is increasing on the interval $\left[0, \frac{\pi}{2}\right]$, so a left-endpoint approximation gives us the lower sum. A left-endpoint approximation is the Riemann sum $\sum_{i=0}^{5} \sin x_i \left(\frac{\pi}{12}\right)$. We have

$$A \approx \sin(0)\left(\tfrac{\pi}{12}\right) + \sin\left(\tfrac{\pi}{12}\right)\left(\tfrac{\pi}{12}\right) + \sin\left(\tfrac{\pi}{6}\right)\left(\tfrac{\pi}{12}\right) + \sin\left(\tfrac{\pi}{4}\right)\left(\tfrac{\pi}{12}\right) + \sin\left(\tfrac{\pi}{3}\right)\left(\tfrac{\pi}{12}\right) + \sin\left(\tfrac{5\pi}{12}\right)\left(\tfrac{\pi}{12}\right) \approx 0.863 \text{ units}^2.$$

Exercise 5.2.6

Using the function $f(x) = \sin x$ over the interval $\left[0, \frac{\pi}{2}\right]$, find an upper sum; let $n = 6$.

Hint

Follow the steps from Example 5.2.6

Answer

$A \approx 1.125 \text{ units}^2$

Key Concepts

- The use of sigma (summation) notation of the form $\sum_{i=1}^{n} a_i$ is useful for expressing long sums of values in compact form.
- For a continuous function defined over an interval $[a, b]$, the process of dividing the interval into n equal parts, extending a rectangle to the graph of the function, calculating the areas of the series of rectangles, and then summing the areas yields an approximation of the area of that region.
- When using a regular partition, the width of each rectangle is $\Delta x = \dfrac{b-a}{n}$.

- Riemann sums are expressions of the form $\sum_{i=1}^{n} f(x_i^*)\Delta x$, and can be used to estimate the area under the curve $y = f(x)$. Left- and right-endpoint approximations are special kinds of Riemann sums where the values of x_i^* are chosen to be the left or right endpoints of the subintervals, respectively.
- Riemann sums allow for much flexibility in choosing the set of points x_i^* at which the function is evaluated, often with an eye to obtaining a lower sum or an upper sum.

Key Equations

- **Properties of Sigma Notation**

$$\sum_{i=1}^{n} c = nc$$

$$\sum_{i=1}^{n} ca_i = c\sum_{i=1}^{n} a_i$$

$$\sum_{i=1}^{n}(a_i + b_i) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i$$

$$\sum_{i=1}^{n}(a_i - b_i) = \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} b_i$$

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{m} a_i + \sum_{i=m+1}^{n} a_i$$

- **Sums and Powers of Integers**

$$\sum_{i=1}^{n} i = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} i^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{i=0}^{n} i^3 = 1^3 + 2^3 + \cdots + n^3 = \frac{n^2(n+1)^2}{4}$$

- **Left-Endpoint Approximation**

$$A \approx L_n = f(x_0)\Delta x + f(x_1)\Delta x + \cdots + f(x_{n-1})\Delta x = \sum_{i=1}^{n} f(x_{i-1})\Delta x$$

- **Right-Endpoint Approximation**

$$A \approx R_n = f(x_1)\Delta x + f(x_2)\Delta x + \cdots + f(x_n)\Delta x = \sum_{i=1}^{n} f(x_i)\Delta x$$

Glossary

left-endpoint approximation

an approximation of the area under a curve computed by using the left endpoint of each subinterval to calculate the height of the vertical sides of each rectangle

lower sum

a sum obtained by using the minimum value of $f(x)$ on each subinterval

partition

a set of points that divides an interval into subintervals

regular partition

a partition in which the subintervals all have the same width

riemann sum

an estimate of the area under the curve of the form $A \approx \sum_{i=1}^{n} f(x_i^*)\Delta x$

right-endpoint approximation

the right-endpoint approximation is an approximation of the area of the rectangles under a curve using the right endpoint of each subinterval to construct the vertical sides of each rectangle

sigma notation

(also, **summation notation**) the Greek letter sigma (Σ) indicates addition of the values; the values of the index above and below the sigma indicate where to begin the summation and where to end it

upper sum

a sum obtained by using the maximum value of $f(x)$ on each subinterval

---

This page titled 5.2: Approximating Areas with Limits of Finite Sums is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin "Jed" Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **5.1: Approximating Areas** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

5.3: The Definite Integral

Learning Objectives

- State the definition of the definite integral.
- Explain the terms integrand, limits of integration, and variable of integration.
- Explain when a function is integrable.
- Describe the relationship between the definite integral and net area.
- Use geometry and the properties of definite integrals to evaluate them.
- Calculate the average value of a function.

In the preceding section we defined the area under a curve in terms of Riemann sums:

$$A = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x.$$

However, this definition came with restrictions. We required $f(x)$ to be continuous and nonnegative. Unfortunately, real-world problems don't always meet these restrictions. In this section, we look at how to apply the concept of the area under the curve to a broader set of functions through the use of the definite integral.

Definition and Notation

The definite integral generalizes the concept of the area under a curve. We lift the requirements that $f(x)$ be continuous and nonnegative, and define the definite integral as follows.

Definition: Definite Integral

If $f(x)$ is a function defined on an interval $[a, b]$, the ***definite integral*** of f from a to b is given by

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x,$$

provided the limit exists. If this limit exists, the function $f(x)$ is said to be integrable on $[a, b]$, or is an integrable function.

The integral symbol in the previous definition should look familiar. We have seen similar notation in the chapter on Applications of Derivatives, where we used the indefinite integral symbol (without the a and b above and below) to represent an antiderivative. Although the notation for indefinite integrals may look similar to the notation for a definite integral, they are not the same. A definite integral is a number. An indefinite integral is a family of functions. Later in this chapter we examine how these concepts are related. However, close attention should always be paid to notation so we know whether we're working with a definite integral or an indefinite integral.

Integral notation goes back to the late seventeenth century and is one of the contributions of Gottfried Wilhelm **Leibniz**, who is often considered to be the codiscoverer of calculus, along with Isaac Newton. The integration symbol $\int$ is an elongated S, suggesting sigma or summation. On a definite integral, above and below the summation symbol are the boundaries of the interval, $[a, b]$. The numbers a and b are x-values and are called the **limits of integration**; specifically, a is the lower limit and b is the upper limit. To clarify, we are using the word limit in two different ways in the context of the definite integral. First, we talk about the limit of a sum as $n \to \infty$. Second, the boundaries of the region are called the *limits of integration.*

We call the function $f(x)$ the **integrand**, and the dx indicates that $f(x)$ is a function with respect to x, called the **variable of integration**. Note that, like the index in a sum, the variable of integration is a **dummy variable**, and has no impact on the computation of the integral. We could use any variable we like as the variable of integration:

$$\int_a^b f(x)\,dx = \int_a^b f(t)\,dt = \int_a^b f(u)\,du$$

Previously, we discussed the fact that if $f(x)$ is continuous on $[a,b]$, then the limit $\lim_{n\to\infty}\sum_{i=1}^{n} f(x_i^*)\Delta x$ exists and is unique. This leads to the following theorem, which we state without proof.

Continuous Functions Are Integrable

If $f(x)$ is continuous on $[a,b]$, then f is integrable on $[a,b]$.

Functions that are not continuous on $[a,b]$ may still be integrable, depending on the nature of the discontinuities. For example, functions with a finite number of jump discontinuities or removable discontinuities on a closed interval are integrable.

It is also worth noting here that we have retained the use of a regular partition in the Riemann sums. This restriction is not strictly necessary. Any partition can be used to form a Riemann sum. However, if a nonregular partition is used to define the definite integral, it is not sufficient to take the limit as the number of subintervals goes to infinity. Instead, we must take the limit as the width of the largest subinterval goes to zero. This introduces a little more complex notation in our limits and makes the calculations more difficult without really gaining much additional insight, so we stick with regular partitions for the Riemann sums.

Example 5.3.1: Evaluating an Integral Using the Definition

Use the definition of the definite integral to evaluate $\int_0^2 x^2\,dx$. Use a right-endpoint approximation to generate the Riemann sum.

Solution

We first want to set up a Riemann sum. Based on the limits of integration, we have $a=0$ and $b=2$. For $i=0,1,2,\ldots,n$, let $P=x_i$ be a regular partition of $[0,2]$. Then

$$\Delta x=\frac{b-a}{n}=\frac{2}{n}.$$

Since we are using a right-endpoint approximation to generate Riemann sums, for each i, we need to calculate the function value at the right endpoint of the interval $[x_{i-1},x_i]$. The right endpoint of the interval is x_i, and since P is a regular partition,

$$x_i=x_0+i\Delta x=0+i\left[\frac{2}{n}\right]=\frac{2i}{n}.$$

Thus, the function value at the right endpoint of the interval is

$$f(x_i)=x_i^2=\left(\frac{2i}{n}\right)^2=\frac{4i^2}{n^2}.$$

Then the Riemann sum takes the form

$$\sum_{i=1}^{n} f(x_i)\Delta x=\sum_{i=1}^{n}\left(\frac{4i^2}{n^2}\right)\frac{2}{n}=\sum_{i=1}^{n}\frac{8i^2}{n^3}=\frac{8}{n^3}\sum_{i=1}^{n} i^2.$$

Using the summation formula for $\sum_{i=1}^{n} i^2$, we have

$$\begin{aligned}\sum_{i=1}^{n} f(x_i)\Delta x &= \frac{8}{n^3}\sum_{i=1}^{n} i^2 \\ &= \frac{8}{n^3}\left[\frac{n(n+1)(2n+1)}{6}\right] \\ &= \frac{8}{n^3}\left[\frac{2n^3+3n^2+n}{6}\right] \\ &= \frac{16n^3+24n^2+n}{6n^3} \\ &= \frac{8}{3}+\frac{4}{n}+\frac{1}{6n^2}.\end{aligned}$$

Now, to calculate the definite integral, we need to take the limit as $n \to \infty$. We get

$$\begin{aligned}\int_0^2 x^2 dx &= \lim_{n\to\infty}\sum_{i=1}^{n} f(x_i)\Delta x \\ &= \lim_{n\to\infty}\left(\frac{8}{3}+\frac{4}{n}+\frac{1}{6n^2}\right) \\ &= \lim_{n\to\infty}\left(\frac{8}{3}\right)+\lim_{n\to\infty}\left(\frac{4}{n}\right)+\lim_{n\to\infty}\left(\frac{1}{6n^2}\right) \\ &= \frac{8}{3}+0+0=\frac{8}{3}.\end{aligned}$$

? Exercise 5.3.1

Use the definition of the definite integral to evaluate $\int_0^3 (2x-1)\,dx$.

Use a right-endpoint approximation to generate the Riemann sum.

Hint

Use the solving strategy from Example 5.3.1

Answer

6

Evaluating Definite Integrals

Evaluating definite integrals this way can be quite tedious because of the complexity of the calculations. Later in this chapter we develop techniques for evaluating definite integrals *without* taking limits of Riemann sums. However, for now, we can rely on the fact that definite integrals represent the area under the curve, and we can evaluate definite integrals by using geometric formulas to calculate that area. We do this to confirm that definite integrals do, indeed, represent areas, so we can then discuss what to do in the case of a curve of a function dropping below the x-axis.

✓ Example 5.3.2: Using Geometric Formulas to Calculate Definite Integrals

Use the formula for the area of a circle to evaluate $\int_3^6 \sqrt{9-(x-3)^2}\,dx$.

Solution

The function describes a semicircle with radius 3. To find

$$\int_3^6 \sqrt{9-(x-3)^2}\,dx$$

we want to find the area under the curve over the interval $[3,6]$. The formula for the area of a circle is $A=\pi r^2$. The area of a semicircle is just one-half the area of a circle, or $A=\frac{1}{2}\pi r^2$. The shaded area in Figure 5.3.1 covers one-half of the semicircle, or $A=\frac{1}{4}\pi r^2$. Thus,

$$\int_3^6 \sqrt{9-(x-3)^2}\,dx=\frac{1}{4}\pi(3)^2=\frac{9}{4}\pi\approx 7.069.$$

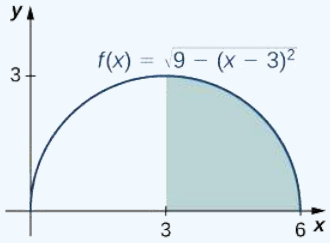


Figure 5.3.1: The value of the integral of the function $f(x)$ over the interval $[3,6]$ is the area of the shaded region.

? Exercise 5.3.2

Use the formula for the area of a trapezoid to evaluate $\int_2^4 (2x+3)\,dx$.

Hint

Graph the function $f(x)$ and calculate the area under the function on the interval $[2,4]$.

Answer

18 square units

Area and the Definite Integral

When we defined the definite integral, we lifted the requirement that $f(x)$ be nonnegative. But how do we interpret "the area under the curve" when $f(x)$ is negative?

Net Signed Area

Let us return to the Riemann sum. Consider, for example, the function $f(x)=2-2x^2$ (shown in Figure 5.3.2) on the interval $[0,2]$. Use $n=8$ and choose $\{x_i^*\}$ as the left endpoint of each interval. Construct a rectangle on each subinterval of height $f(x_i^*)$ and width Δx. When $f(x_i^*)$ is positive, the product $f(x_i^*)\Delta x$ represents the area of the rectangle, as before. When $f(x_i^*)$ is negative, however, the product $f(x_i^*)\Delta x$ represents the negative of the area of the rectangle. The Riemann sum then becomes

$$\sum_{i=1}^{8} f(x_i^*)\Delta x=(\text{Area of rectangles above the } x\text{-axis})-(\text{Area of rectangles below the } x\text{-axis})$$

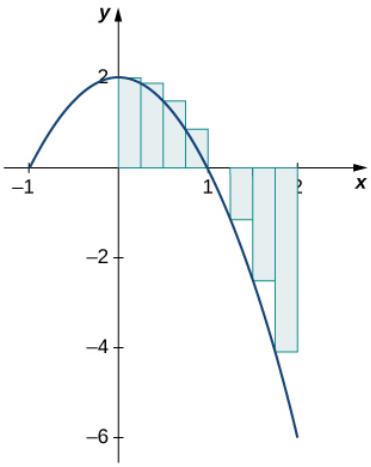


Figure 5.3.2: For a function that is partly negative, the Riemann sum is the area of the rectangles above the x-axis less the area of the rectangles below the x-axis.

Taking the limit as $n \to \infty$, the Riemann sum approaches the area between the curve above the x-axis and the x-axis, less the area between the curve below the x-axis and the x-axis, as shown in Figure 5.3.3. Then,

$$\int_0^2 f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(c_i)\Delta x = A_1 - A_2.$$

The quantity $A_1 - A_2$ is called the **net signed area.**

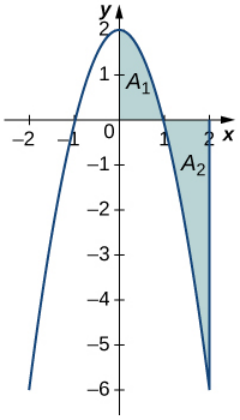


Figure 5.3.3: In the limit, the definite integral equals area A_1 less area A_2, or the net signed area.

Notice that net signed area can be positive, negative, or zero. If the area above the x-axis is larger, the net signed area is positive. If the area below the x-axis is larger, the net signed area is negative. If the areas above and below the x-axis are equal, the net signed area is zero.

Example 5.3.3: Finding the Net Signed Area

Find the net signed area between the curve of the function $f(x) = 2x$ and the x-axis over the interval $[-3, 3]$.

Solution

The function produces a straight line that forms two triangles: one from $x = -3$ to $x = 0$ and the other from $x = 0$ to $x = 3$ (Figure 5.3.4). Using the geometric formula for the area of a triangle, $A = \frac{1}{2}bh$, the area of triangle A_1, above the axis, is

$$A_1 = \frac{1}{2}3(6) = 9\,,$$

where 3 is the base and $2(3) = 6$ is the height. The area of triangle A_2, below the axis, is

$$A_2 = \frac{1}{2}(3)(6) = 9,$$

where 3 is the base and 6 is the height. Thus, the net area is

$$\int_{-3}^{3} 2x\,dx = A_1 - A_2 = 9 - 9 = 0.$$

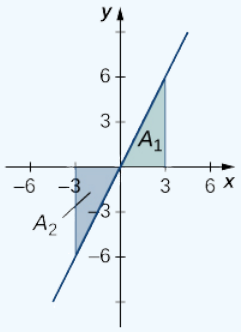


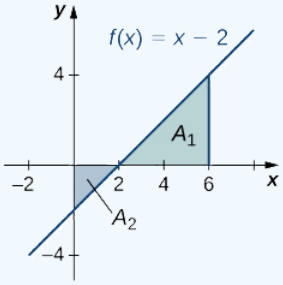
Figure 5.3.4: The area above the curve and below the x-axis equals the area below the curve and above the x-axis.

Analysis

If A_1 is the area above the x-axis and A_2 is the area below the x-axis, then the net area is $A_1 - A_2$. Since the areas of the two triangles are equal, the net area is zero.

? Exercise 5.3.3

Find the net signed area of $f(x) = x - 2$ over the interval $[0, 6]$, illustrated in the following image.



Hint

Use the solving method described in Example 5.3.3

Answer

6

Total Area

One application of the definite integral is finding **displacement** when given a velocity function. If $v(t)$ represents the velocity of an object as a function of time, then the area under the curve tells us how far the object is from its original position. This is a very important application of the definite integral, and we examine it in more detail later in the chapter. For now, we're just going to look at some basics to get a feel for how this works by studying constant velocities.

When velocity is a constant, the area under the curve is just velocity times time. This idea is already very familiar. If a car travels away from its starting position in a straight line at a speed of 70 mph for 2 hours, then it is 140 miles away from its original position (Figure 5.3.5). Using integral notation, we have

$$\int_0^2 70\,dt = 140 \text{ miles}.$$

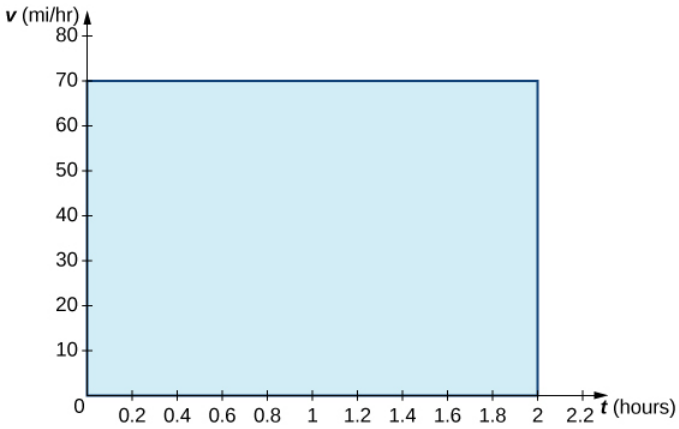


Figure 5.3.5: The area under the curve $v(t) = 70$ tells us how far the car is from its starting point at a given time.

In the context of displacement, net signed area allows us to take direction into account. If a car travels straight north at a speed of 60 mph for 2 hours, it is 120 miles north of its starting position. If the car then turns around and travels south at a speed of 40 mph for 3 hours, it will be back at it starting position (Figure 5.3.6). Again, using integral notation, we have

$$\int_0^2 60\,dt + \int_2^5 -40\,dt = 120 - 120 = 0.$$

In this case the displacement is zero.

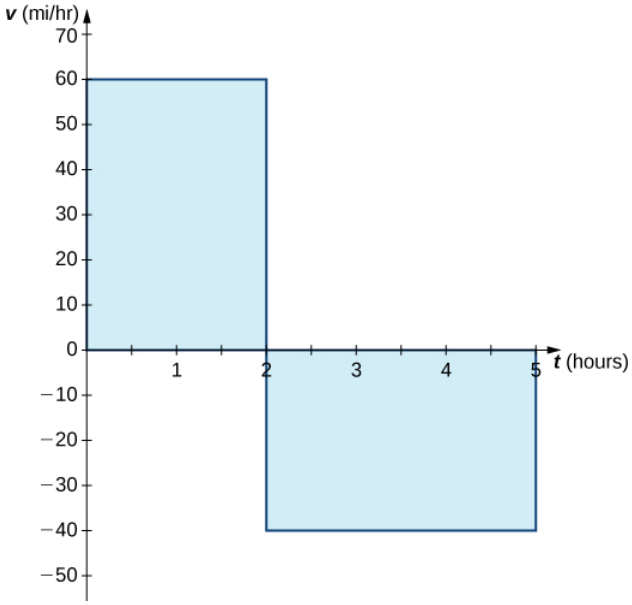


Figure 5.3.6: The area above the axis and the area below the axis are equal, so the net signed area is zero.

Suppose we want to know how far the car travels overall, regardless of direction. In this case, we want to know the area between the curve and the t-axis, regardless of whether that area is above or below the axis. This is called the **total area.**

Graphically, it is easiest to think of calculating total area by adding the areas above the axis and the areas below the axis (rather than subtracting the areas below the axis, as we did with net signed area). To accomplish this mathematically, we use the absolute value function. Thus, the total distance traveled by the car is

$$\int_0^2 |60|\,dt + \int_2^5 |-40|\,dt = \int_0^2 60\,dt + \int_2^5 40\,dt = 120 + 120 = 240.$$

Bringing these ideas together formally, we state the following definitions.

Definition: Net Signed Area

Let $f(x)$ be an integrable function defined on an interval $[a, b]$. Let A_1 represent the area between $f(x)$ and the x-axis that lies above the axis and let A_2 represent the area between $f(x)$ and the x-axis that lies below the axis. Then, the *net signed area* between $f(x)$ and the x-axis is given by

$$\int_a^b f(x)\,dx = A_1 - A_2.$$

The total area between $f(x)$ and the x-axis is given by

$$\int_a^b |f(x)|\,dx = A_1 + A_2.$$

Example 5.3.4: Finding the Total Area

Find the total area between $f(x) = x - 2$ and the x-axis over the interval $[0, 6]$.

Solution

Calculate the x-intercept as $(2, 0)$ (set $y = 0$, solve for x). To find the total area, take the area below the x-axis over the subinterval $[0, 2]$ and add it to the area above the x-axis on the subinterval $[2, 6]$ (Figure 5.3.7).

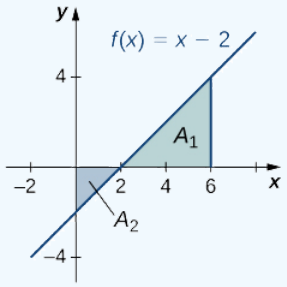


Figure 5.3.7: The total area between the line and the x-axis over $[0, 6]$ is A_2 plus A_1.

We have

$$\int_0^6 |(x-2)|\,dx = A_2 + A_1.$$

Then, using the formula for the area of a triangle, we obtain

$$A_2 = \frac{1}{2}bh = \frac{1}{2}\cdot 2 \cdot 2 = 2$$

$$A_1 = \frac{1}{2}bh = \frac{1}{2}\cdot 4 \cdot 4 = 8.$$

The total area, then, is

$$A_1 + A_2 = 8 + 2 = 10 \text{ units}^2.$$

Exercise 5.3.4

Find the total area between the function $f(x) = 2x$ and the x-axis over the interval $[-3, 3]$.

Hint

Review the solving strategy in Example 5.3.4

Answer

18 units^2

Properties of the Definite Integral

The properties of indefinite integrals apply to definite integrals as well. Definite integrals also have properties that relate to the limits of integration. These properties, along with the rules of integration that we examine later in this chapter, help us manipulate expressions to evaluate definite integrals.

Rule: Properties of the Definite Integral

1.

$$\int_a^a f(x)\,dx = 0$$

If the limits of integration are the same, the integral is just a line and contains no area.

2.

$$\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx$$

If the limits are reversed, then place a negative sign in front of the integral.

3.

$$\int_a^b [f(x)+g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx$$

The integral of a sum is the sum of the integrals.

4.

$$\int_a^b [f(x)-g(x)]\,dx = \int_a^b f(x)\,dx - \int_a^b g(x)\,dx$$

The integral of a difference is the difference of the integrals.

5.

$$\int_a^b cf(x)\,dx = c\int_a^b f(x)\,dx$$

for constant c. The integral of the product of a constant and a function is equal to the constant multiplied by the integral of the function.

6.

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

Although this formula normally applies when c is between a and b, the formula holds for all values of a, b, and c, provided $f(x)$ is integrable on the largest interval.

Example 5.3.5: Using the Properties of the Definite Integral

Use the properties of the definite integral to express the definite integral of $f(x) = -3x^3 + 2x + 2$ over the interval $[-2, 1]$ as the sum of three definite integrals.

Solution

Using integral notation, we have $\int_{-2}^{1} (-3x^3 + 2x + 2)\,dx$. We apply properties 3. and 5. to get

$$\int_{-2}^{1}(-3x^3+2x+2)\,dx=\int_{-2}^{1}-3x^3\,dx+\int_{-2}^{1}2x\,dx+\int_{-2}^{1}2\,dx$$
$$=-3\int_{-2}^{1}x^3\,dx+2\int_{-2}^{1}x\,dx+\int_{-2}^{1}2\,dx.$$

Exercise 5.3.5

Use the properties of the definite integral to express the definite integral of $f(x)=6x^3-4x^2+2x-3$ over the interval $[1,3]$ as the sum of four definite integrals.

Hint

Use the solving strategy from Example 5.3.5and the properties of definite integrals.

Answer

$$6\int_1^3 x^3\,dx-4\int_1^3 x^2\,dx+2\int_1^3 x\,dx-\int_1^3 3\,dx$$

Example 5.3.6: Using the Properties of the Definite Integral

If it is known that $\int_0^8 f(x)\,dx=10$ and $\int_0^5 f(x)\,dx=5$, find the value of $\int_5^8 f(x)\,dx$.

Solution

By property 6,

$$\int_a^b f(x)\,dx=\int_a^c f(x)\,dx+\int_c^b f(x)\,dx.$$

Thus,

$$\int_0^8 f(x)\,dx=\int_0^5 f(x)\,dx+\int_5^8 f(x)\,dx$$
$$10=5+\int_5^8 f(x)\,dx$$
$$5=\int_5^8 f(x)\,dx.$$

Exercise 5.3.6

If it is known that $\int_1^5 f(x)\,dx=-3$ and $\int_2^5 f(x)\,dx=4$, find the value of $\int_1^2 f(x)\,dx$.

Hint

Use the solving strategy from Example 5.3.6and the rule on properties of definite integrals.

Answer

-7

Comparison Properties of Integrals

A picture can sometimes tell us more about a function than the results of computations. Comparing functions by their graphs as well as by their algebraic expressions can often give new insight into the process of integration. Intuitively, we might say that if a

function $f(x)$ is above another function $g(x)$, then the area between $f(x)$ and the x-axis is greater than the area between $g(x)$ and the x-axis. This is true depending on the interval over which the comparison is made. The properties of definite integrals are valid whether $a < b$, $a = b$, or $a > b$. The following properties, however, concern only the case $a \leq b$, and are used when we want to compare the sizes of integrals.

Comparison Theorem

i. If $f(x) \geq 0$ for $a \leq x \leq b$, then

$$\int_a^b f(x)\,dx \geq 0.$$

ii. If $f(x) \geq g(x)$ for $a \leq x \leq b$, then

$$\int_a^b f(x)\,dx \geq \int_a^b g(x)\,dx.$$

iii. If m and M are constants such that $m \leq f(x) \leq M$ for $a \leq x \leq b$, then

$$m(b-a) \leq \int_a^b f(x)\,dx \leq M(b-a).$$

Example 5.3.7: Comparing Two Functions over a Given Interval

Compare $f(x) = \sqrt{1+x^2}$ and $g(x) = \sqrt{1+x}$ over the interval $[0, 1]$.

Solution

Graphing these functions is necessary to understand how they compare over the interval $[0, 1]$. Initially, when graphed on a graphing calculator, $f(x)$ appears to be above $g(x)$ everywhere. However, on the interval $[0, 1]$, the graphs appear to be on top of each other. We need to zoom in to see that, on the interval $[0, 1]$, $g(x)$ is above $f(x)$. The two functions intersect at $x = 0$ and $x = 1$ (Figure 5.3.8).

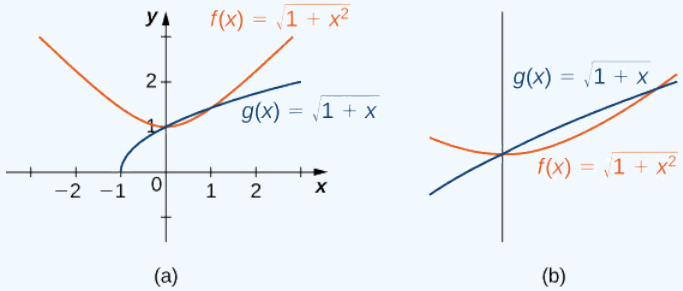


Figure 5.3.8: (a) The function $f(x)$ appears above the function $g(x)$ except over the interval $[0, 1]$ (b) Viewing the same graph with a greater zoom shows this more clearly.

We can see from the graph that over the interval $[0, 1]$, $g(x) \geq f(x)$. Comparing the integrals over the specified interval $[0, 1]$, we also see that $\int_0^1 g(x)\,dx \geq \int_0^1 f(x)\,dx$ (Figure 5.3.9). The thin, red-shaded area shows just how much difference there is between these two integrals over the interval $[0, 1]$.

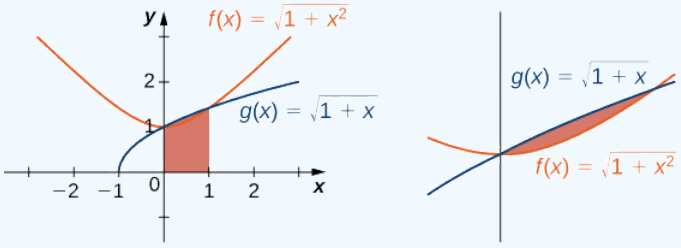


Figure 5.3.9: (a) The graph shows that over the interval $[0,1], g(x) \geq f(x)$, where equality holds only at the endpoints of the interval. (b) Viewing the same graph with a greater zoom shows this more clearly.

Average Value of a Function

We often need to find the average of a set of numbers, such as an average test grade. Suppose you received the following test scores in your algebra class: 89, 90, 56, 78, 100, and 69. Your semester grade is your average of test scores and you want to know what grade to expect. We can find the average by adding all the scores and dividing by the number of scores. In this case, there are six test scores. Thus,

$$\frac{89+90+56+78+100+69}{6} = \frac{482}{6} \approx 80.33.$$

Therefore, your average test grade is approximately 80.33, which translates to a B− at most schools.

Suppose, however, that we have a function $v(t)$ that gives us the speed of an object at any time t, and we want to find the object's average speed. The function $v(t)$ takes on an infinite number of values, so we can't use the process just described. Fortunately, we can use a definite integral to find the average value of a function such as this.

Let $f(x)$ be continuous over the interval $[a,b]$ and let $[a,b]$ be divided into n subintervals of width $\Delta x = (b-a)/n$. Choose a representative x_i^* in each subinterval and calculate $f(x_i^*)$ for $i = 1, 2, \ldots, n$. In other words, consider each $f(x_i^*)$ as a sampling of the function over each subinterval. The average value of the function may then be approximated as

$$f_{ave} \approx \frac{f(x_1^*) + f(x_2^*) + \cdots + f(x_n^*)}{n},$$

which is basically the same expression used to calculate the average of discrete values.

But we know $\Delta x = \dfrac{b-a}{n}$, so $n = \dfrac{b-a}{\Delta x}$, and we get

$$f_{ave} \approx \frac{f(x_1^*) + f(x_2^*) + \cdots + f(x_n^*)}{n} = \frac{f(x_1^*) + f(x_2^*) + \cdots + f(x_n^*)}{\left(\dfrac{b-a}{\Delta x}\right)}.$$

Following through with the algebra, the numerator is a sum that is represented as $\sum_{i=1}^{n} f(x_i^*)$, and we are dividing by a fraction. To divide by a fraction, invert the denominator and multiply. Thus, an approximate value for the average value of the function is given by

$$\frac{\sum_{i=1}^{n} f(x_i^*)}{\left(\dfrac{b-a}{\Delta x}\right)} = \left(\frac{\Delta x}{b-a}\right) \sum_{i=1}^{n} f(x_i^*) = \left(\frac{1}{b-a}\right) \sum_{i=1}^{n} f(x_i^*) \Delta x.$$

This is a Riemann sum. Then, to get the exact average value, take the limit as n goes to infinity. Thus, the average value of a function is given by

$$\frac{1}{b-a} \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i) \Delta x = \frac{1}{b-a} \int_a^b f(x)\,dx.$$

Definition: Average Value of a Function

Let $f(x)$ be continuous over the interval $[a,b]$. Then, the *average value of the function* $f(x)$ (or f_{ave}) on $[a,b]$ is given by

$$f_{ave} = \frac{1}{b-a}\int_a^b f(x)\,dx. \tag{5.3.1}$$

Example 5.3.8: Finding the Average Value of a Linear Function

Find the average value of $f(x) = x+1$ over the interval $[0,5]$.

Solution

First, graph the function on the stated interval, as shown in Figure 5.3.10

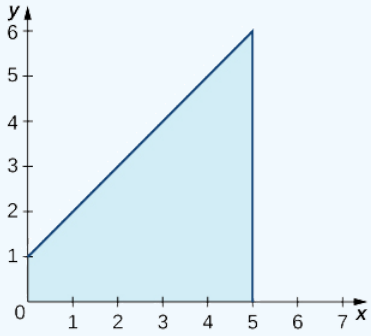


Figure 5.3.10:The graph shows the area under the function $(x) = x+1$ over $[0,5]$.

The region is a trapezoid lying on its side, so we can use the area formula for a trapezoid $A = \frac{1}{2}h(a+b)$, where h represents height, and a and b represent the two parallel sides. Then,

$$\int_0^5 x+1\,dx = \frac{1}{2}h(a+b) = \frac{1}{2}\cdot 5\cdot(1+6) = \frac{35}{2}\quad .$$

Thus the average value of the function is

$$\frac{1}{5-0}\int_0^5 x+1\,dx = \frac{1}{5}\cdot\frac{35}{2} = \frac{7}{2}\ .$$

Exercise 5.3.7

Find the average value of $f(x) = 6-2x$ over the interval $[0,3]$.

Hint

Use the average value formula (Equation 5.3.1), and use geometry to evaluate the integral.

Answer

3

Key Concepts

- The definite integral can be used to calculate net signed area, which is the area above the x-axis less the area below the x-axis. Net signed area can be positive, negative, or zero.
- The component parts of the definite integral are the integrand, the variable of integration, and the limits of integration.
- Continuous functions on a closed interval are integrable. Functions that are not continuous may still be integrable, depending on the nature of the discontinuities.
- The properties of definite integrals can be used to evaluate integrals.
- The area under the curve of many functions can be calculated using geometric formulas.

- The average value of a function can be calculated using definite integrals.

Key Equations

- **Definite Integral**

$$\int_a^b f(x)\,dx = \lim_{n\to\infty}\sum_{i=1}^{n} f(x_i^*)\Delta x$$

- **Properties of the Definite Integral**

$$\int_a^a f(x)\,dx = 0$$

$$\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx$$

$$\int_a^b [f(x)+g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx$$

$$\int_a^b [f(x)-g(x)]\,dx = \int_a^b f(x)\,dx - \int_a^b g(x)\,dx$$

$$\int_a^b cf(x)\,dx = c\int_a^b f(x)\,dx, \text{ for constant } c$$

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

Glossary

average value of a function

(or f_{ave}) the average value of a function on an interval can be found by calculating the definite integral of the function and dividing that value by the length of the interval

definite integral

a primary operation of calculus; the area between the curve and the x-axis over a given interval is a definite integral

integrable function

a function is integrable if the limit defining the integral exists; in other words, if the limit of the Riemann sums as n goes to infinity exists

integrand

the function to the right of the integration symbol; the integrand includes the function being integrated

limits of integration

these values appear near the top and bottom of the integral sign and define the interval over which the function should be integrated

net signed area

the area between a function and the x-axis such that the area below the x-axis is subtracted from the area above the x-axis; the result is the same as the definite integral of the function

total area

total area between a function and the x-axis is calculated by adding the area above the x-axis and the area below the x-axis; the result is the same as the definite integral of the absolute value of the function

variable of integration

indicates which variable you are integrating with respect to; if it is x, then the function in the integrand is followed by dx

This page titled 5.3: The Definite Integral is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin “Jed” Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **5.2: The Definite Integral** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

5.4: The Fundamental Theorem of Calculus

Learning Objectives

- Describe the meaning of the Mean Value Theorem for Integrals.
- State the meaning of the Fundamental Theorem of Calculus, Part 1.
- Use the Fundamental Theorem of Calculus, Part 1, to evaluate derivatives of integrals.
- State the meaning of the Fundamental Theorem of Calculus, Part 2.
- Use the Fundamental Theorem of Calculus, Part 2, to evaluate definite integrals.
- Explain the relationship between differentiation and integration.

In the previous two sections, we looked at the definite integral and its relationship to the area under the curve of a function. Unfortunately, so far, the only tools we have available to calculate the value of a definite integral are geometric area formulas and limits of Riemann sums, and both approaches are extremely cumbersome. In this section we look at some more powerful and useful techniques for evaluating definite integrals.

These new techniques rely on the relationship between differentiation and integration. This relationship was discovered and explored by both Sir Isaac Newton and Gottfried Wilhelm Leibniz (among others) during the late 1600s and early 1700s, and it is codified in what we now call the **Fundamental Theorem of Calculus**, which has two parts that we examine in this section. Its very name indicates how central this theorem is to the entire development of calculus.

Isaac Newton's contributions to mathematics and physics changed the way we look at the world. The relationships he discovered, codified as Newton's laws and the law of universal gravitation, are still taught as foundational material in physics today, and his calculus has spawned entire fields of mathematics.

Before we get to this crucial theorem, however, let's examine another important theorem, the Mean Value Theorem for Integrals, which is needed to prove the Fundamental Theorem of Calculus.

The Mean Value Theorem for Integrals

The Mean Value Theorem for Integrals states that a continuous function on a closed interval takes on its average value at some point in that interval. The theorem guarantees that if $f(x)$ is continuous, a point c exists in an interval $[a, b]$ such that the value of the function at c is equal to the average value of $f(x)$ over $[a, b]$. We state this theorem mathematically with the help of the formula for the average value of a function that we presented at the end of the preceding section.

Theorem 5.4.1: The Mean Value Theorem for Integrals

If $f(x)$ is continuous over an interval $[a, b]$, then there is at least one point $c \in [a, b]$ such that

$$f(c) = \frac{1}{b-a} \int_a^b f(x)\, dx.$$

This formula can also be stated as

$$\int_a^b f(x)\, dx = f(c)(b-a). \tag{5.4.1}$$

Since $f(x)$ is continuous on $[a, b]$, by the extreme value theorem (see section on Maxima and Minima), it assumes minimum and maximum values—m and M, respectively—on $[a, b]$. Then, for all x in $[a, b]$, we have $m \leq f(x) \leq M$. Therefore, by the comparison theorem (see Section on The Definite Integral), we have

$$m(b-a) \leq \int_a^b f(x)\, dx \leq M(b-a).$$

Proof

Since $f(x)$ is continuous on $[a,b]$, by the extreme value theorem (see section on Maxima and Minima), it assumes minimum and maximum values—m and M, respectively—on $[a,b]$. Then, for all x in $[a,b]$, we have $m \le f(x) \le M$. Therefore, by the comparison theorem (see Section on The Definite Integral), we have

$$m(b-a) \le \int_a^b f(x)\,dx \le M(b-a).$$

Dividing by $b-a$ gives us

$$m \le \frac{1}{b-a}\int_a^b f(x)\,dx \le M.$$

Since $\dfrac{1}{b-a}\displaystyle\int_a^b f(x)\,dx$ is a number between m and M, and since $f(x)$ is continuous and assumes the values m and M over $[a,b]$, by the Intermediate Value Theorem, there is a number c over $[a,b]$ such that

$$f(c) = \frac{1}{b-a}\int_a^b f(x)\,dx,$$

and the proof is complete.

□

Example 5.4.1: Finding the Average Value of a Function

Find the average value of the function $f(x) = 8 - 2x$ over the interval $[0,4]$ and find c such that $f(c)$ equals the average value of the function over $[0,4]$.

Solution

The formula states the mean value of $f(x)$ is given by

$$\frac{1}{4-0}\int_0^4 (8-2x)\,dx.$$

We can see in Figure 5.4.1 that the function represents a straight line and forms a right triangle bounded by the x- and y-axes. The area of the triangle is $A = \frac{1}{2}(base)(height)$. We have

$$A = \frac{1}{2}(4)(8) = 16.$$

The average value is found by multiplying the area by $1/(4-0)$. Thus, the average value of the function is

$$\frac{1}{4}(16) = 4$$

Set the average value equal to $f(c)$ and solve for c.

$$8 - 2c = 4$$
$$c = 2$$

At $c = 2$, $f(2) = 4$.

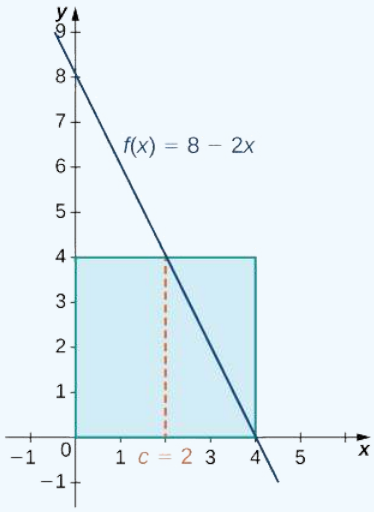


Figure 5.4.1: By the Mean Value Theorem, the continuous function $f(x)$ takes on its average value at c at least once over a closed interval.

Exercise 5.4.1

Find the average value of the function $f(x) = \dfrac{x}{2}$ over the interval $[0, 6]$ and find c such that $f(c)$ equals the average value of the function over $[0, 6]$.

Hint

Use the procedures from Example 5.4.1to solve the problem

Answer

The average value is 1.5 and $c = 3$.

Example 5.4.2: Finding the Point Where a Function Takes on Its Average Value

Given $\int_0^3 x^2\,dx = 9$, find c such that $f(c)$ equals the average value of $f(x) = x^2$ over $[0, 3]$.

Solution

We are looking for the value of c such that

$$f(c) = \frac{1}{3-0}\int_0^3 x^2\,dx = \frac{1}{3}(9) = 3.$$

Replacing $f(c)$ with c^2, we have

$$c^2 = 3$$
$$c = \pm\sqrt{3}.$$

Since $-\sqrt{3}$ is outside the interval, take only the positive value. Thus, $c = \sqrt{3}$ (Figure 5.4.2).

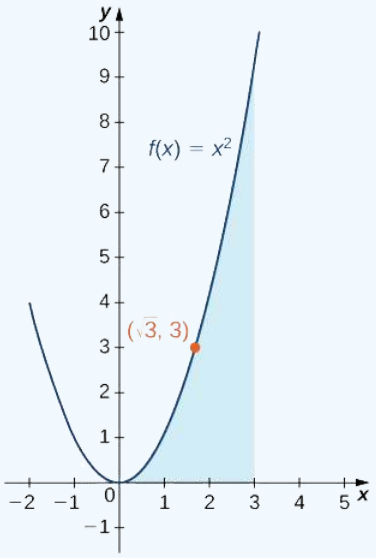


Figure 5.4.2: Over the interval $[0,3]$, the function $f(x)=x^2$ takes on its average value at $c=\sqrt{3}$.

? Exercise 5.4.2

Given $\int_0^3 (2x^2-1)\,dx = 15$, find c such that $f(c)$ equals the average value of $f(x)=2x^2-1$ over $[0,3]$.

Hint

Use the procedures from Example 5.4.2to solve the problem.

Answer

$c=\sqrt{3}$

Fundamental Theorem of Calculus Part 1: Integrals and Antiderivatives

As mentioned earlier, the Fundamental Theorem of Calculus is an extremely powerful theorem that establishes the relationship between differentiation and integration, and gives us a way to evaluate definite integrals without using Riemann sums or calculating areas. The theorem is comprised of two parts, the first of which, the **Fundamental Theorem of Calculus, Part 1,** is stated here. Part 1 establishes the relationship between differentiation and integration.

Theorem 5.4.2: The Fundamental Theorem of Calculus, Part 1

If $f(x)$ is continuous over an interval $[a,b]$, and the function $F(x)$ is defined by

$$F(x)=\int_a^x f(t)\,dt,$$

then $F'(x)=f(x)$ over $[a,b]$.

Before we delve into the proof, a couple of subtleties are worth mentioning here. First, a comment on the notation. Note that we have defined a function, $F(x)$, as the definite integral of another function, $f(t)$, from the point a to the point x. At first glance, this is confusing, because we have said several times that a definite integral is a number, and here it looks like it's a function. The key here is to notice that for any particular value of x, the definite integral is a number. So the function $F(x)$ returns a number (the value of the definite integral) for each value of x.

Second, it is worth commenting on some of the key implications of this theorem. There is a reason it is called the *Fundamental Theorem of Calculus*. Not only does it establish a relationship between integration and differentiation, but also it guarantees that any integrable function has an antiderivative. Specifically, it guarantees that any continuous function has an antiderivative.

Proof: Fundamental Theorem of Calculus, Part 1

Applying the definition of the derivative, we have

$$\begin{aligned} F'(x) &= \lim_{h\to 0} \frac{F(x+h)-F(x)}{h} \\ &= \lim_{h\to 0} \frac{1}{h}\left[\int_a^{x+h} f(t)dt - \int_a^x f(t)\,dt\right] \\ &= \lim_{h\to 0} \frac{1}{h}\left[\int_a^{x+h} f(t)\,dt + \int_x^a f(t)\,dt\right] \\ &= \lim_{h\to 0} \frac{1}{h}\int_x^{x+h} f(t)\,dt. \end{aligned}$$

Looking carefully at this last expression, we see $\frac{1}{h}\int_x^{x+h} f(t)\,dt$ is just the average value of the function $f(x)$ over the interval $[x, x+h]$. Therefore, by Equation 5.4.1, there is some number c in $[x, x+h]$ such that

$$\frac{1}{h}\int_x^{x+h} f(t)\,dt = f(c).$$

In addition, since c is between x and h, c approaches x as h approaches zero. Also, since $f(x)$ is continuous, we have

$$\lim_{h\to 0} f(c) = \lim_{c\to x} f(c) = f(x)$$

Putting all these pieces together, we have

$$F'(x) = \lim_{h\to 0} \frac{1}{h}\int_x^{x+h} f(t)\,dt = \lim_{h\to 0} f(c) = f(x),$$

and the proof is complete.

□

Example 5.4.3: Finding a Derivative with the Fundamental Theorem of Calculus

Use the Fundamental Theorem of Calculus, Part 1 to find the derivative of

$$g(x) = \int_1^x \frac{1}{t^3+1}\,dt.$$

Solution

According to the Fundamental Theorem of Calculus, the derivative is given by

$$g'(x) = \frac{1}{x^3+1}.$$

Exercise 5.4.3

Use the Fundamental Theorem of Calculus, Part 1 to find the derivative of $g(r) = \int_0^r \sqrt{x^2+4}\,dx$.

Hint

Follow the procedures from Example 5.4.3to solve the problem.

Answer

$g'(r) = \sqrt{r^2+4}$

Example 5.4.4: Using the Fundamental Theorem and the Chain Rule to Calculate Derivatives

Let $F(x) = \int_1^{\sqrt{x}} \sin t \, dt$. Find $F'(x)$.

Solution

Letting $u(x) = \sqrt{x}$, we have $F(x) = \int_1^{u(x)} \sin t \, dt$.

Thus, by the Fundamental Theorem of Calculus and the chain rule,

$$F'(x) = \sin(u(x))\frac{du}{dx} = \sin(u(x)) \cdot \left(\frac{1}{2}x^{-1/2}\right) = \frac{\sin\sqrt{x}}{2\sqrt{x}}.$$

Exercise 5.4.4

Let $F(x) = \int_1^{x^3} \cos t \, dt$. Find $F'(x)$.

Hint

Use the chain rule to solve the problem.

Answer

$F'(x) = 3x^2 \cos x^3$

Example 5.4.5: Using the Fundamental Theorem of Calculus with Two Variable Limits of Integration

Let $F(x) = \int_x^{2x} t^3 \, dt$. Find $F'(x)$.

Solution

We have $F(x) = \int_x^{2x} t^3 \, dt$. Both limits of integration are variable, so we need to split this into two integrals. We get

$$\begin{aligned} F(x) &= \int_x^{2x} t^3 \, dt = \int_x^0 t^3 \, dt + \int_0^{2x} t^3 \, dt \\ &= -\int_0^x t^3 \, dt + \int_0^{2x} t^3 \, dt. \end{aligned}$$

Differentiating the first term, we obtain

$$\frac{d}{dx}\left[-\int_0^x t^3 \, dt\right] = -x^3.$$

Differentiating the second term, we first let $u(x) = 2x$. Then,

$$\begin{aligned} \frac{d}{dx}\left[\int_0^{2x} t^3 \, dt\right] &= \frac{d}{dx}\left[\int_0^{u(x)} t^3 \, dt\right] \\ &= (u(x))^3 \frac{du}{dx} \\ &= (2x)^3 \cdot 2 = 16x^3. \end{aligned}$$

Thus,

$$F'(x) = \frac{d}{dx}\left[-\int_0^x t^3\,dt\right] + \frac{d}{dx}\left[\int_0^{2x} t^3\,dt\right]$$
$$= -x^3 + 16x^3 = 15x^3$$

? Exercise 5.4.5

Let $F(x) = \int_x^{x^2} \cos t\,dt$. Find $F'(x)$.

Hint

Use the procedures from Example 5.4.5to solve the problem

Answer

$F'(x) = 2x\cos x^2 - \cos x$

Fundamental Theorem of Calculus, Part 2: The Evaluation Theorem

The Fundamental Theorem of Calculus, Part 2, is perhaps the most important theorem in calculus. After tireless efforts by mathematicians for approximately 500 years, new techniques emerged that provided scientists with the necessary tools to explain many phenomena. Using calculus, astronomers could finally determine distances in space and map planetary orbits. Everyday financial problems such as calculating marginal costs or predicting total profit could now be handled with simplicity and accuracy. Engineers could calculate the bending strength of materials or the three-dimensional motion of objects. Our view of the world was forever changed with calculus.

After finding approximate areas by adding the areas of n rectangles, the application of this theorem is straightforward by comparison. It almost seems too simple that the area of an entire curved region can be calculated by just evaluating an antiderivative at the first and last endpoints of an interval.

Theorem 5.4.3: The Fundamental Theorem of Calculus, Part 2

If $f(x)$ is continuous over the interval $[a, b]$ and $F(x)$ is any antiderivative of $f(x)$, then

$$\int_a^b f(x)\,dx = F(b) - F(a). \tag{5.4.2}$$

We often see the notation $F(x)|_a^b$ to denote the expression $F(b) - F(a)$. We use this vertical bar and associated limits a and b to indicate that we should evaluate the function $F(x)$ at the upper limit (in this case, b), and subtract the value of the function $F(x)$ evaluated at the lower limit (in this case, a).

The **Fundamental Theorem of Calculus, Part 2** (also known as the **evaluation theorem**) states that if we can find an antiderivative for the integrand, then we can evaluate the definite integral by evaluating the antiderivative at the endpoints of the interval and subtracting.

Proof

Let $P = x_i, i = 0, 1, \ldots, n$ be a regular partition of $[a, b]$. Then, we can write

$$\begin{aligned} F(b) - F(a) &= F(x_n) - F(x_0) \\ &= [F(x_n) - F(x_{n-1})] + [F(x_{n-1}) - F(x_{n-2})] + \ldots + [F(x_1) - F(x_0)] \\ &= \sum_{i=1}^{n} [F(x_i) - F(x_{i-1})]. \end{aligned}$$

Now, we know F is an antiderivative of f over $[a, b]$, so by the Mean Value Theorem for derivatives (see The Mean Value Theorem) for $i = 0, 1, \ldots, n$ we can find c_i in $[x_{i-1}, x_i]$ such that

$$F(x_i) - F(x_{i-1}) = F'(c_i)(x_i - x_{i-1}) = f(c_i)\,\Delta x.$$

Then, substituting into the previous equation, we have

$$F(b) - F(a) = \sum_{i=1}^{n} f(c_i)\,\Delta x.$$

Taking the limit of both sides as $n \to \infty$, we obtain

$$F(b) - F(a) = \lim_{n\to\infty} \sum_{i=1}^{n} f(c_i)\Delta x = \int_a^b f(x)\,dx.$$

□

Example 5.4.6: Evaluating an Integral with the Fundamental Theorem of Calculus

Use Equation 5.4.2 to evaluate

$$\int_{-2}^{2} (t^2 - 4)\,dt.$$

Solution

Recall the power rule for Antiderivatives:

If $y = x^n$,

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C.$$

Use this rule to find the antiderivative of the function and then apply the theorem. We have

$$\begin{aligned}
\int_{-2}^{2} (t^2 - 4)dt &= \left(\frac{t^3}{3} - 4t\right)\Big|_{-2}^{2} \\
&= \left[\frac{(2)^3}{3} - 4(2)\right] - \left[\frac{(-2)^3}{3} - 4(-2)\right] \\
&= \left[\frac{8}{3} - 8\right] - \left[-\frac{8}{3} + 8\right] \\
&= \frac{8}{3} - 8 + \frac{8}{3} - 8 \\
&= \frac{16}{3} - 16 = -\frac{32}{3}.
\end{aligned}$$

Analysis

Notice that we did not include the "$+C$" term when we wrote the antiderivative. The reason is that, according to the Fundamental Theorem of Calculus, Part 2 (Equation 5.4.2), any antiderivative works. So, for convenience, we chose the antiderivative with $C = 0$. If we had chosen another antiderivative, the constant term would have canceled out. This always happens when evaluating a definite integral.

The region of the area we just calculated is depicted in Figure 5.4.3. Note that the region between the curve and the x-axis is all below the x-axis. Area is always positive, but a definite integral can still produce a negative number (a net signed area). For example, if this were a profit function, a negative number indicates the company is operating at a loss over the given interval.

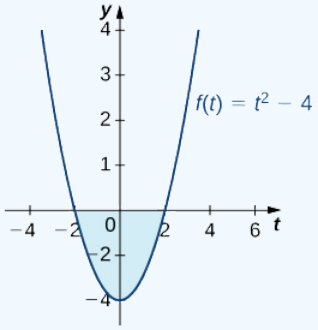


Figure 5.4.3: The evaluation of a definite integral can produce a negative value, even though area is always positive.

Example 5.4.7: Evaluating a Definite Integral Using the Fundamental Theorem of Calculus, Part 2

Evaluate the following integral using the Fundamental Theorem of Calculus, Part 2 (Equation 5.4.2):

$$\int_1^9 \frac{x-1}{\sqrt{x}}\,dx.$$

Solution

First, eliminate the radical by rewriting the integral using rational exponents. Then, separate the numerator terms by writing each one over the denominator:

$$\int_1^9 \frac{x-1}{x^{1/2}}\,dx = \int_1^9 \left(\frac{x}{x^{1/2}} - \frac{1}{x^{1/2}}\right)\,dx.$$

Use the properties of exponents to simplify:

$$\int_1^9 \left(\frac{x}{x^{1/2}} - \frac{1}{x^{1/2}}\right)\,dx = \int_1^9 (x^{1/2} - x^{-1/2})\,dx.$$

Now, integrate using the power rule:

$$\begin{aligned}\int_1^9 (x^{1/2} - x^{-1/2})\,dx &= \left(\frac{x^{3/2}}{\frac{3}{2}} - \frac{x^{1/2}}{\frac{1}{2}}\right)\Bigg|_1^9 \\ &= \left[\frac{(9)^{3/2}}{\frac{3}{2}} - \frac{(9)^{1/2}}{\frac{1}{2}}\right] - \left[\frac{(1)^{3/2}}{\frac{3}{2}} - \frac{(1)^{1/2}}{\frac{1}{2}}\right] \\ &= \left[\frac{2}{3}(27) - 2(3)\right] - \left[\frac{2}{3}(1) - 2(1)\right] \\ &= 18 - 6 - \frac{2}{3} + 2 = \frac{40}{3}.\end{aligned}$$

See Figure 5.4.4.

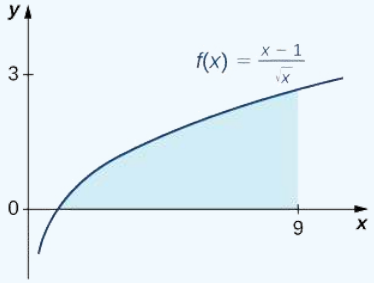


Figure 5.4.4: The area under the curve from $x = 1$ to $x = 9$ can be calculated by evaluating a definite integral.

Exercise 5.4.6

Use Note to evaluate $\int_1^2 x^{-4}\,dx$.

Hint

Use the power rule.

Answer

$\frac{7}{24}$

Applications of Definite Integrals

Example 5.4.8: A Roller-Skating Race

James and Kathy are racing on roller skates. They race along a long, straight track, and whoever has gone the farthest after 5 sec wins a prize. If James can skate at a velocity of $f(t) = 5 + 2t$ ft/sec and Kathy can skate at a velocity of $g(t) = 10 + \cos\left(\frac{\pi}{2}t\right)$ ft/sec, who is going to win the race?

Solution

We need to integrate both functions over the interval $[0, 5]$ and see which value is bigger. For James, we want to calculate

$$\int_0^5 (5 + 2t)\,dt.$$

Using the power rule, we have

$$\begin{aligned}\int_0^5 (5 + 2t)\,dt &= \left(5t + t^2\right)\Big|_0^5 \\ &= (25 + 25) \\ &= 50.\end{aligned}$$

Thus, James has skated 50 ft after 5 sec. Turning now to Kathy, we want to calculate

$$\int_0^5 10 + \cos\left(\frac{\pi}{2}t\right)\,dt.$$

We know $\sin t$ is an antiderivative of $\cos t$, so it is reasonable to expect that an antiderivative of $\cos\left(\frac{\pi}{2}t\right)$ would involve $\sin\left(\frac{\pi}{2}t\right)$. However, when we differentiate $\sin\left(\pi^2 t\right)$, we get $\pi^2 \cos\left(\pi^2 t\right)$ as a result of the chain rule, so we have to account for this additional coefficient when we integrate. We obtain

$$\begin{aligned}\int_0^5 10 + \cos\left(\frac{\pi}{2}t\right)\,dt &= \left(10t + \frac{2}{\pi}\sin\left(\frac{\pi}{2}t\right)\right)\Big|_0^5 \\ &= \left(50 + \frac{2}{\pi}\right) - \left(0 - \frac{2}{\pi}\sin 0\right) \approx 50.6.\end{aligned}$$

Kathy has skated approximately 50.6 ft after 5 sec. Kathy wins, but not by much!

Exercise 5.4.7

Suppose James and Kathy have a rematch, but this time the official stops the contest after only 3 sec. Does this change the outcome?

Hint

Change the limits of integration from those in Example 5.4.7.

Answer

Kathy still wins, but by a much larger margin: James skates 24 ft in 3 sec, but Kathy skates 29.3634 ft in 3 sec.

A Parachutist in Free Fall

Julie is an avid skydiver with more than 300 jumps under her belt and has mastered the art of making adjustments to her body position in the air to control how fast she falls. If she arches her back and points her belly toward the ground, she reaches a terminal velocity of approximately 120 mph (176 ft/sec). If, instead, she orients her body with her head straight down, she falls faster, reaching a terminal velocity of 150 mph (220 ft/sec).



Figure 5.4.5: Skydivers can adjust the velocity of their dive by changing the position of their body during the free fall. (credit: Jeremy T. Lock)

Since Julie will be moving (falling) in a downward direction, we assume the downward direction is positive to simplify our calculations. Julie executes her jumps from an altitude of 12,500 ft. After she exits the aircraft, she immediately starts falling at a velocity given by $v(t) = 32t$.

She continues to accelerate according to this velocity function until she reaches terminal velocity. After she reaches terminal velocity, her speed remains constant until she pulls her ripcord and slows down to land.

On her first jump of the day, Julie orients herself in the slower “belly down” position (terminal velocity is 176 ft/sec). Using this information, answer the following questions.

1. How long after she exits the aircraft does Julie reach terminal velocity?
2. Based on your answer to question 1, set up an expression involving one or more integrals that represents the distance Julie falls after 30 sec.
3. If Julie pulls her ripcord at an altitude of 3000 ft, how long does she spend in a free fall?
4. Julie pulls her ripcord at 3000 ft. It takes 5 sec for her parachute to open completely and for her to slow down, during which time she falls another 400 ft. After her canopy is fully open, her speed is reduced to 16 ft/sec. Find the total time Julie spends in the air, from the time she leaves the airplane until the time her feet touch the ground. On Julie’s second jump of the day, she decides she wants to fall a little faster and orients herself in the “head down” position. Her terminal velocity in this position is 220 ft/sec. Answer these questions based on this velocity:
5. How long does it take Julie to reach terminal velocity in this case?
6. Before pulling her ripcord, Julie reorients her body in the “belly down” position so she is not moving quite as fast when her parachute opens. If she begins this maneuver at an altitude of 4000 ft, how long does she spend in a free fall before beginning the reorientation?

Some jumpers wear “**wingsuits**” (Figure 5.4.6). These suits have fabric panels between the arms and legs and allow the wearer to glide around in a free fall, much like a flying squirrel. (Indeed, the suits are sometimes called “flying squirrel suits.”) When wearing these suits, terminal velocity can be reduced to about 30 mph (44 ft/sec), allowing the wearers a much longer time in

the air. Wingsuit flyers still use parachutes to land; although the vertical velocities are within the margin of safety, horizontal velocities can exceed 70 mph, much too fast to land safely.



Figure 5.4.6: The fabric panels on the arms and legs of a wingsuit work to reduce the vertical velocity of a skydiver's fall. (credit: Richard Schneider)

Answer the following question based on the velocity in a wingsuit.

7. If Julie dons a wingsuit before her third jump of the day, and she pulls her ripcord at an altitude of 3000 ft, how long does she get to spend gliding around in the air

Key Concepts

- The Mean Value Theorem for Integrals states that for a continuous function over a closed interval, there is a value c such that $f(c)$ equals the average value of the function.
- The Fundamental Theorem of Calculus, Part 1 shows the relationship between the derivative and the integral.
- The Fundamental Theorem of Calculus, Part 2 is a formula for evaluating a definite integral in terms of an antiderivative of its integrand. The total area under a curve can be found using this formula.

Key Equations

- **Mean Value Theorem for Integrals**

  If $f(x)$ is continuous over an interval $[a, b]$, then there is at least one point $c \in [a, b]$ such that

  $$f(c) = \frac{1}{b-a} \int_a^b f(x)\, dx.$$

- **Fundamental Theorem of Calculus, Part 1**

  If $f(x)$ is continuous over an interval $[a, b]$, and the function $F(x)$ is defined by

  $$F(x) = \int_a^x f(t)\, dt,$$

  then

  $$F\,'(x) = f(x).$$

- **Fundamental Theorem of Calculus, Part 2**

  If f is continuous over the interval $[a, b]$ and $F(x)$ is any antiderivative of $f(x)$, then

  $$\int_a^b f(x)\, dx = F(b) - F(a).$$

Glossary

fundamental theorem of calculus

the theorem, central to the entire development of calculus, that establishes the relationship between differentiation and integration

fundamental theorem of calculus, part 1

uses a definite integral to define an antiderivative of a function

fundamental theorem of calculus, part 2

(also, **evaluation theorem**) we can evaluate a definite integral by evaluating the antiderivative of the integrand at the endpoints of the interval and subtracting

mean value theorem for integrals

guarantees that a point c exists such that $f(c)$ is equal to the average value of the function

---

This page titled 5.4: The Fundamental Theorem of Calculus is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin "Jed" Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **5.3: The Fundamental Theorem of Calculus** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

5.5: Substitution with Indefinite Integrals

Learning Objectives

- Use substitution to evaluate indefinite integrals.

The Fundamental Theorem of Calculus gave us a method to evaluate integrals without using Riemann sums. The drawback of this method, though, is that we must be able to find an antiderivative, and this is not always easy. In this section we examine a technique, called *integration by substitution*, to help us find antiderivatives. Specifically, this method helps us find antiderivatives when the integrand is the result of a chain-rule derivative.

At first, the approach to the substitution procedure may not appear very obvious. However, it is primarily a visual task—that is, the integrand shows you what to do; it is a matter of recognizing the form of the function. So, what are we supposed to see? We are looking for an integrand of the form $f\left[g(x)\right]g'(x)\,dx$. For example, in the integral

$$\int (x^2-3)^3\,2x\,dx. \tag{5.5.1}$$

we have

$$f(x)=x^3$$

and

$$g(x)=x^2-3.$$

Then

$$g'(x)=2x.$$

and

$$f[g(x)]g'(x)=(x^2-3)^3(2x),$$

and we see that our integrand is in the correct form. The method is called substitution because we substitute part of the integrand with the variable u and part of the integrand with du. It is also referred to as *change of variables* because we are changing variables to obtain an expression that is easier to work with for applying the integration rules.

Substitution with Indefinite Integrals

Let $u=g(x)$,, where $g'(x)$ is continuous over an interval, let $f(x)$ be continuous over the corresponding range of g, and let $F(x)$ be an antiderivative of $f(x)$. Then,

$$\begin{aligned}\int f[g(x)]g'(x)\,dx &= \int f(u)\,du \\ &= F(u)+C \\ &= F(g(x))+C\end{aligned}$$

Proof

Let f, g, u, and F be as specified in the theorem. Then

$$\frac{d}{dx}\left[F(g(x))\right]=F'(g(x))g'(x)=f[g(x)]g'(x).$$

Integrating both sides with respect to x, we see that

$$\int f[g(x)]g'(x)\,dx=F(g(x))+C.$$

If we now substitute $u = g(x)$, and $du = g'(x)\,dx$, we get

$$\int f[g(x)]g'(x)\,dx = \int f(u)\,du = F(u) + C = F(g(x)) + C.$$

□

Returning to the problem we looked at originally, we let $u = x^2 - 3$ and then $du = 2x\,dx$.

Rewrite the integral (Equation 5.5.1) in terms of u:

$$\int (x^2 - 3)^3 (2x\,dx) = \int u^3\,du.$$

Using the power rule for integrals, we have

$$\int u^3\,du = \frac{u^4}{4} + C.$$

Substitute the original expression for x back into the solution:

$$\frac{u^4}{4} + C = \frac{(x^2 - 3)^4}{4} + C.$$

We can generalize the procedure in the following Problem-Solving Strategy.

Problem-Solving Strategy: Integration by Substitution

1. Look carefully at the integrand and select an expression $g(x)$ within the integrand to set equal to u. Let's select $g(x)$. such that $g'(x)$ is also part of the integrand.
2. Substitute $u = g(x)$ and $du = g'(x)dx$. into the integral.
3. We should now be able to evaluate the integral with respect to u. If the integral can't be evaluated we need to go back and select a different expression to use as u.
4. Evaluate the integral in terms of u.
5. Write the result in terms of x and the expression $g(x)$.

Example 5.5.1: Using Substitution to Find an Antiderivative

Use substitution to find the antiderivative of $\int 6x(3x^2 + 4)^4\,dx$.

Solution

The first step is to choose an expression for u. We choose $u = 3x^2 + 4$ because then $du = 6x\,dx$ and we already have du in the integrand. Write the integral in terms of u:

$$\int 6x(3x^2 + 4)^4\,dx = \int u^4\,du.$$

Remember that du is the derivative of the expression chosen for u, regardless of what is inside the integrand. Now we can evaluate the integral with respect to u:

$$\int u^4\,du = \frac{u^5}{5} + C = \frac{(3x^2 + 4)^5}{5} + C.$$

Analysis

We can check our answer by taking the derivative of the result of integration. We should obtain the integrand. Picking a value for C of 1, we let $y = \frac{1}{5}(3x^2 + 4)^5 + 1$. We have

$$y = \frac{1}{5}(3x^2 + 4)^5 + 1,$$

so

$$y' = \left(\frac{1}{5}\right) 5(3x^2+4)^4 6x$$
$$= 6x(3x^2+4)^4.$$

This is exactly the expression we started with inside the integrand.

? Exercise 5.5.1

Use substitution to find the antiderivative of $\int 3x^2(x^3-3)^2\,dx$.

Hint

Let $u = x^3 - 3$.

Answer

$$\int 3x^2(x^3-3)^2\,dx = \frac{1}{3}(x^3-3)^3 + C$$

Sometimes we need to adjust the constants in our integral if they don't match up exactly with the expressions we are substituting.

✓ Example 5.5.2: Using Substitution with Alteration

Use substitution to find the antiderivative of

$$\int z\sqrt{z^2-5}\,dz.$$

Solution

Rewrite the integral as $\int z(z^2-5)^{1/2}\,dz$. Let $u = z^2-5$ and $du = 2z\,dz$. Now we have a problem because $du = 2z\,dz$ and the original expression has only $z\,dz$. We have to alter our expression for du or the integral in u will be twice as large as it should be. If we multiply both sides of the du equation by $\frac{1}{2}$. we can solve this problem. Thus,

$$u = z^2 - 5$$
$$du = 2z\,dz$$
$$\frac{1}{2}du = \frac{1}{2}(2z)\,dz = z\,dz.$$

Write the integral in terms of u, but pull the $\frac{1}{2}$ outside the integration symbol:

$$\int z(z^2-5)^{1/2}\,dz = \frac{1}{2}\int u^{1/2}\,du.$$

Integrate the expression in u:

$$\begin{aligned}\frac{1}{2}\int u^{1/2}\,du &= \left(\frac{1}{2}\right)\frac{u^{3/2}}{\frac{3}{2}}+C\\ &= \left(\frac{1}{2}\right)\left(\frac{2}{3}\right)u^{3/2}+C\\ &= \frac{1}{3}u^{3/2}+C\\ &= \frac{1}{3}(z^2-5)^{3/2}+C\end{aligned}$$

Exercise 5.5.2

Use substitution to find the antiderivative of $\int x^2(x^3+5)^9\,dx$.

Hint

Multiply the du equation by $\frac{1}{3}$.

Answer

$$\int x^2(x^3+5)^9\,dx = \frac{(x^3+5)^{10}}{30}+C$$

Example 5.5.3: Using Substitution with Integrals of Trigonometric Functions

Use substitution to evaluate the integral $\int \frac{\sin t}{\cos^3 t}\,dt$.

Solution

We know the derivative of $\cos t$ is $-\sin t$, so we set $u=\cos t$. Then $du=-\sin t\,dt$.

Substituting into the integral, we have

$$\int \frac{\sin t}{\cos^3 t}\,dt = -\int \frac{du}{u^3}.$$

Evaluating the integral, we get

$$-\int \frac{du}{u^3} = -\int u^{-3}\,du = -\left(-\frac{1}{2}\right)u^{-2}+C.$$

Putting the answer back in terms of t, we get

$$\int \frac{\sin t}{\cos^3 t}\,dt = \frac{1}{2u^2}+C = \frac{1}{2\cos^2 t}+C.$$

Exercise 5.5.3

Use substitution to evaluate the integral $\int \frac{\cos t}{\sin^2 t}\,dt$.

Hint

Use the process from Example 5.5.3to solve the problem.

Answer

$$\int \frac{\cos t}{\sin^2 t}\,dt = -\frac{1}{\sin t} + C$$

Exercise 5.5.4

Use substitution to evaluate the indefinite integral $\int \cos^3 t \sin t\,dt$.

Hint

Use the process from Example 5.5.3to solve the problem.

Answer

$$\int \cos^3 t \sin t\,dt = -\frac{\cos^4 t}{4} + C$$

Sometimes we need to manipulate an integral in ways that are more complicated than just multiplying or dividing by a constant. We need to eliminate all the expressions within the integrand that are in terms of the original variable. When we are done, u should be the only variable in the integrand. In some cases, this means solving for the original variable in terms of u. This technique should become clear in the next example.

Example 5.5.4: Finding an Antiderivative Using u-Substitution

Use substitution to find the antiderivative of

$$\int \frac{x}{\sqrt{x-1}}\,dx.$$

Solution

If we let $u = x - 1$, then $du = dx$. But this does not account for the x in the numerator of the integrand. We need to express x in terms of u. If $u = x - 1$, then $x = u + 1$. Now we can rewrite the integral in terms of u :

$$\int \frac{x}{\sqrt{x-1}}\,dx = \int \frac{u+1}{\sqrt{u}}\,du = \int \left(\sqrt{u} + \frac{1}{\sqrt{u}}\right)\,du = \int \left(u^{1/2} + u^{-1/2}\right)\,du.$$

Then we integrate in the usual way, replace u with the original expression, and factor and simplify the result. Thus,

$$\begin{aligned}
\int (u^{1/2} + u^{-1/2})\,du &= \frac{2}{3}u^{3/2} + 2u^{1/2} + C \\
&= \frac{2}{3}(x-1)^{3/2} + 2(x-1)^{1/2} + C \\
&= (x-1)^{1/2}\left[\frac{2}{3}(x-1) + 2\right] + C \\
&= (x-1)^{1/2}\left(\frac{2}{3}x - \frac{2}{3} + \frac{6}{3}\right) \\
&= (x-1)^{1/2}\left(\frac{2}{3}x + \frac{4}{3}\right) \\
&= \frac{2}{3}(x-1)^{1/2}(x+2) + C.
\end{aligned}$$

This page titled 5.5: Substitution with Indefinite Integrals is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Gilbert Strang & Edwin "Jed" Herman (OpenStax) via source content that was edited to the style and standards of the LibreTexts platform.

- **5.5: Substitution** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

5.6: Substitution with Definite Integrals

Learning Objectives

- Use substitution to evaluate definite integrals.

Substitution with Definite Integrals

Let $u = g(x)$ and let g' be continuous over an interval $[a, b]$, and let f be continuous over the range of $u = g(x)$. Then,

$$\int_a^b f(g(x))g'(x)\,dx = \int_{g(a)}^{g(b)} f(u)\,du.$$

Although we will not formally prove this theorem, we justify it with some calculations here. From the substitution rule for indefinite integrals, if $F(x)$ is an antiderivative of $f(x)$, we have

$$\int f(g(x))g'(x)\,dx = F(g(x)) + C.$$

Then

$$\begin{aligned}\int_a^b f[g(x)]g'(x)\,dx &= F(g(x))\Big|_{x=a}^{x=b} \\ &= F(g(b)) - F(g(a)) \\ &= F(u)\Big|_{u=g(a)}^{u=g(b)} \\ &= \int_{g(a)}^{g(b)} f(u)\,du\end{aligned}$$

and we have the desired result.

Example 5.6.1: Using Substitution to Evaluate a Definite Integral

Use substitution to evaluate

$$\int_0^1 x^2(1+2x^3)^5\,dx.$$

Solution

Let $u = 1 + 2x^3$, so $du = 6x^2\,dx$. Since the original function includes one factor of x^2 and $du = 6x^2\,dx$, multiply both sides of the du equation by $1/6$. Then,

$$du = 6x^2\,dx$$

$$\text{becomes} \quad \frac{1}{6}du = x^2\,dx.$$

To adjust the limits of integration, note that when $x = 0, u = 1 + 2(0) = 1,$ and when $x = 1,\ u = 1 + 2(1) = 3.$

Then

$$\int_0^1 x^2(1+2x^3)^5 dx = \frac{1}{6}\int_1^3 u^5\,du.$$

Evaluating this expression, we get

$$\begin{aligned}\frac{1}{6}\int_1^3 u^5\,du &= \left(\frac{1}{6}\right)\left(\frac{u^6}{6}\right)\Big|_1^3 \\ &= \frac{1}{36}\left[(3)^6-(1)^6\right] \\ &= \frac{182}{9}.\end{aligned}$$

? Exercise 5.6.1

Use substitution to evaluate the definite integral $\int_{-1}^{0} y(2y^2-3)^5\,dy$.

Hint

Use the steps from Example 5.6.5to solve the problem.

Answer

$$\int_{-1}^{0} y(2y^2-3)^5\,dy = \frac{91}{3}$$

? Exercise 5.6.2

Use substitution to evaluate $\int_0^1 x^2\cos\left(\frac{\pi}{2}x^3\right)\,dx$.

Hint

Use the process from Example 5.6.5to solve the problem.

Answer

$$\int_0^1 x^2\cos\left(\frac{\pi}{2}x^3\right)\,dx = \frac{2}{3\pi} \approx 0.2122$$

✓ Example 5.6.2: Using Substitution with an Exponential Function

Use substitution to evaluate

$$\int_0^1 xe^{4x^2+3}\,dx.$$

Solution

Let $u = 4x^3+3$. Then, $du = 8x\,dx$. To adjust the limits of integration, we note that when $x=0$, $u=3$, and when $x=1$, $u=7$. So our substitution gives

$$\begin{aligned}\int_0^1 xe^{4x^2+3}\,dx &= \frac{1}{8}\int_3^7 e^u\,du \\ &= \frac{1}{8}e^u\Big|_3^7 \\ &= \frac{e^7-e^3}{8} \\ &\approx 134.568\end{aligned}$$

Substitution may be only one of the techniques needed to evaluate a definite integral. All of the properties and rules of integration apply independently, and trigonometric functions may need to be rewritten using a trigonometric identity before we can apply substitution. Also, we have the option of replacing the original expression for u after we find the antiderivative, which means that we do not have to change the limits of integration. These two approaches are shown in Example 5.6.7.

Example 5.6.3: Using Substitution to Evaluate a Trigonometric Integral

Use substitution to evaluate

$$\int_0^{\pi/2} \cos^2\theta\, d\theta.$$

Solution

Let us first use a trigonometric identity to rewrite the integral. The trig identity $\cos^2\theta = \dfrac{1+\cos 2\theta}{2}$ allows us to rewrite the integral as

$$\int_0^{\pi/2} \cos^2\theta\, d\theta = \int_0^{\pi/2} \frac{1+\cos 2\theta}{2}\, d\theta.$$

Then,

$$\begin{aligned}\int_0^{\pi/2} \left(\frac{1+\cos 2\theta}{2}\right) d\theta &= \int_0^{\pi/2} \left(\frac{1}{2}+\frac{1}{2}\cos 2\theta\right) d\theta \\ &= \frac{1}{2}\int_0^{\pi/2} d\theta + \int_0^{\pi/2} \cos 2\theta\, d\theta.\end{aligned}$$

We can evaluate the first integral as it is, but we need to make a substitution to evaluate the second integral. Let $u = 2\theta$. Then, $du = 2\,d\theta$, or $\dfrac{1}{2}\,du = d\theta$. Also, when $\theta = 0$, $u = 0$, and when $\theta = \pi/2$, $u = \pi$. Expressing the second integral in terms of u, we have

$$\begin{aligned}\frac{1}{2}\int_0^{\pi/2} d\theta + \frac{1}{2}\int_0^{\pi/2} \cos 2\theta\, d\theta &= \frac{1}{2}\int_0^{\pi/2} d\theta + \frac{1}{2}\left(\frac{1}{2}\right)\int_0^{\pi} \cos u\, du \\ &= \frac{\theta}{2}\Big|_{\theta=0}^{\theta=\pi/2} + \frac{1}{4}\sin u\Big|_{u=0}^{u=\theta} \\ &= \left(\frac{\pi}{4}-0\right)+(0-0) = \frac{\pi}{4}\end{aligned}$$

5.6: Substitution with Definite Integrals is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **5.5: Substitution** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

CHAPTER OVERVIEW

6: Applications of Definite Integrals

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

6.1: Volumes Using Cross-Sections
6.2: Volumes Using Cylindrical Shells
6.3: Arc Length
6.4: Areas of Surfaces of Revolution
6.5: Work
6.6: Moments and Centers of Mass

6: Applications of Definite Integrals is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

6.1: Volumes Using Cross-Sections

6.1: Volumes Using Cross-Sections is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

6.2: Volumes Using Cylindrical Shells

Learning Objectives

- Calculate the volume of a solid of revolution by using the method of cylindrical shells.
- Compare the different methods for calculating a volume of revolution.

In this section, we examine the method of cylindrical shells, the final method for finding the volume of a solid of revolution. We can use this method on the same kinds of solids as the disk method or the washer method; however, with the disk and washer methods, we integrate along the coordinate axis parallel to the axis of revolution. With the method of cylindrical shells, we integrate along the coordinate *axis perpendicular* to the axis of revolution. The ability to choose which variable of integration we want to use can be a significant advantage with more complicated functions. Also, the specific geometry of the solid sometimes makes the method of using cylindrical shells more appealing than using the washer method. In the last part of this section, we review all the methods for finding volume that we have studied and lay out some guidelines to help you determine which method to use in a given situation.

The Method of Cylindrical Shells

Again, we are working with a solid of revolution. As before, we define a region R, bounded above by the graph of a function $y = f(x)$, below by the $\boldsymbol{x}$**-axis**, and on the left and right by the lines $x = a$ and $x = b$, respectively, as shown in Figure $6.2.1a$. We then revolve this region around the y-axis, as shown in Figure $6.2.1b$. Note that this is different from what we have done before. Previously, regions defined in terms of functions of x were revolved around the $\boldsymbol{x}$**-axis** or a line parallel to it.

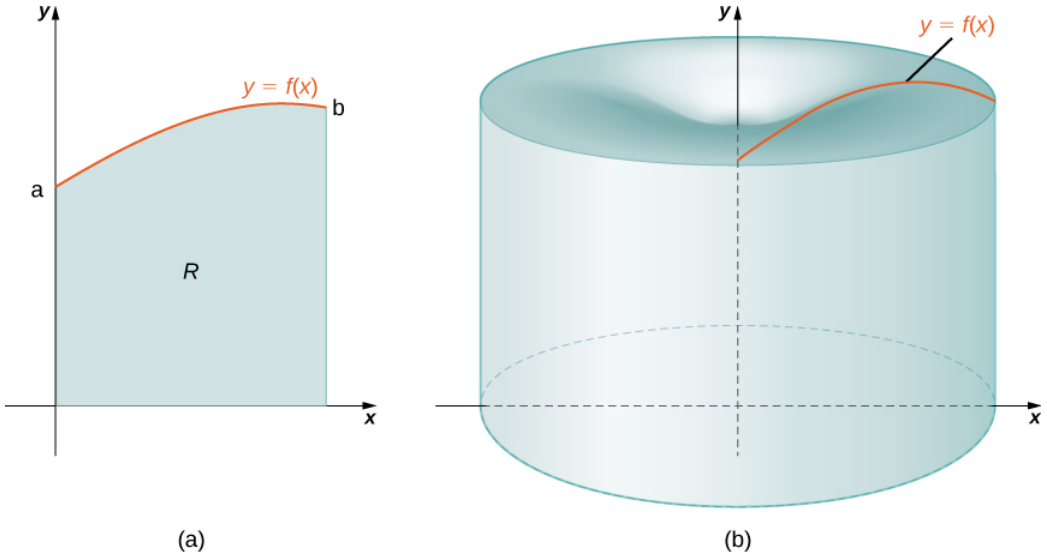


Figure 6.2.1: (a) A region bounded by the graph of a function of x. (b) The solid of revolution formed when the region is revolved around the $\boldsymbol{y}$**-axis.**

As we have done many times before, partition the interval $[a, b]$ using a regular partition, $P = x_0, x_1, \ldots, x_n$ and, for $i = 1, 2, \ldots, n$, choose a point $x_i^* \in [x_{i-1}, x_i]$. Then, construct a rectangle over the interval $[x_{i-1}, x_i]$ of height $f(x_i^*)$ and width Δx. A representative rectangle is shown in Figure $6.2.2a$. When that rectangle is revolved around the y-axis, instead of a disk or a washer, we get a cylindrical shell, as shown in Figure $6.2.2$.

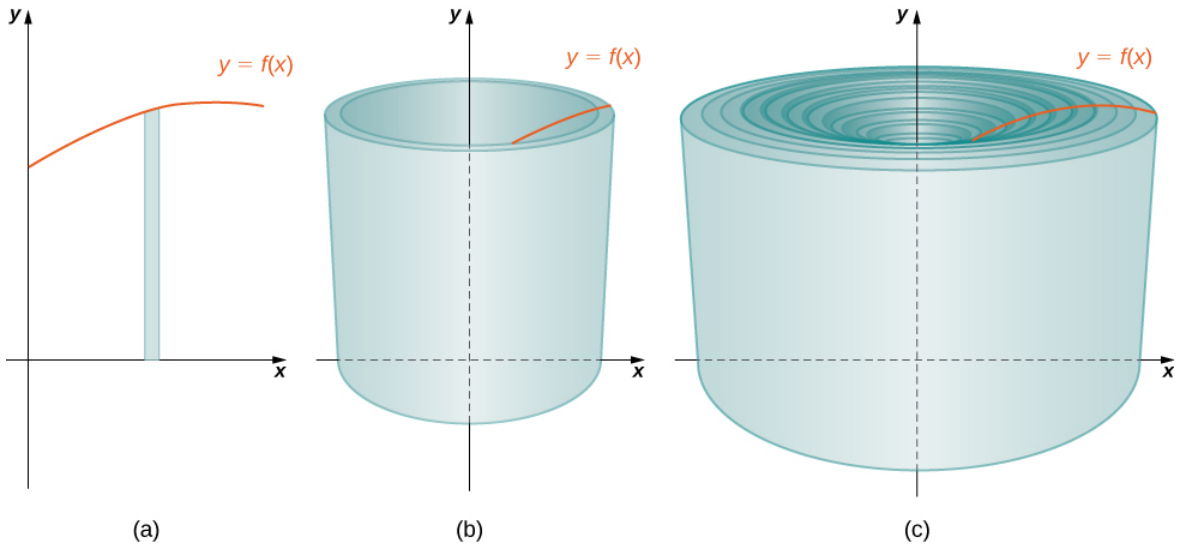


Figure 6.2.2: (a) A representative rectangle. (b) When this rectangle is revolved around the ***y*-axis**, the result is a cylindrical shell. (c) When we put all the shells together, we get an approximation of the original solid.

To calculate the volume of this shell, consider Figure 6.2.3.

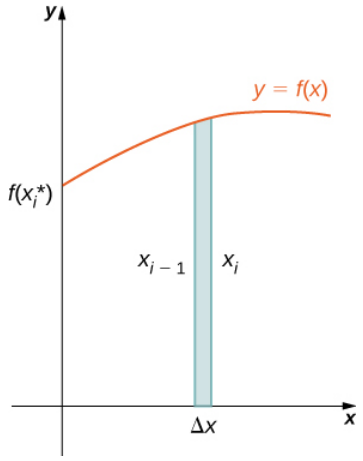


Figure 6.2.3: Calculating the volume of the shell.

The shell is a cylinder, so its volume is the cross-sectional area multiplied by the height of the cylinder. The cross-sections are annuli (ring-shaped regions—essentially, circles with a hole in the center), with outer radius x_i and inner radius x_{i-1}. Thus, the cross-sectional area is $\pi x_i^2 - \pi x_{i-1}^2$. The height of the cylinder is $f(x_i^*)$. Then the volume of the shell is

$$\begin{aligned} V_{shell} &= f(x_i^*)(\pi\, x_i^2 - \pi\, x_{i-1}^2) \\ &= \pi\, f(x_i^*)(x_i^2 - x_{i-1}^2) \\ &= \pi\, f(x_i^*)(x_i + x_{i-1})(x_i - x_{i-1}) \\ &= 2\pi\, f(x_i^*)\left(\frac{x_i + x_{i-1}}{2}\right)(x_i - x_{i-1}). \end{aligned}$$

Note that $x_i - x_{i-1} = \Delta x,$ so we have

$$V_{shell} = 2\pi\, f(x_i^*)\left(\frac{x_i + x_{i-1}}{2}\right)\Delta x.$$

Furthermore, $\dfrac{x_i + x_{i-1}}{2}$ is both the midpoint of the interval $[x_{i-1}, x_i]$ and the average radius of the shell, and we can approximate this by x_i^*. We then have

$$V_{shell} \approx 2\pi\, f(x_i^*) x_i^*\, \Delta x.$$

Another way to think of this is to think of making a vertical cut in the shell and then opening it up to form a flat plate (Figure 6.2.4).

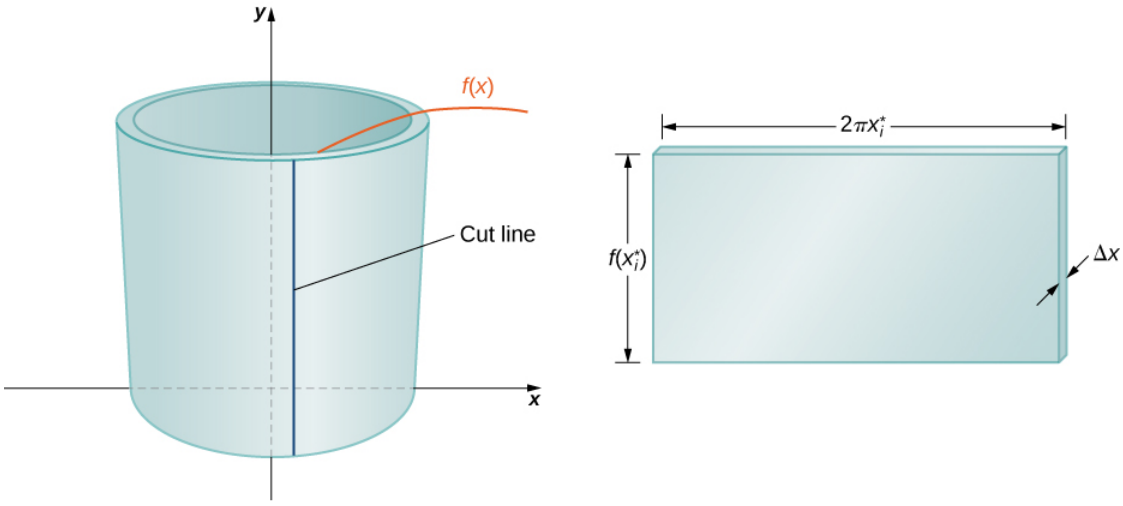


Figure 6.2.4: (a) Make a vertical cut in a representative shell. (b) Open the shell up to form a flat plate.

In reality, the outer radius of the shell is greater than the inner radius, and hence the back edge of the plate would be slightly longer than the front edge of the plate. However, we can approximate the flattened shell by a flat plate of height $f(x_i^*)$, width $2\pi x_i^*$, and thickness Δx (Figure). The volume of the shell, then, is approximately the volume of the flat plate. Multiplying the height, width, and depth of the plate, we get

$$V_{shell} \approx f(x_i^*)(2\pi\, x_i^*)\,\Delta x,$$

which is the same formula we had before.

To calculate the volume of the entire solid, we then add the volumes of all the shells and obtain

$$V \approx \sum_{i=1}^{n} (2\pi\, x_i^* f(x_i^*)\,\Delta x).$$

Here we have another Riemann sum, this time for the function $2\pi\, x\, f(x)$.Taking the limit as $n \to \infty$ gives us

$$V = \lim_{n\to\infty} \sum_{i=1}^{n} (2\pi\, x_i^* f(x_i^*)\,\Delta x) = \int_a^b (2\pi\, x\, f(x))\, dx.$$

This leads to the following rule for the *method of cylindrical shells*.

Rule: The Method of Cylindrical Shells

Let $f(x)$ be continuous and nonnegative. Define R as the region bounded above by the graph of $f(x)$, below by the **x-axis,** on the left by the line $x = a$, and on the right by the line $x = b$. Then the volume of the solid of revolution formed by revolving R around the y-axis is given by

$$V = \int_a^b (2\pi\, x\, f(x))\, dx.$$

Now let's consider an example.

Example 6.2.1: The Method of Cylindrical Shells I

Define R as the region bounded above by the graph of $f(x) = 1/x$ and below by the **x-axis** over the interval $[1, 3]$. Find the volume of the solid of revolution formed by revolving R around the y-axis.

Solution

First we must graph the region R and the associated solid of revolution, as shown in Figure $6.2.5$.

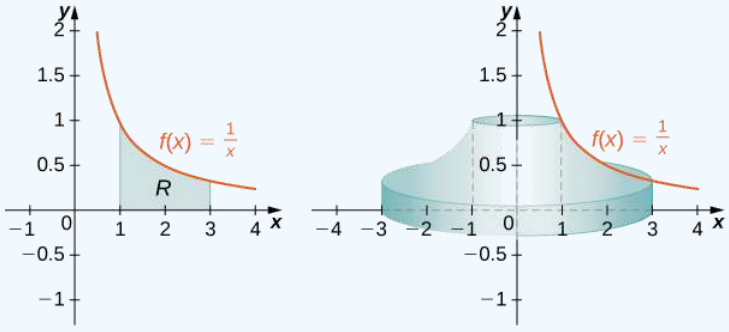


Figure $6.2.5$: (a) The region R under the graph of $f(x) = 1/x$ over the interval $[1, 3]$. (b) The solid of revolution generated by revolving R about the y-axis.

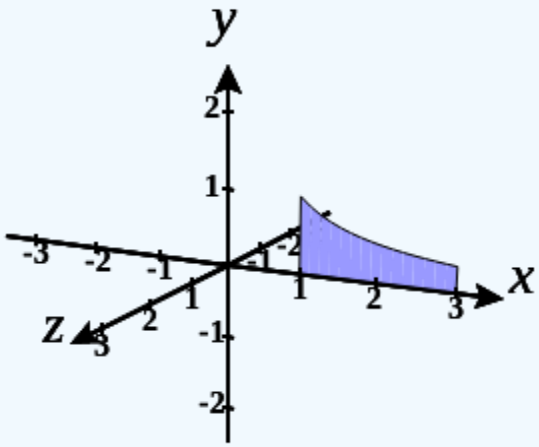


Figure $6.2.5$ (c) Visualizing the solid of revolution with CalcPlot3D.

Then the volume of the solid is given by

$$V = \int_a^b (2\pi\, x\, f(x))\, dx$$
$$= \int_1^3 \left(2\pi\, x \left(\frac{1}{x}\right)\right) dx$$
$$= \int_1^3 2\pi\, dx$$
$$= 2\pi\, x\Big|_1^3 = 4\pi \text{ units}^3.$$

Exercise 6.2.1

Define R as the region bounded above by the graph of $f(x) = x^2$ and below by the x-axis over the interval $[1, 2]$. Find the volume of the solid of revolution formed by revolving R around the y-axis.

Hint

Use the procedure from Example 6.2.1

Answer

$\frac{15\pi}{2}$ units^3

Example 6.2.2: The Method of Cylindrical Shells II

Define R as the region bounded above by the graph of $f(x) = 2x - x^2$ and below by the x-axis over the interval $[0, 2]$. Find the volume of the solid of revolution formed by revolving R around the **y-axis.**

Solution

First graph the region R and the associated solid of revolution, as shown in Figure 6.2.6.

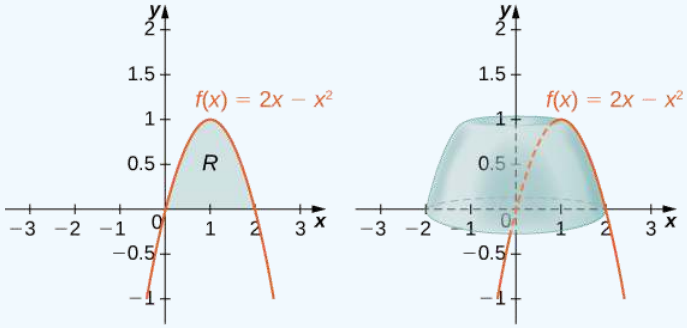


Figure 6.2.6: (a) The region R *under the graph of* $f(x) = 2x - x^2$ *over the interval* $[0, 2]$. *(b) The volume of revolution obtained by revolving* R *about the* y*-axis.*

Then the volume of the solid is given by

$$
\begin{aligned}
V &= \int_a^b (2\pi\, x\, f(x))\, dx \\
&= \int_0^2 (2\pi\, x(2x - x^2))\, dx \\
&= 2\pi \int_0^2 (2x^2 - x^3)\, dx \\
&= 2\pi \left[\frac{2x^3}{3} - \frac{x^4}{4} \right] \Big|_0^2 \\
&= \frac{8\pi}{3} \text{ units}^3
\end{aligned}
$$

Exercise 6.2.2

Define R as the region bounded above by the graph of $f(x) = 3x - x^2$ and below by the x-axis over the interval $[0, 2]$. Find the volume of the solid of revolution formed by revolving R around the y-axis.

Hint

Use the process from Example 6.2.2

Answer

$8\pi \text{ units}^3$

As with the disk method and the washer method, we can use the method of cylindrical shells with solids of revolution, revolved around the x-axis, when we want to integrate with respect to y. The analogous rule for this type of solid is given here.

Rule: The Method of Cylindrical Shells for Solids of Revolution around the x-axis

Let $g(y)$ be continuous and nonnegative. Define Q as the region bounded on the right by the graph of $g(y)$, on the left by the y-axis, below by the line $y = c$, and above by the line $y = d$. Then, the volume of the solid of revolution formed by revolving Q around the x-axis is given by

$$V = \int_c^d (2\pi\, y\, g(y))\, dy.$$

Example 6.2.3: The Method of Cylindrical Shells for a Solid Revolved around the x-axis

Define Q as the region bounded on the right by the graph of $g(y) = 2\sqrt{y}$ and on the left by the y-axis for $y \in [0, 4]$. Find the volume of the solid of revolution formed by revolving Q around the x-axis.

Solution

First, we need to graph the region Q and the associated solid of revolution, as shown in Figure 6.2.7.

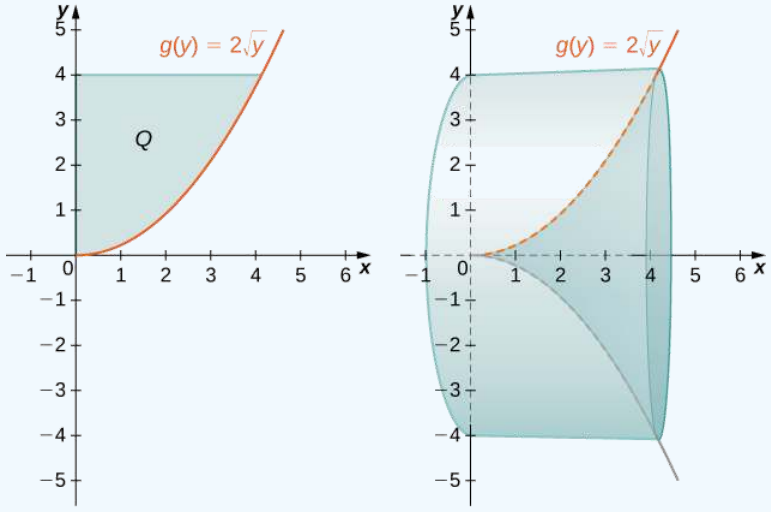


Figure 6.2.7: (a) The region Q to the left of the function $g(y)$ over the interval $[0,4]$. (b) The solid of revolution generated by revolving Q around the x-axis.

Label the shaded region Q. Then the volume of the solid is given by

$$\begin{aligned} V &= \int_c^d (2\pi\, y\, g(y))\, dy \\ &= \int_0^4 (2\pi\, y(2\sqrt{y}))\, dy \\ &= 4\pi \int_0^4 y^{3/2}\, dy \\ &= 4\pi \left[\frac{2y^{5/2}}{5}\right]\Bigg|_0^4 \\ &= \frac{256\pi}{5} \text{ units}^3 \end{aligned}$$

? Exercise 6.2.3

Define Q as the region bounded on the right by the graph of $g(y) = 3/y$ and on the left by the y-axis for $y \in [1,3]$. Find the volume of the solid of revolution formed by revolving Q around the x-axis.

Hint

Use the process from Example 6.2.3

Answer

12π units^3

For the next example, we look at a solid of revolution for which the graph of a function is revolved around a line other than one of the two coordinate axes. To set this up, we need to revisit the development of the method of cylindrical shells. Recall that we found the volume of one of the shells to be given by

$$\begin{aligned} V_{shell} &= f(x_i^*)(\pi\, x_i^2 - \pi\, x_{i-1}^2) \\ &= \pi\, f(x_i^*)(x_i^2 - x_{i-1}^2) \\ &= \pi\, f(x_i^*)(x_i + x_{i-1})(x_i - x_{i-1}) \\ &= 2\pi\, f(x_i^*)\left(\frac{x_i + x_{i-1}}{2}\right)(x_i - x_{i-1}). \end{aligned}$$

This was based on a shell with an outer radius of x_i and an inner radius of x_{i-1}. If, however, we rotate the region around a line other than the y-axis, we have a different outer and inner radius. Suppose, for example, that we rotate the region around the line $x = -k$, where k is some positive constant. Then, the outer radius of the shell is $x_i + k$ and the inner radius of the shell is $x_{i-1} + k$. Substituting these terms into the expression for volume, we see that when a plane region is rotated around the line $x = -k$, the volume of a shell is given by

$$\begin{aligned} V_{shell} &= 2\pi\, f(x_i^*)\left(\frac{(x_i + k) + (x_{i-1} + k)}{2}\right)((x_i + k) - (x_{i-1} + k)) \\ &= 2\pi\, f(x_i^*)\left(\left(\frac{x_i + x_{i-2}}{2}\right) + k\right)\Delta x. \end{aligned}$$

As before, we notice that $\dfrac{x_i + x_{i-1}}{2}$ is the midpoint of the interval $[x_{i-1}, x_i]$ and can be approximated by x_i^*. Then, the approximate volume of the shell is

$$V_{shell} \approx 2\pi(x_i^* + k) f(x_i^*)\Delta x.$$

The remainder of the development proceeds as before, and we see that

$$V = \int_a^b (2\pi(x + k) f(x))dx.$$

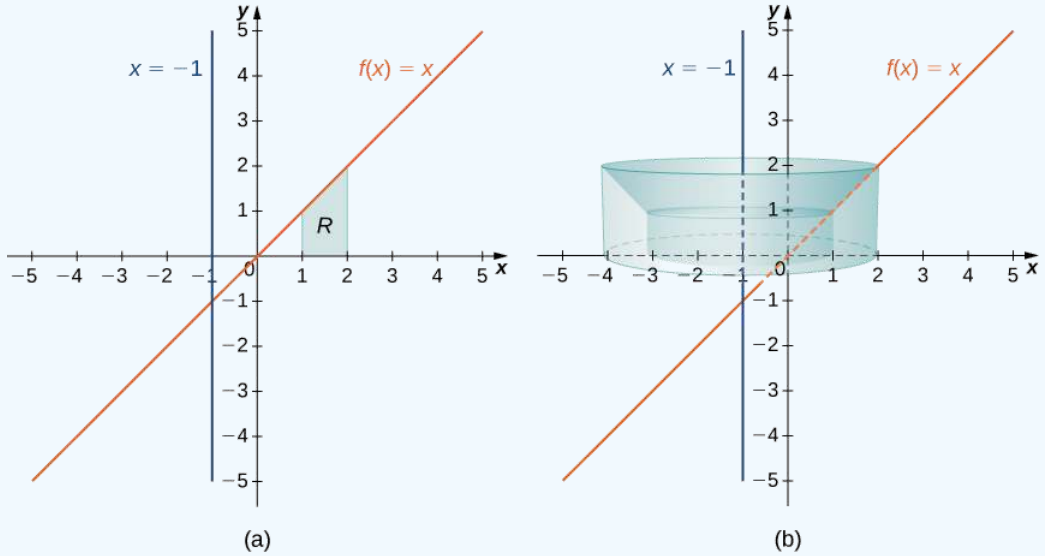
We could also rotate the region around other horizontal or vertical lines, such as a vertical line in the right half plane. In each case, the volume formula must be adjusted accordingly. Specifically, the x-term in the integral must be replaced with an expression representing the radius of a shell. To see how this works, consider the following example.

Example 6.2.4: A Region of Revolution Revolved around a Line

Define R as the region bounded above by the graph of $f(x) = x$ and below by the x-axis over the interval $[1, 2]$. Find the volume of the solid of revolution formed by revolving R around the line $x = -1$.

Solution

First, graph the region R and the associated solid of revolution, as shown in Figure 6.2.8.



(a) (b)

Figure 6.2.8: (a) The region R between the graph of $f(x)$ and the x-axis over the interval $[1,2]$. (b) The solid of revolution generated by revolving R around the line $x=-1$.

Note that the radius of a shell is given by $x+1$. Then the volume of the solid is given by

$$\begin{aligned} V &= \int_1^2 2\pi(x+1)f(x)\,dx \\ &= \int_1^2 2\pi(x+1)x\,dx = 2\pi\int_1^2 x^2+x\,dx \\ &= 2\pi\left[\frac{x^3}{3}+\frac{x^2}{2}\right]\Big|_1^2 \\ &= \frac{23\pi}{3}\text{ units}^3 \end{aligned}$$

Exercise 6.2.4

Define R as the region bounded above by the graph of $f(x)=x^2$ and below by the x-axis over the interval $[0,1]$. Find the volume of the solid of revolution formed by revolving R around the line $x=-2$.

Hint

Use the process from Example 6.2.4

Answer

$\frac{11\pi}{6}$ units^3

For our final example in this section, let's look at the volume of a solid of revolution for which the region of revolution is bounded by the graphs of two functions.

Example 6.2.5: A Region of Revolution Bounded by the Graphs of Two Functions

Define R as the region bounded above by the graph of the function $f(x)=\sqrt{x}$ and below by the graph of the function $g(x)=1/x$ over the interval $[1,4]$. Find the volume of the solid of revolution generated by revolving R around the y-axis.

Solution

First, graph the region R and the associated solid of revolution, as shown in Figure 6.2.9.

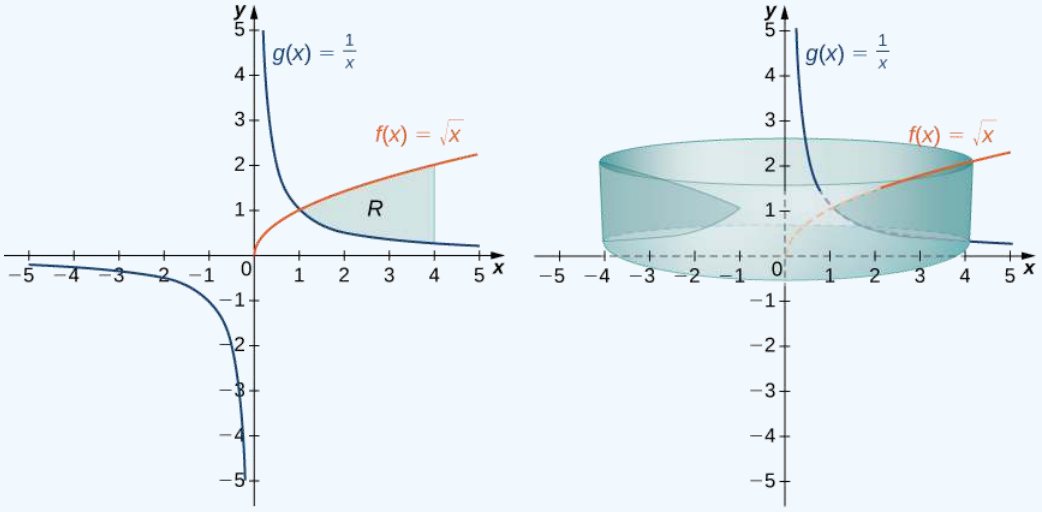


Figure 6.2.9: (a) The region R between the graph of $f(x)$ and the graph of $g(x)$ over the interval $[1,4]$. (b) The solid of revolution generated by revolving R around the y-axis.

Note that the axis of revolution is the y-axis, so the radius of a shell is given simply by x. We don't need to make any adjustments to the x-term of our integrand. The height of a shell, though, is given by $f(x)-g(x)$, so in this case we need to adjust the $f(x)$ term of the integrand. Then the volume of the solid is given by

$$V = \int_1^4 (2\pi\, x(f(x)-g(x)))\, dx$$

$$= \int_1^4 (2\pi\, x(\sqrt{x} - \frac{1}{x}))\, dx = 2\pi \int_1^4 (x^{3/2} - 1) dx$$

$$= 2\pi \left[\frac{2x^{5/2}}{5} - x\right]\Big|_1^4 = \frac{94\pi}{5} \text{ units}^3.$$

? Exercise 6.2.5

Define R as the region bounded above by the graph of $f(x) = x$ and below by the graph of $g(x) = x^2$ over the interval $[0,1]$. Find the volume of the solid of revolution formed by revolving R around the y-axis.

Hint

Hint: Use the process from Example 6.2.5

Answer

$\frac{\pi}{6}$ units^3

Which Method Should We Use?

We have studied several methods for finding the volume of a solid of revolution, but how do we know which method to use? It often comes down to a choice of which integral is easiest to evaluate. Figure 6.2.10 describes the different approaches for solids of revolution around the x-axis. It's up to you to develop the analogous table for solids of revolution around the y-axis.

Comparing the Methods for Finding the Volume of a Solid Revolution around the x-axis

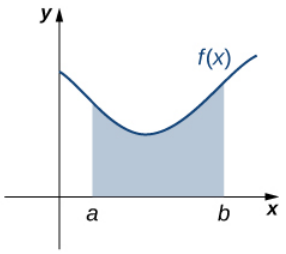
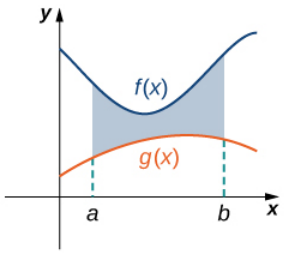
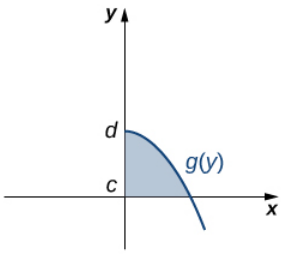
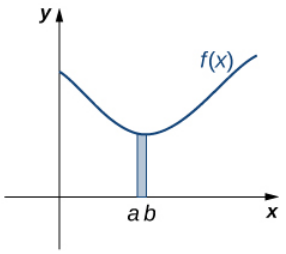
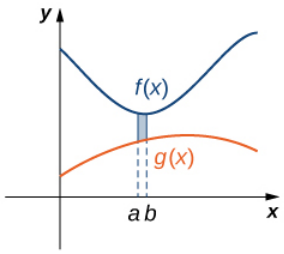
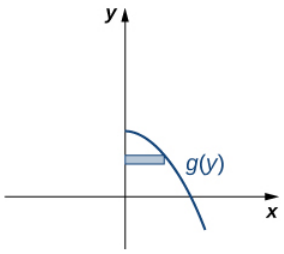
Compare	Disk Method	Washer Method	Shell Method
Volume formula	$V = \int_a^b \pi[f(x)]^2\,dx$	$V = \int_a^b \pi\left[(f(x))^2 - (g(x))^2\right]dx$	$V = \int_c^d 2\pi y\, g(y)\,dy$
Solid	No cavity in the center	Cavity in the center	With or without a cavity in the center
Interval to partition	$[a, b]$ on x-axis	$[a, b]$ on x-axis	$[c, d]$ on y-axis
Rectangle	Vertical	Vertical	Horizontal
Typical region			
Typical element			

Figure 6.2.10

Let's take a look at a couple of additional problems and decide on the best approach to take for solving them.

Example 6.2.6: Selecting the Best Method

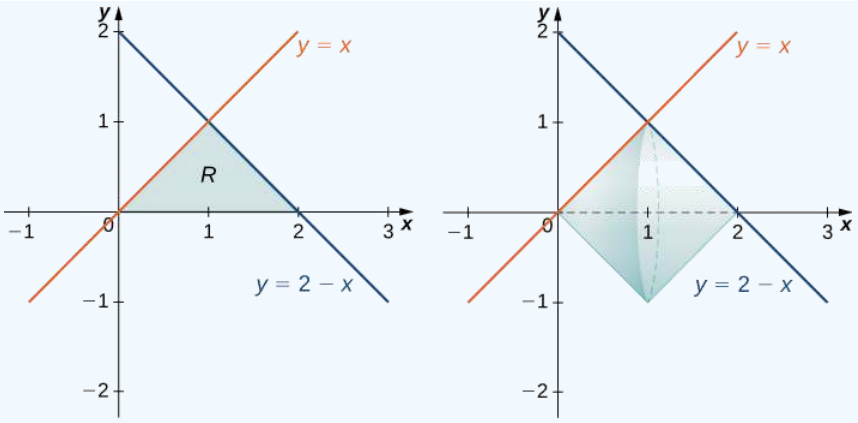
For each of the following problems, select the best method to find the volume of a solid of revolution generated by revolving the given region around the x-axis, and set up the integral to find the volume (do not evaluate the integral).

a. The region bounded by the graphs of $y = x, y = 2 - x,$ and the x-axis.
b. The region bounded by the graphs of $y = 4x - x^2$ and the x-axis.

Solution

a.

First, sketch the region and the solid of revolution as shown.



(a) (b)

Figure 6.2.11: (a) The region R bounded by two lines and the x-axis. (b) The solid of revolution generated by revolving R about the x-axis.

Looking at the region, if we want to integrate with respect to x, we would have to break the integral into two pieces, because we have different functions bounding the region over $[0, 1]$ and $[1, 2]$. In this case, using the disk method, we would have

$$V = \int_0^1 \pi x^2 \, dx + \int_1^2 \pi (2 - x)^2 \, dx.$$

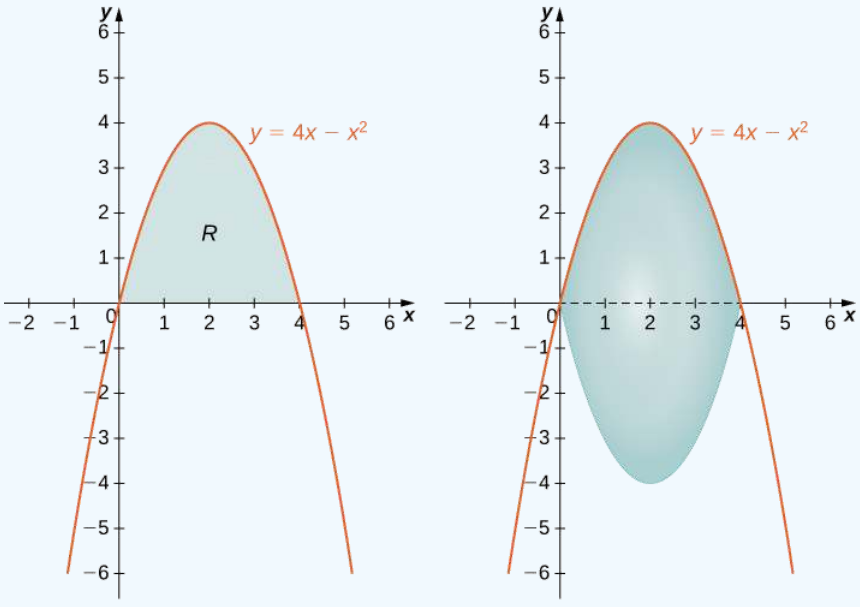
If we used the shell method instead, we would use functions of y to represent the curves, producing

$$V = \int_0^1 2\pi y[(2 - y) - y] \, dy = \int_0^1 2\pi y[2 - 2y] \, dy.$$

Neither of these integrals is particularly onerous, but since the shell method requires only one integral, and the integrand requires less simplification, we should probably go with the shell method in this case.

b.

First, sketch the region and the solid of revolution as shown.



(a) (b)

Figure 6.2.12: (a) The region R between the curve and the x-axis. (b) The solid of revolution generated by revolving R about the x-axis.

Looking at the region, it would be problematic to define a horizontal rectangle; the region is bounded on the left and right by the same function. Therefore, we can dismiss the method of shells. The solid has no cavity in the middle, so we can use the method of disks. Then

$$V = \int_0^4 \pi \left(4x - x^2\right)^2 \, dx$$

? Exercise 6.2.6

Select the best method to find the volume of a solid of revolution generated by revolving the given region around the x-axis, and set up the integral to find the volume (do not evaluate the integral): the region bounded by the graphs of $y = 2 - x^2$ and $y = x^2$.

Hint

Sketch the region and use Figure $6.2.12$to decide which integral is easiest to evaluate.

Answer

Use the method of washers;

$$V = \int_{-1}^1 \pi \left[\left(2 - x^2\right)^2 - \left(x^2\right)^2 \right] \, dx$$

Key Concepts

- The method of cylindrical shells is another method for using a definite integral to calculate the volume of a solid of revolution. This method is sometimes preferable to either the method of disks or the method of washers because we integrate with respect to the other variable. In some cases, one integral is substantially more complicated than the other.
- The geometry of the functions and the difficulty of the integration are the main factors in deciding which integration method to use.

Key Equations

- **Method of Cylindrical Shells**

$$V = \int_a^b \left(2\pi \, x \, f(x)\right) \, dx$$

Glossary

method of cylindrical shells

a method of calculating the volume of a solid of revolution by dividing the solid into nested cylindrical shells; this method is different from the methods of disks or washers in that we integrate with respect to the opposite variable

6.2: Volumes Using Cylindrical Shells is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **6.3: Volumes of Revolution - Cylindrical Shells** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

6.3: Arc Length

Learning Objectives

- Determine the length of a curve, $y = f(x)$, between two points.
- Determine the length of a curve, $x = g(y)$, between two points.
- Find the surface area of a solid of revolution.

In this section, we use definite integrals to find the arc length of a curve. We can think of **arc length** as the distance you would travel if you were walking along the path of the curve. Many real-world applications involve arc length. If a rocket is launched along a parabolic path, we might want to know how far the rocket travels. Or, if a curve on a map represents a road, we might want to know how far we have to drive to reach our destination.

We begin by calculating the arc length of curves defined as functions of x, then we examine the same process for curves defined as functions of y. (The process is identical, with the roles of x and y reversed.) The techniques we use to find arc length can be extended to find the surface area of a surface of revolution, and we close the section with an examination of this concept.

Arc Length of the Curve y = f(x)

In previous applications of integration, we required the function $f(x)$ to be integrable, or at most continuous. However, for calculating arc length we have a more stringent requirement for $f(x)$. Here, we require $f(x)$ to be differentiable, and furthermore we require its derivative, $f'(x)$, to be continuous. Functions like this, which have continuous derivatives, are called **smooth.** (This property comes up again in later chapters.)

Let $f(x)$ be a smooth function defined over $[a, b]$. We want to calculate the length of the curve from the point $(a, f(a))$ to the point $(b, f(b))$. We start by using line segments to approximate the length of the curve. For $i = 0, 1, 2, \ldots, n$, let $P = x_i$ be a regular partition of $[a, b]$. Then, for $i = 1, 2, \ldots, n$, construct a line segment from the point $(x_{i-1}, f(x_{i-1}))$ to the point $(x_i, f(x_i))$. Although it might seem logical to use either horizontal or vertical line segments, we want our line segments to approximate the curve as closely as possible. Figure 6.3.1 depicts this construct for $n = 5$.

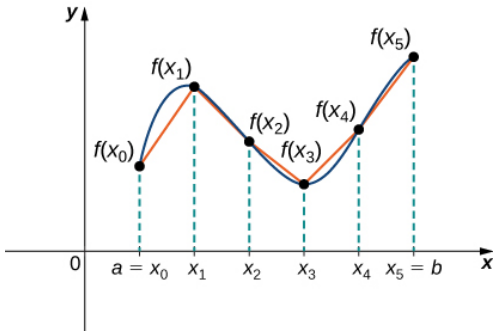


Figure 6.3.1: We can approximate the length of a curve by adding line segments.

To help us find the length of each line segment, we look at the change in vertical distance as well as the change in horizontal distance over each interval. Because we have used a regular partition, the change in horizontal distance over each interval is given by Δx. The change in vertical distance varies from interval to interval, though, so we use $\Delta y_i = f(x_i) - f(x_{i-1})$ to represent the change in vertical distance over the interval $[x_{i-1}, x_i]$, as shown in Figure 6.3.2. Note that some (or all) Δy_i may be negative.

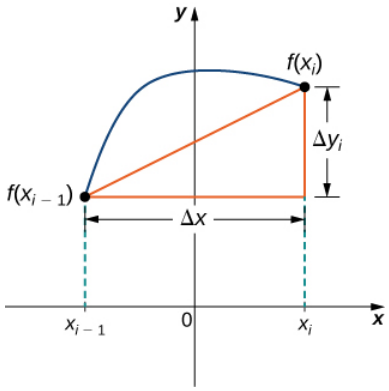


Figure 6.3.2: A representative line segment approximates the curve over the interval $[x_{i-1}, x_i]$.

By the Pythagorean theorem, the length of the line segment is

$$\sqrt{(\Delta x)^2 + (\Delta y_i)^2}.$$

We can also write this as

$$\Delta x \sqrt{1 + ((\Delta y_i)/(\Delta x))^2}.$$

Now, by the Mean Value Theorem, there is a point $x_i^* \in [x_{i-1}, x_i]$ such that $f'(x_i^*) = (\Delta y_i)/(\Delta x)$. Then the length of the line segment is given by

$$\Delta x \sqrt{1 + [f'(x_i^*)]^2}.$$

Adding up the lengths of all the line segments, we get

$$\text{Arc Length} \approx \sum_{i=1}^{n} \sqrt{1 + [f'(x_i^*)]^2}\, \Delta x.$$

This is a Riemann sum. Taking the limit as $n \to \infty$, we have

$$\begin{aligned} \text{Arc Length} &= \lim_{n\to\infty} \sum_{i=1}^{n} \sqrt{1 + [f'(x_i^*)]^2}\, \Delta x \\ &= \int_a^b \sqrt{1 + [f'(x)]^2}\, dx. \end{aligned}$$

We summarize these findings in the following theorem.

Arc Length for $y = f(x)$

Let $f(x)$ be a smooth function over the interval $[a, b]$. Then the arc length of the portion of the graph of $f(x)$ from the point $(a, f(a))$ to the point $(b, f(b))$ is given by

$$\text{Arc Length} = \int_a^b \sqrt{1 + [f'(x)]^2}\, dx.$$

Note that we are integrating an expression involving $f'(x)$, so we need to be sure $f'(x)$ is integrable. This is why we require $f(x)$ to be smooth. The following example shows how to apply the theorem.

Example 6.3.1: Calculating the Arc Length of a Function of x

Let $f(x) = 2x^{3/2}$. Calculate the arc length of the graph of $f(x)$ over the interval $[0, 1]$. Round the answer to three decimal places.

Solution

We have $f'(x) = 3x^{1/2}$, so $[f'(x)]^2 = 9x$. Then, the arc length is

$$\text{Arc Length} = \int_a^b \sqrt{1+[f'(x)]^2}dx$$
$$= \int_0^1 \sqrt{1+9x}\, dx.$$

Substitute $u = 1+9x$. Then, $du = 9\, dx$. When $x = 0$, then $u = 1$, and when $x = 1$, then $u = 10$. Thus,

$$\text{Arc Length} = \int_0^1 \sqrt{1+9x}\, dx$$
$$= \frac{1}{9}\int_0^1 \sqrt{1+9x}\cdot 9\, dx$$
$$= \frac{1}{9}\int_1^{10} \sqrt{u}\, du$$
$$= \frac{1}{9}\cdot\frac{2}{3}u^{3/2}\Big|_1^{10} = \frac{2}{27}[10\sqrt{10}-1]$$
$$\approx 2.268 \text{ units}.$$

Exercise 6.3.1

Let $f(x) = (4/3)x^{3/2}$. Calculate the arc length of the graph of $f(x)$ over the interval $[0, 1]$. Round the answer to three decimal places.

Hint

Use the process from the previous example. Don't forget to change the limits of integration.

Answer

$$\frac{1}{6}(5\sqrt{5}-1) \approx 1.697 \text{ units}$$

Although it is nice to have a formula for calculating arc length, this particular theorem can generate expressions that are difficult to integrate. We study some techniques for integration in Introduction to Techniques of Integration. In some cases, we may have to use a computer or calculator to approximate the value of the integral.

Example 6.3.2: Using a Computer or Calculator to Determine the Arc Length of a Function of x

Let $f(x) = x^2$. Calculate the arc length of the graph of $f(x)$ over the interval $[1, 3]$.

Solution

We have $f'(x) = 2x$, so $[f'(x)]^2 = 4x^2$. Then the arc length is given by

$$\text{Arc Length} = \int_a^b \sqrt{1+[f'(x)]^2}\, dx$$
$$= \int_1^3 \sqrt{1+4x^2}\, dx.$$

Using a computer to approximate the value of this integral, we get

$$\int_1^3 \sqrt{1+4x^2}\, dx \approx 8.26815 \text{ units}.$$

Exercise 6.3.2

Let $f(x) = \sin x$. Calculate the arc length of the graph of $f(x)$ over the interval $[0, \pi]$. Use a computer or calculator to approximate the value of the integral.

Hint

Use the process from the previous example.

Answer

$$\text{Arc Length} \approx 3.8202 \text{ units}$$

Arc Length of the Curve x = g(y)

We have just seen how to approximate the length of a curve with line segments. If we want to find the arc length of the graph of a function of y, we can repeat the same process, except we partition the y-axis instead of the x-axis. Figure 6.3.3 shows a representative line segment.

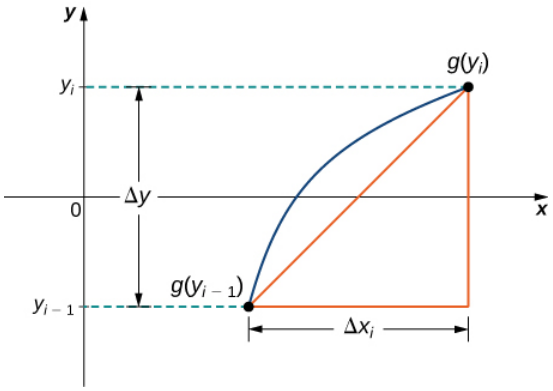


Figure 6.3.3: A representative line segment over the interval $[y_{i-1}, y_i]$.

Then the length of the line segment is

$$\sqrt{(\Delta y)^2 + (\Delta x_i)^2},$$

which can also be written as

$$\Delta y \sqrt{1 + \left(\frac{\Delta x_i}{\Delta y}\right)^2}.$$

If we now follow the same development we did earlier, we get a formula for arc length of a function $x = g(y)$.

Arc Length for $x = g(y)$

Let $g(y)$ be a smooth function over an interval $[c, d]$. Then, the arc length of the graph of $g(y)$ from the point $(c, g(c))$ to the point $(d, g(d))$ is given by

$$\text{Arc Length} = \int_c^d \sqrt{1 + [g'(y)]^2}\, dy.$$

Example 6.3.3: Calculating the Arc Length of a Function of y

Let $g(y) = 3y^3$. Calculate the arc length of the graph of $g(y)$ over the interval $[1, 2]$.

Solution

We have $g'(y) = 9y^2$, so $[g'(y)]^2 = 81y^4$. Then the arc length is

$$\text{Arc Length} = \int_c^d \sqrt{1+[g'(y)]^2}\, dy$$
$$= \int_1^2 \sqrt{1+81y^4}\, dy.$$

Using a computer to approximate the value of this integral, we obtain

$$\int_1^2 \sqrt{1+81y^4} dy \approx 21.0277 \text{ units}.$$

? Exercise 6.3.3

Let $g(y) = 1/y$. Calculate the arc length of the graph of $g(y)$ over the interval $[1, 4]$. Use a computer or calculator to approximate the value of the integral.

Hint

Use the process from the previous example.

Answer

$$\text{Arc Length} = 3.15018 \text{ units}$$

Area of a Surface of Revolution

The concepts we used to find the arc length of a curve can be extended to find the surface area of a surface of revolution. **Surface area** is the total area of the outer layer of an object. For objects such as cubes or bricks, the surface area of the object is the sum of the areas of all of its faces. For curved surfaces, the situation is a little more complex. Let $f(x)$ be a nonnegative smooth function over the interval $[a, b]$. We wish to find the surface area of the surface of revolution created by revolving the graph of $y = f(x)$ around the x-axis as shown in the following figure.

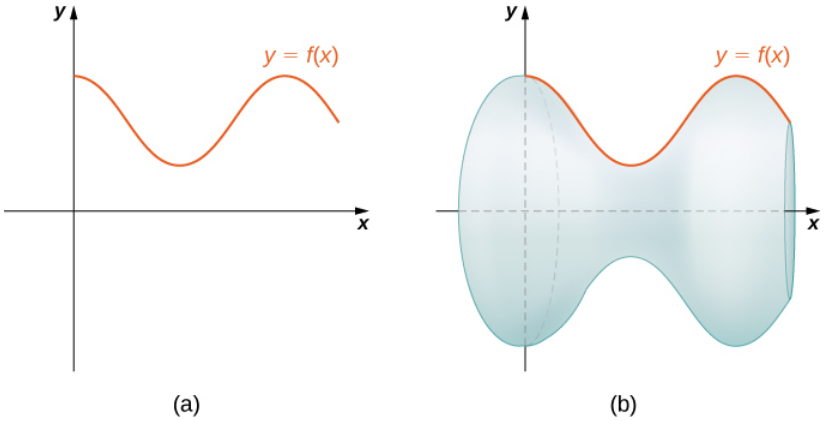


Figure 6.3.4: (a) A curve representing the function $f(x)$. (b) The surface of revolution formed by revolving the graph of $f(x)$ around the $x-axis$.

As we have done many times before, we are going to partition the interval $[a, b]$ and approximate the surface area by calculating the surface area of simpler shapes. We start by using line segments to approximate the curve, as we did earlier in this section. For $i = 0, 1, 2, \ldots, n$, let $P = x_i$ be a regular partition of $[a, b]$. Then, for $i = 1, 2, \ldots, n$, construct a line segment from the point $(x_{i-1}, f(x_{i-1}))$ to the point $(x_i, f(x_i))$. Now, revolve these line segments around the x-axis to generate an approximation of the surface of revolution as shown in the following figure.

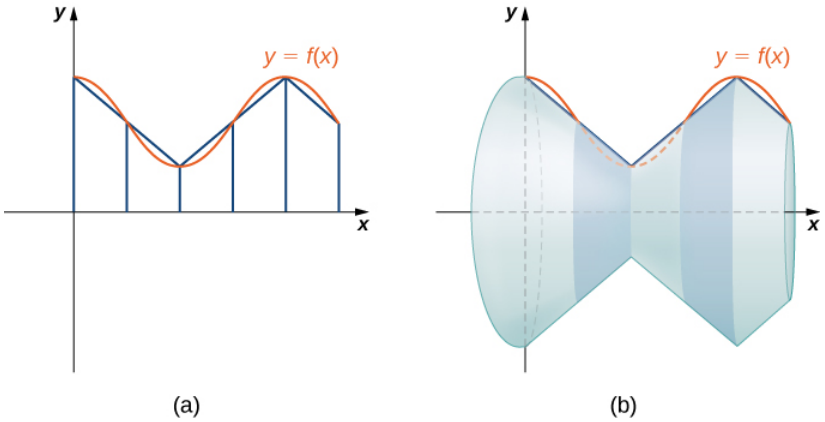


Figure 6.3.5: (a) Approximating $f(x)$ with line segments. (b) The surface of revolution formed by revolving the line segments around the x-axis.

Notice that when each line segment is revolved around the axis, it produces a band. These bands are actually pieces of cones (think of an ice cream cone with the pointy end cut off). A piece of a cone like this is called a **frustum** of a cone.

To find the surface area of the band, we need to find the lateral surface area, S, of the frustum (the area of just the slanted outside surface of the frustum, not including the areas of the top or bottom faces). Let r_1 and r_2 be the radii of the wide end and the narrow end of the frustum, respectively, and let l be the slant height of the frustum as shown in the following figure.

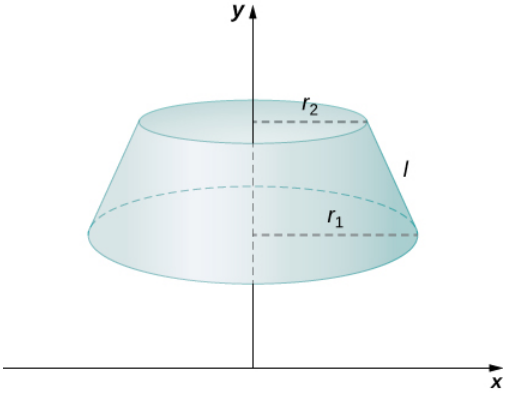


Figure 6.3.6: A frustum of a cone can approximate a small part of surface area.

We know the lateral surface area of a cone is given by

$$\text{Lateral Surface Area } = \pi r s,$$

where r is the radius of the base of the cone and s is the **slant height** (Figure 6.3.7).

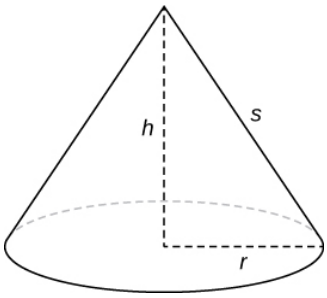


Figure 6.3.7: The lateral surface area of the cone is given by $\pi r s$.

Since a frustum can be thought of as a piece of a cone, the lateral surface area of the frustum is given by the lateral surface area of the whole cone less the lateral surface area of the smaller cone (the pointy tip) that was cut off (Figure 6.3.8).

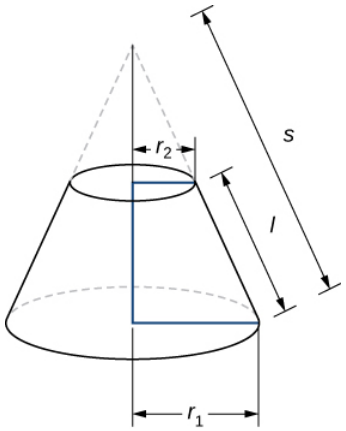


Figure 6.3.8: Calculating the lateral surface area of a frustum of a cone.

The cross-sections of the small cone and the large cone are similar triangles, so we see that

$$\frac{r_2}{r_1} = \frac{s-l}{s}$$

Solving for s, we get

$$\begin{aligned} \frac{r_2}{r_1} &= \frac{s-l}{s} \\ r_2 s &= r_1(s-l) \\ r_2 s &= r_1 s - r_1 l \\ r_1 l &= r_1 s - r_2 s \\ r_1 l &= (r_1 - r_2)s \\ \frac{r_1 l}{r_1 - r_2} &= s \end{aligned}$$

Then the lateral surface area (SA) of the frustum is

$$\begin{aligned} S &= (\text{Lateral SA of large cone}) - (\text{Lateral SA of small cone}) \\ &= \pi r_1 s - \pi r_2 (s-l) \\ &= \pi r_1 \left(\frac{r_1 l}{r_1 - r_2}\right) - \pi r_2 \left(\frac{r_1 l}{r_1 - r_2} - l\right) \\ &= \frac{\pi r_1^2 l}{r^1 - r^2} - \frac{\pi r_1 r_2 l}{r_1 - r_2} + \pi r_2 l \\ &= \frac{\pi r_1^2 l}{r_1 - r_2} - \frac{\pi r_1 r 2_l}{r_1 - r_2} + \frac{\pi r_2 l (r_1 - r_2)}{r_1 - r_2} \\ &= \frac{\pi r_1^2}{l r_1 - r_2} - \frac{\pi r_1 r_2 l}{r_1 - r_2} + \frac{\pi r_1 r_2 l}{r_1 - r_2} - \frac{\pi r_2^2 l}{r_1 - r_3} \\ &= \frac{\pi(r_1^2 - r_2^2)l}{r_1 - r_2} = \frac{\pi(r_1 - r + 2)(r1 + r2)l}{r_1 - r_2} \\ &= \pi(r_1 + r_2)l. \end{aligned}$$

Let's now use this formula to calculate the surface area of each of the bands formed by revolving the line segments around the x-axis. A representative band is shown in the following figure.

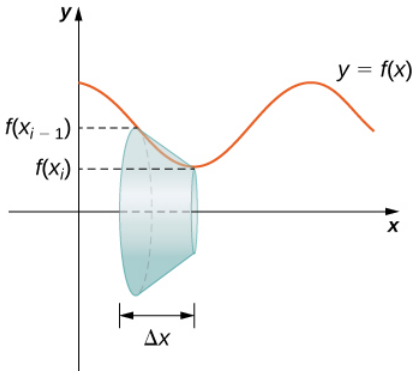


Figure 6.3.9: A representative band used for determining surface area.

Note that the slant height of this frustum is just the length of the line segment used to generate it. So, applying the surface area formula, we have

$$\begin{aligned} S &= \pi(r_1 + r_2)l \\ &= \pi\big(f(x_{i-1}) + f(x_i)\big)\sqrt{\Delta x^2 + (\Delta y_i)^2} \\ &= \pi\big(f(x_{i-1}) + f(x_i)\big)\Delta x\sqrt{1 + \left(\frac{\Delta y_i}{\Delta x}\right)^2} \end{aligned}$$

Now, as we did in the development of the arc length formula, we apply the Mean Value Theorem to select $x_i^* \in [x_{i-1}, x_i]$ such that $f\,'(x_i^*) = (\Delta y_i)/\Delta x$. This gives us

$$S = \pi\big(f(x_{i-1}) + f(x_i)\big)\Delta x\sqrt{1 + \big(f\,'(x_i^*)\big)^2}$$

Furthermore, since $f(x)$ is continuous, by the *Intermediate Value Theorem*, there is a point $x_i^{**} \in [x_{i-1}, x_i]$ such that $f(x_i^{**}) = \frac{1}{2}\left[f(x_{i-1}) + f(x_i)\right]$,

so we get

$$S = 2\pi f(x_i^{**})\Delta x\sqrt{1 + \big(f\,'(x_i^*)\big)^2}.$$

Then the approximate surface area of the whole surface of revolution is given by

$$\text{Surface Area} \approx \sum_{i=1}^{n} 2\pi f(x_i^{**})\Delta x\sqrt{1 + \big(f\,'(x_i^*)\big)^2}.$$

This almost looks like a Riemann sum, except we have functions evaluated at two different points, x_i^* and x_i^{**}, over the interval $[x_{i-1}, x_i]$. Although we do not examine the details here, it turns out that because $f(x)$ is smooth, if we let $n \to \infty$, the limit works the same as a Riemann sum even with the two different evaluation points. This makes sense intuitively. Both x_i^* and x_i^{**} are in the interval $[x_{i-1}, x_i]$, so it makes sense that as $n \to \infty$, both x_i^* and x_i^{**} approach x Those of you who are interested in the details should consult an advanced calculus text.

Taking the limit as $n \to \infty$, we get

$$\begin{aligned} \text{Surface Area} &= \lim_{n\to\infty} \sum_{i=1}^{n} 2\pi f(x_i^{**})\Delta x\sqrt{1 + \big(f\,'(x_i^*)\big)^2} \\ &= \int_a^b \left(2\pi f(x)\sqrt{1 + \big(f\,'(x)\big)^2}\right) dx \end{aligned}$$

As with arc length, we can conduct a similar development for functions of y to get a formula for the surface area of surfaces of revolution about the y-axis. These findings are summarized in the following theorem.

Surface Area of a Surface of Revolution

Let $f(x)$ be a nonnegative smooth function over the interval $[a, b]$. Then, the surface area of the surface of revolution formed by revolving the graph of $f(x)$ around the x-axis is given by

$$\text{Surface Area} = \int_a^b \left(2\pi f(x)\sqrt{1+\left(f'(x)\right)^2}\right) dx$$

Similarly, let $g(y)$ be a nonnegative smooth function over the interval $[c, d]$. Then, the surface area of the surface of revolution formed by revolving the graph of $g(y)$ around the y-axis is given by

$$\text{Surface Area} = \int_c^d \left(2\pi g(y)\sqrt{1+\left(g'(y)\right)^2}\right) dy$$

Example 6.3.4: Calculating the Surface Area of a Surface of Revolution 1.

Let $f(x) = \sqrt{x}$ over the interval $[1, 4]$. Find the surface area of the surface generated by revolving the graph of $f(x)$ around the x-axis. Round the answer to three decimal places.

Solution

The graph of $f(x)$ and the surface of rotation are shown in Figure 6.3.10.

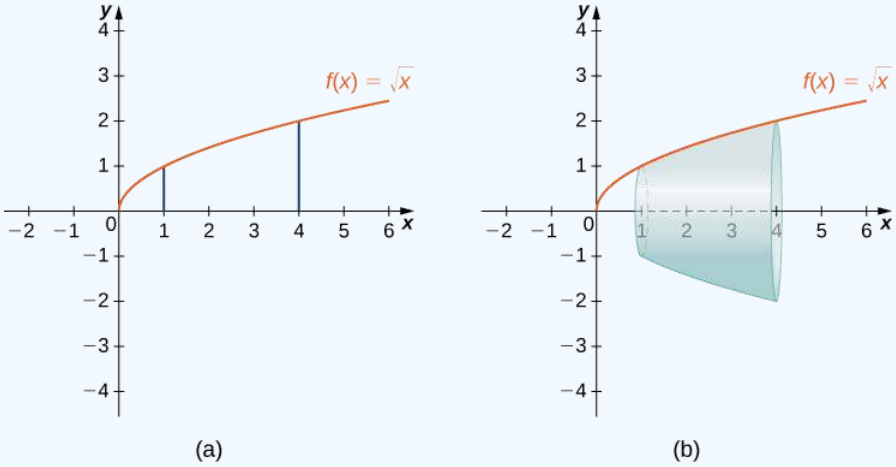


Figure 6.3.10: (a) The graph of $f(x)$. (b) The surface of revolution.

We have $f(x) = \sqrt{x}$. Then, $f'(x) = \frac{1}{2\sqrt{x}}$ and $\left(f'(x)\right)^2 = \frac{1}{4x}$. Then,

$$\begin{aligned}\text{Surface Area} &= \int_a^b \left(2\pi f(x)\sqrt{1+\left(f'(x)\right)^2}\right) dx \\ &= \int_1^4 \left(2\pi\sqrt{x}\sqrt{1+\frac{1}{4x}}\right) dx \\ &= \int_1^4 \left(2\pi\sqrt{x+\frac{1}{4}}\right) dx.\end{aligned}$$

Let $u = x + 1/4$. Then, $du = dx$. When $x = 1,\ u = 5/4$, and when $x = 4,\ u = 17/4$. This gives us

$$\begin{aligned}\int_0^1 \left(2\pi\sqrt{x+\frac{1}{4}}\right) dx &= \int_{5/4}^{17/4} 2\pi\sqrt{u}\, du \\ &= 2\pi\left[\frac{2}{3}u^{3/2}\right]\Big|_{5/4}^{17/4} \\ &= \frac{\pi}{6}[17\sqrt{17} - 5\sqrt{5}] \approx 30.846 \text{ units}^2\end{aligned}$$

Exercise 6.3.4

Let $f(x) = \sqrt{1-x}$ over the interval $[0, 1/2]$. Find the surface area of the surface generated by revolving the graph of $f(x)$ around the x-axis. Round the answer to three decimal places.

Hint

Use the process from the previous example.

Answer

$$\frac{\pi}{6}(5\sqrt{5} - 3\sqrt{3}) \approx 3.133 \text{ units}^2$$

Example 6.3.5: Calculating the Surface Area of a Surface of Revolution 2

Let $f(x) = y = \sqrt[3]{3x}$. Consider the portion of the curve where $0 \le y \le 2$. Find the surface area of the surface generated by revolving the graph of $f(x)$ around the y-axis.

Solution

Notice that we are revolving the curve around the y-axis, and the interval is in terms of y, so we want to rewrite the function as a function of y. We get $x = g(y) = (1/3)y^3$. The graph of $g(y)$ and the surface of rotation are shown in the following figure.

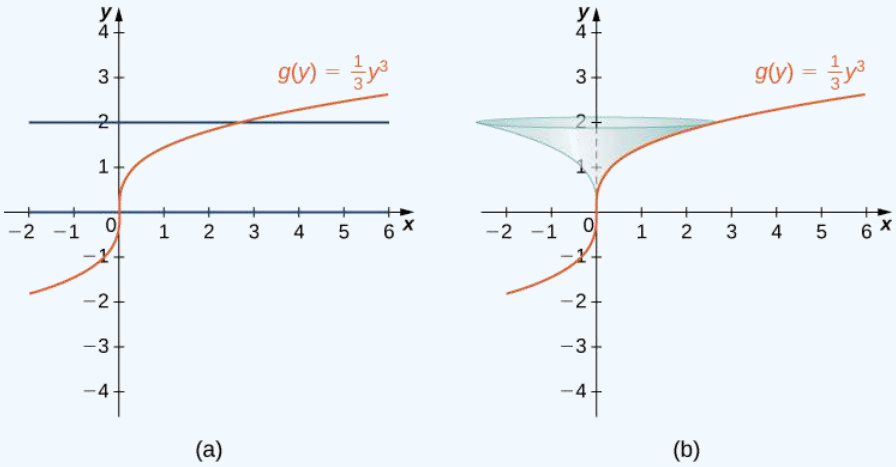


Figure 6.3.11: (a) The graph of $g(y)$. (b) The surface of revolution.

We have $g(y) = (1/3)y^3$, so $g'(y) = y^2$ and $\left(g'(y)\right)^2 = y^4$. Then

$$\begin{aligned} \text{Surface Area} &= \int_c^d \left(2\pi g(y)\sqrt{1+\left(g'(y)\right)^2}\right) dy \\ &= \int_0^2 \left(2\pi\left(\frac{1}{3}y^3\right)\sqrt{1+y^4}\right) dy \\ &= \frac{2\pi}{3}\int_0^2 \left(y^3\sqrt{1+y^4}\right) dy. \end{aligned}$$

Let $u = y^4 + 1$. Then $du = 4y^3\, dy$. When $y = 0$, $u = 1$, and when $y = 2$, $u = 17$. Then

$$\begin{aligned} \frac{2\pi}{3}\int_0^2 \left(y^3\sqrt{1+y^4}\right) dy &= \frac{2\pi}{3}\int_1^{17} \frac{1}{4}\sqrt{u}\, du \\ &= \frac{\pi}{6}\left[\frac{2}{3}u^{3/2}\right]\Big|_1^{17} = \frac{\pi}{9}\left[(17)^{3/2} - 1\right] \approx 24.118 \text{ units}^2. \end{aligned}$$

Exercise 6.3.5

Let $g(y) = \sqrt{9 - y^2}$ over the interval $y \in [0, 2]$. Find the surface area of the surface generated by revolving the graph of $g(y)$ around the y-axis.

Hint

Use the process from the previous example.

Answer

12π units^2

Key Concepts

- The arc length of a curve can be calculated using a definite integral.
- The arc length is first approximated using line segments, which generates a Riemann sum. Taking a limit then gives us the definite integral formula. The same process can be applied to functions of y.
- The concepts used to calculate the arc length can be generalized to find the surface area of a surface of revolution.
- The integrals generated by both the arc length and surface area formulas are often difficult to evaluate. It may be necessary to use a computer or calculator to approximate the values of the integrals.

Key Equations

- **Arc Length of a Function of x**

$$\textbf{Arc Length} = \int_a^b \sqrt{1 + [f'(x)]^2}\, dx$$

- **Arc Length of a Function of y**

$$\textbf{Arc Length} = \int_c^d \sqrt{1 + [g'(y)]^2}\, dy$$

- **Surface Area of a Function of x**

$$\textbf{Surface Area} = \int_a^b \left(2\pi f(x)\sqrt{1 + \left(f'(x)\right)^2}\right) dx$$

Glossary

arc length

the arc length of a curve can be thought of as the distance a person would travel along the path of the curve

frustum

a portion of a cone; a frustum is constructed by cutting the cone with a plane parallel to the base

surface area

the surface area of a solid is the total area of the outer layer of the object; for objects such as cubes or bricks, the surface area of the object is the sum of the areas of all of its faces

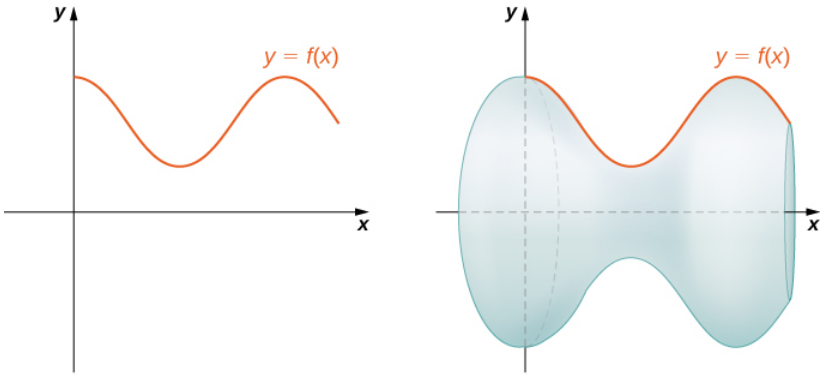
---

6.3: Arc Length is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **6.4: Arc Length of a Curve and Surface Area** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

6.4: Areas of Surfaces of Revolution

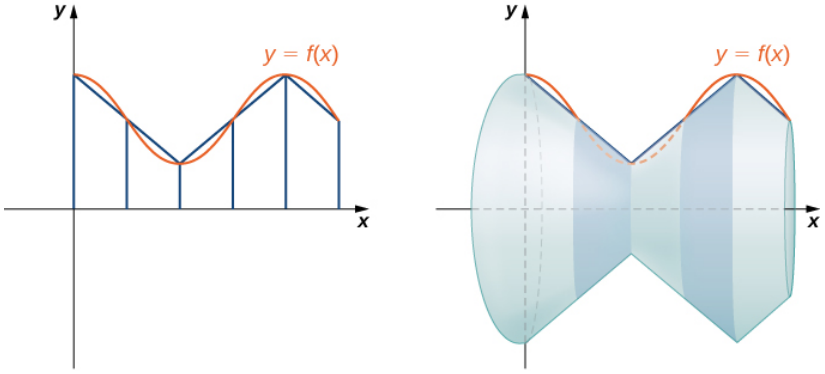
The concepts we used to find the arc length of a curve can be extended to find the surface area of a surface of revolution. **Surface area** is the total area of the outer layer of an object. For objects such as cubes or bricks, the surface area of the object is the sum of the areas of all of its faces. For curved surfaces, the situation is a little more complex. Let $f(x)$ be a nonnegative smooth function over the interval $[a, b]$. We wish to find the surface area of the surface of revolution created by revolving the graph of $y = f(x)$ around the x-axis as shown in the following figure.



(a) (b)

Figure 6.4.4: (a) A curve representing the function $f(x)$. (b) The surface of revolution formed by revolving the graph of $f(x)$ around the $x - axis$.

As we have done many times before, we are going to partition the interval $[a, b]$ and approximate the surface area by calculating the surface area of simpler shapes. We start by using line segments to approximate the curve, as we did earlier in this section. For $i = 0, 1, 2, \ldots, n$, let $P = x_i$ be a regular partition of $[a, b]$. Then, for $i = 1, 2, \ldots, n$, construct a line segment from the point $(x_{i-1}, f(x_{i-1}))$ to the point $(x_i, f(x_i))$. Now, revolve these line segments around the x-axis to generate an approximation of the surface of revolution as shown in the following figure.



(a) (b)

Figure 6.4.5: (a) Approximating $f(x)$ with line segments. (b) The surface of revolution formed by revolving the line segments around the x-axis.

Notice that when each line segment is revolved around the axis, it produces a band. These bands are actually pieces of cones (think of an ice cream cone with the pointy end cut off). A piece of a cone like this is called a **frustum** of a cone.

To find the surface area of the band, we need to find the lateral surface area, S, of the frustum (the area of just the slanted outside surface of the frustum, not including the areas of the top or bottom faces). Let r_1 and r_2 be the radii of the wide end and the narrow end of the frustum, respectively, and let l be the slant height of the frustum as shown in the following figure.

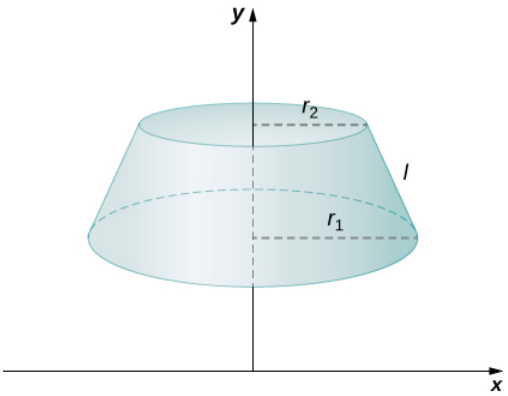


Figure 6.4.6: A frustum of a cone can approximate a small part of surface area.

We know the lateral surface area of a cone is given by

$$\text{Lateral Surface Area } = \pi r s,$$

where r is the radius of the base of the cone and s is the **slant height** (Figure 6.4.7).

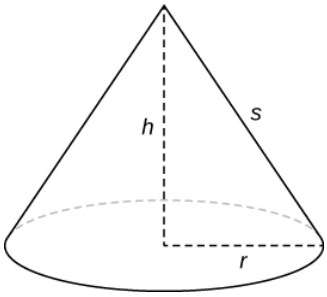


Figure 6.4.7: The lateral surface area of the cone is given by $\pi r s$.

Since a frustum can be thought of as a piece of a cone, the lateral surface area of the frustum is given by the lateral surface area of the whole cone less the lateral surface area of the smaller cone (the pointy tip) that was cut off (Figure 6.4.8).

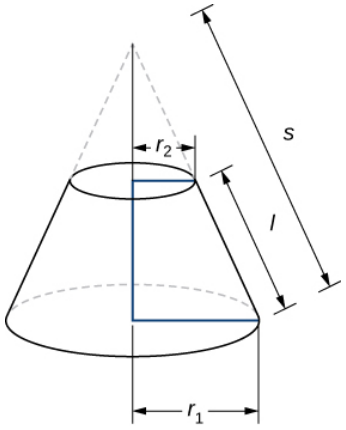


Figure 6.4.8: Calculating the lateral surface area of a frustum of a cone.

The cross-sections of the small cone and the large cone are similar triangles, so we see that

$$\frac{r_2}{r_1} = \frac{s-l}{s}$$

Solving for s, we get

$$\begin{aligned}\frac{r_2}{r_1} &= \frac{s-l}{s} \\ r_2 s &= r_1(s-l) \\ r_2 s &= r_1 s - r_1 l \\ r_1 l &= r_1 s - r_2 s \\ r_1 l &= (r_1 - r_2)s \\ \frac{r_1 l}{r_1 - r_2} &= s\end{aligned}$$

Then the lateral surface area (SA) of the frustum is

$$\begin{aligned}S &= (\text{Lateral SA of large cone}) - (\text{Lateral SA of small cone}) \\ &= \pi r_1 s - \pi r_2 (s-l) \\ &= \pi r_1 \left(\frac{r_1 l}{r_1 - r_2}\right) - \pi r_2 \left(\frac{r_1 l}{r_1 - r_2} - l\right) \\ &= \frac{\pi r_1^2 l}{r^1 - r^2} - \frac{\pi r_1 r_2 l}{r_1 - r_2} + \pi r_2 l \\ &= \frac{\pi r_1^2 l}{r_1 - r_2} - \frac{\pi r_1 r 2_l}{r_1 - r_2} + \frac{\pi r_2 l (r_1 - r_2)}{r_1 - r_2} \\ &= \frac{\pi r_1^2}{l r_1 - r_2} - \frac{\pi r_1 r_2 l}{r_1 - r_2} + \frac{\pi r_1 r_2 l}{r_1 - r_2} - \frac{\pi r_2^2 l}{r_1 - r_3} \\ &= \frac{\pi(r_1^2 - r_2^2)l}{r_1 - r_2} = \frac{\pi(r_1 - r + 2)(r1 + r2)l}{r_1 - r_2} \\ &= \pi(r_1 + r_2)l.\end{aligned}$$

Let's now use this formula to calculate the surface area of each of the bands formed by revolving the line segments around the x-axis. A representative band is shown in the following figure.

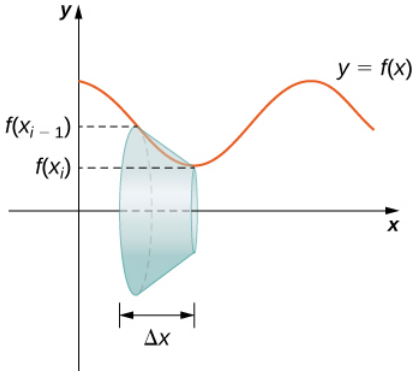


Figure 6.4.9: A representative band used for determining surface area.

Note that the slant height of this frustum is just the length of the line segment used to generate it. So, applying the surface area formula, we have

$$\begin{aligned}S &= \pi(r_1 + r_2)l \\ &= \pi\big(f(x_{i-1}) + f(x_i)\big)\sqrt{\Delta x^2 + (\Delta y_i)^2} \\ &= \pi\big(f(x_{i-1}) + f(x_i)\big)\Delta x\sqrt{1 + \left(\frac{\Delta y_i}{\Delta x}\right)^2}\end{aligned}$$

Now, as we did in the development of the arc length formula, we apply the Mean Value Theorem to select $x_i^* \in [x_{i-1}, x_i]$ such that $f'(x_i^*) = (\Delta y_i)/\Delta x$. This gives us

$$S = \pi\big(f(x_{i-1}) + f(x_i)\big)\Delta x\sqrt{1 + \big(f\,'(x_i^*)\big)^2}$$

Furthermore, since $f(x)$ is continuous, by the *Intermediate Value Theorem*, there is a point $x_i^{**} \in [x_{i-1}, x_i]$ such that $f(x_i^{**}) = \frac{1}{2}\left[f(x_{i-1}) + f(x_i)\right]$,

so we get

$$S = 2\pi f(x_i^{**})\Delta x\sqrt{1 + \big(f\,'(x_i^*)\big)^2}.$$

Then the approximate surface area of the whole surface of revolution is given by

$$\text{Surface Area} \approx \sum_{i=1}^{n} 2\pi f(x_i^{**})\Delta x\sqrt{1 + \big(f\,'(x_i^*)\big)^2}.$$

This almost looks like a Riemann sum, except we have functions evaluated at two different points, x_i^* and x_i^{**}, over the interval $[x_{i-1}, x_i]$. Although we do not examine the details here, it turns out that because $f(x)$ is smooth, if we let $n \to \infty$, the limit works the same as a Riemann sum even with the two different evaluation points. This makes sense intuitively. Both x_i^* and x_i^{**} are in the interval $[x_{i-1}, x_i]$, so it makes sense that as $n \to \infty$, both x_i^* and x_i^{**} approach x Those of you who are interested in the details should consult an advanced calculus text.

Taking the limit as $n \to \infty$, we get

$$\begin{aligned}\text{Surface Area} &= \lim_{n\to\infty}\sum_{i=1}^{n} 2\pi f(x_i^{**})\Delta x\sqrt{1 + \big(f\,'(x_i^*)\big)^2} \\ &= \int_a^b \left(2\pi f(x)\sqrt{1 + \big(f\,'(x)\big)^2}\right) dx\end{aligned}$$

As with arc length, we can conduct a similar development for functions of y to get a formula for the surface area of surfaces of revolution about the y-axis. These findings are summarized in the following theorem.

Surface Area of a Surface of Revolution

Let $f(x)$ be a nonnegative smooth function over the interval $[a, b]$. Then, the surface area of the surface of revolution formed by revolving the graph of $f(x)$ around the x-axis is given by

$$\text{Surface Area} = \int_a^b \left(2\pi f(x)\sqrt{1 + \big(f\,'(x)\big)^2}\right) dx$$

Similarly, let $g(y)$ be a nonnegative smooth function over the interval $[c, d]$. Then, the surface area of the surface of revolution formed by revolving the graph of $g(y)$ around the y-axis is given by

$$\text{Surface Area} = \int_c^d \left(2\pi g(y)\sqrt{1 + \big(g'(y)\big)^2}\right) dy$$

Example 6.4.4: Calculating the Surface Area of a Surface of Revolution 1.

Let $f(x) = \sqrt{x}$ over the interval $[1, 4]$. Find the surface area of the surface generated by revolving the graph of $f(x)$ around the x-axis. Round the answer to three decimal places.

Solution

The graph of $f(x)$ and the surface of rotation are shown in Figure 6.4.10

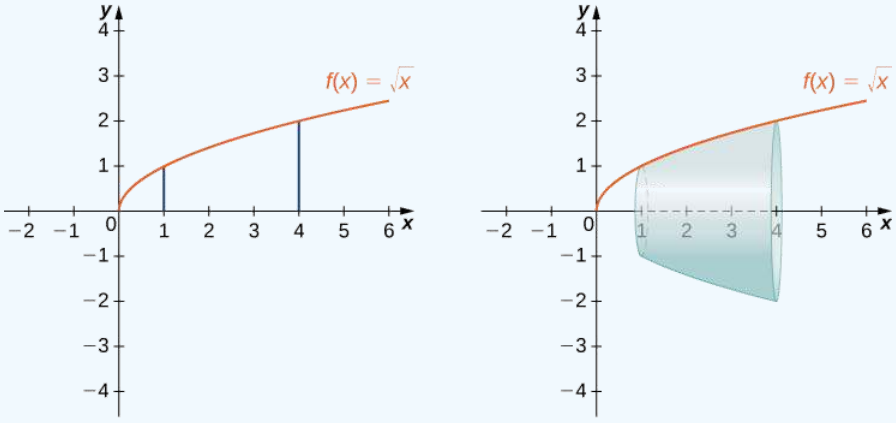


Figure 6.4.10: (a) The graph of $f(x)$. (b) The surface of revolution.

We have $f(x)=\sqrt{x}$. Then, $f'(x)=\frac{1}{2\sqrt{x}}$ and $\left(f'(x)\right)^2=\frac{1}{4x}$. Then,

$$\begin{aligned}\text{Surface Area} &= \int_a^b \left(2\pi f(x)\sqrt{1+\left(f'(x)\right)^2}\right)dx \\ &= \int_1^4 \left(2\pi\sqrt{x}\sqrt{1+\frac{1}{4x}}\right)dx \\ &= \int_1^4 \left(2\pi\sqrt{x+\frac{1}{4}}\right)dx.\end{aligned}$$

Let $u=x+1/4$. Then, $du=dx$. When $x=1$, $u=5/4$, and when $x=4$, $u=17/4$. This gives us

$$\begin{aligned}\int_0^1 \left(2\pi\sqrt{x+\frac{1}{4}}\right)dx &= \int_{5/4}^{17/4} 2\pi\sqrt{u}\,du \\ &= 2\pi\left[\frac{2}{3}u^{3/2}\right]\Big|_{5/4}^{17/4} \\ &= \frac{\pi}{6}[17\sqrt{17}-5\sqrt{5}] \approx 30.846 \text{ units}^2\end{aligned}$$

Exercise 6.4.4

Let $f(x)=\sqrt{1-x}$ over the interval $[0,1/2]$. Find the surface area of the surface generated by revolving the graph of $f(x)$ around the x-axis. Round the answer to three decimal places.

Hint

Use the process from the previous example.

Answer

$$\frac{\pi}{6}\left(5\sqrt{5}-3\sqrt{3}\right) \approx 3.133 \text{ units}^2$$

Example 6.4.5: Calculating the Surface Area of a Surface of Revolution 2

Let $f(x)=y=\sqrt[3]{3x}$. Consider the portion of the curve where $0\le y\le 2$. Find the surface area of the surface generated by revolving the graph of $f(x)$ around the y-axis.

Solution

Notice that we are revolving the curve around the y-axis, and the interval is in terms of y, so we want to rewrite the function as a function of y. We get $x = g(y) = (1/3)y^3$. The graph of $g(y)$ and the surface of rotation are shown in the following figure.

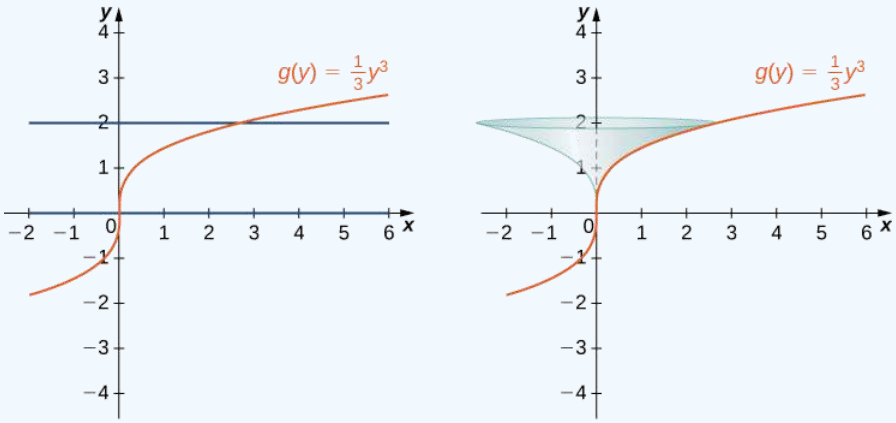


Figure 6.4.11: (a) The graph of $g(y)$. (b) The surface of revolution.

We have $g(y) = (1/3)y^3$, so $g'(y) = y^2$ and $\left(g'(y)\right)^2 = y^4$. Then

$$\begin{aligned} \text{Surface Area} &= \int_c^d \left(2\pi g(y)\sqrt{1+\left(g'(y)\right)^2}\right)\, dy \\ &= \int_0^2 \left(2\pi\left(\frac{1}{3}y^3\right)\sqrt{1+y^4}\right)\, dy \\ &= \frac{2\pi}{3}\int_0^2 \left(y^3\sqrt{1+y^4}\right)\, dy. \end{aligned}$$

Let $u = y^4 + 1$. Then $du = 4y^3\, dy$. When $y = 0$, $u = 1$, and when $y = 2$, $u = 17$. Then

$$\begin{aligned} \frac{2\pi}{3}\int_0^2 \left(y^3\sqrt{1+y^4}\right)\, dy &= \frac{2\pi}{3}\int_1^{17} \frac{1}{4}\sqrt{u}\, du \\ &= \frac{\pi}{6}\left[\frac{2}{3}u^{3/2}\right]\Big|_1^{17} = \frac{\pi}{9}\left[(17)^{3/2} - 1\right] \approx 24.118 \text{ units}^2. \end{aligned}$$

Exercise 6.4.5

Let $g(y) = \sqrt{9-y^2}$ over the interval $y \in [0, 2]$. Find the surface area of the surface generated by revolving the graph of $g(y)$ around the y-axis.

Hint

Use the process from the previous example.

Answer

$12\pi \text{ units}^2$

6.4: Areas of Surfaces of Revolution is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **6.4: Arc Length of a Curve and Surface Area** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

6.5: Work

Learning Objectives

In this section, we strive to understand the ideas generated by the following important questions:

- How do we measure the work accomplished by a varying force that moves an object a certain distance?
- What is the total force exerted by water against a dam?
- How are both of the above concepts and their corresponding use of definite integrals similar to problems we have encountered in the past involving formulas such as “distance equals rate times time” and “mass equals density times volume”?

In our work to date with the definite integral, we have seen several different circumstances where the integral enables us to measure the accumulation of a quantity that varies, provided the quantity is approximately constant over small intervals. For instance, based on the fact that the area of a rectangle is $A = l \cdot w$, if we wish to find the area bounded by a nonnegative curve $y = f(x)$ and the x-axis on an interval $[a, b]$, a representative slice of width Δx has area $A_{\text{slice}} = f(x)\Delta x$, and thus as we let the width of the representative slice tend to zero, we find that the exact area of the region is

$$A = \int_a^b f(x)dx. \tag{6.5.1}$$

In a similar way, if we know that the velocity of a moving object is given by the function $y = v(t)$, and we wish to know the distance the object travels on an interval $[a, b]$ where $v(t)$ is nonnegative, we can use a definite integral to generalize the fact that $d = r \cdot t$ when the rate, r, is constant. More specifically, on a short time interval Δt, $v(t)$ is roughly constant, and hence for a small slice of time, $d_{\text{slice}} = v(t)\Delta t$, and so as the width of the time interval Δt tends to zero, the exact distance traveled is given by the definite integral

$$d = \int_a^b v(t)dt. \tag{6.5.2}$$

Finally, when we recently learned about the mass of an object of non-constant density, we saw that since $M = D \cdot V$ (mass equals density times volume, provided that density is constant), if we can consider a small slice of an object on which the density is approximately constant, a definite integral may be used to determine the exact mass of the object. For instance, if we have a thin rod whose cross sections have constant density, but whose density is distributed along the x axis according to the function $y = \rho(x)$, it follows that for a small slice of the rod that is Δx thick, $M_{\text{slice}} = \rho(x)\Delta x$. In the limit as $\Delta x \to 0$, we then find that the total mass is given by

$$M = \int_a^b \rho(x)dx. \tag{6.5.3}$$

Note that all three of these situations are similar in that we have a basic rule ($A = l \cdot w, d = r \cdot t, M = D \cdot V$) where one of the two quantities being multiplied is no longer constant; in each, we consider a small interval for the other variable in the formula, calculate the approximate value of the desired quantity (area, distance, or mass) over the small interval, and then use a definite integral to sum the results as the length of the small intervals is allowed to approach zero. It should be apparent that this approach will work effectively for other situations where we have a quantity of interest that varies. We next turn to the notion of work: from physics, a basic principal is that work is the product of force and distance. For example, if a person exerts a force of 20 pounds to lift a 20-pound weight 4 feet off the ground, the total work accomplished is

$$W = F \cdot d \tag{6.5.4}$$

$$= 20 \cdot 4 \tag{6.5.5}$$

$$= 80 \text{ foot-pounds}. \tag{6.5.6}$$

If force and distance are measured in English units (pounds and feet), then the units on work are *foot-pounds*. If instead we work in metric units, where forces are measured in Newtons and distances in meters, the units on work are Newton-meters.

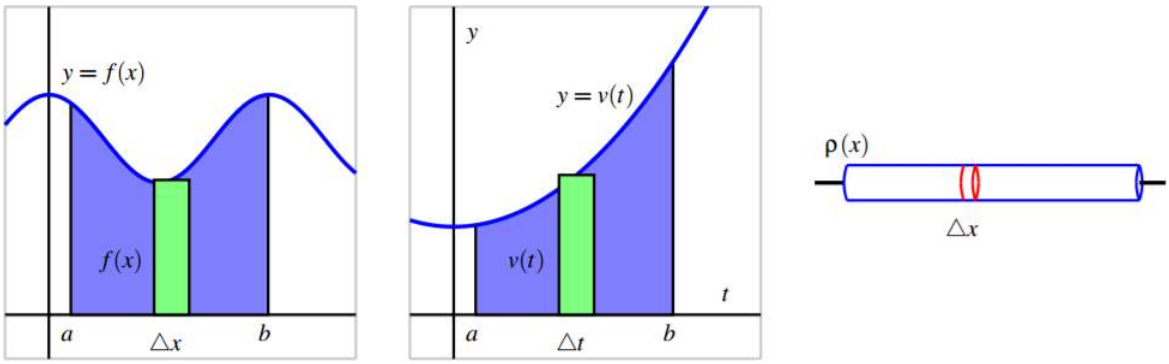


Figure 6.14: *Three settings where we compute the accumulation of a varying quantity: the area under $y = f(x)$, the distance traveled by an object with velocity $y = v(t)$, and the mass of a bar with density function $y = \rho(x)$.*

Of course, the formula $W = F \cdot d$ only applies when the force is constant while it is exerted over the distance d. In Preview Activity 6.4, we explore one way that we can use a definite integral to compute the total work accomplished when the force exerted varies.

Preview Activity 6.5.1

A bucket is being lifted from the bottom of a 50-foot deep well; its weight (including the water), B, in pounds at a height h feet above the water is given by the function $B(h)$. When the bucket leaves the water, the bucket and water together weigh $B(0) = 20$ pounds, and when the bucket reaches the top of the well, $B(50) = 12$ pounds. Assume that the bucket loses water at a constant rate (as a function of height, h) throughout its journey from the bottom to the top of the well.

a. Find a formula for $B(h)$.
b. Compute the value of the product $B(5)\Delta h$, where $\Delta h = 2$ feet. Include units on your answer. Explain why this product represents the approximate work it took to move the bucket of water from $h = 5$ to $h = 7$.
c. Is the value in (b) an over- or under-estimate of the actual amount of work it took to move the bucket from $h = 5$ to $h = 7$? Why?
d. Compute the value of the product $B(22)\Delta h$, where $\Delta h = 0.25$ feet. Include units on your answer. What is the meaning of the value you found?
e. More generally, what does the quantity $W_{\text{slice}} = B(h)\Delta h$ measure for a given value of h and a small positive value of Δh?
f. Evaluate the definite integral $\int^5 0_0 B(h)dh$. What is the meaning of the value you find? Why?

Work

Because work is calculated by the rule $W = F \cdot d$, whenever the force F is constant, it follows that we can use a definite integral to compute the work accomplished by a varying force. For example, suppose that in a setting similar to the problem posed in Preview Activity 6.4, we have a bucket being lifted in a 50-foot well whose weight at height h is given by

$$B(h) = 12 + 8e^{-0.1h}. \tag{6.5.7}$$

In contrast to the problem in the preview activity, this bucket is not leaking at a constant rate; but because the weight of the bucket and water is not constant, we have to use a definite integral to determine the total work that results from lifting the bucket. Observe that at a height h above the water, the approximate work to move the bucket a small distance Δh is

$$W_{\text{slice}} = B(h)\Delta h = (12 + 8e^{-0.1h})\Delta h. \tag{6.5.8}$$

Hence, if we let Δh tend to 0 and take the sum of all of the slices of work accomplished on these small intervals, it follows that the total work is given by

$$W = \int_0^{50} B(h)\, dh = \int_0^{50} (12 + 8e^{-0.1h})dh. \tag{6.5.9}$$

While is a straightforward exercise to evaluate this integral exactly using the First Fundamental Theorem of Calculus, in applied settings such as this one we will typically use computing technology to find accurate approximations of integrals that are of interest to us. Here, it turns out that

$$W = \int_0^{50} (12 + 8e^{-0.1h}) dh \approx 679.461 \text{ foot-pounds.} \tag{6.5.10}$$

Our work in Preview Activity 6.1 and in the most recent example above employs the following important general principle.

For an object being moved in the positive direction along an axis, x, by a force $F(x)$, the total work to move the object from a to b is given by

$$W = \int_a^b F(x) dx. \tag{6.5.11}$$

Activity 6.5.1

Consider the following situations in which a varying force accomplishes work.

a. Suppose that a heavy rope hangs over the side of a cliff. The rope is 200 feet long and weighs 0.3 pounds per foot; initially the rope is fully extended. How much work is required to haul in the entire length of the rope? (Hint: set up a function $F(h)$ whose value is the weight of the rope remaining over the cliff after h feet have been hauled in.)

b. A leaky bucket is being hauled up from a 100 foot deep well. When lifted from the water, the bucket and water together weigh 40 pounds. As the bucket is being hauled upward at a constant rate, the bucket leaks water at a constant rate so that it is losing weight at a rate of 0.1 pounds per foot. What function $B(h)$ tells the weight of the bucket after the bucket has been lifted h feet? What is the total amount of work accomplished in lifting the bucket to the top of the well?

c. Now suppose that the bucket in (b) does not leak at a constant rate, but rather that its weight at a height h feet above the water is given by $B(h) = 25 + 15e^{-0.05h}$. What is the total work required to lift the bucket 100 feet? What is the average force exerted on the bucket on the interval $h = 0$ to $h = 100$?

d. From physics, *Hooke's Law* for springs states that the amount of force required to hold a spring that is compressed (or extended) to a particular length is proportionate to the distance the spring is compressed (or extended) from its natural length. That is, the force to compress (or extend) a spring x units from its natural length is $F(x) = kx$ for some constant k (which is called the spring constant.) For springs, we choose to measure the force in pounds and the distance the spring is compressed in feet. Suppose that a force of 5 pounds extends a particular spring 4 inches (1/3 foot) beyond its natural length.

   i. Use the given fact that $F(1/3) = 5$ to find the spring constant k.
   ii. Find the work done to extend the spring from its natural length to 1 foot beyond its natural length.
   iii. Find the work required to extend the spring from 1 foot beyond its natural length to 1.5 feet beyond its natural length.

Work: Pumping Liquid from a Tank

In certain geographic locations where the water table is high, residential homes with basements have a peculiar feature: in the basement, one finds a large hole in the floor, and in the hole, there is water. For example, in Figure 6.15 where we see a *sump crock*.



Figure 6.15: A sump crock. Image credit to www.warreninspect.com/basement-moisture.

Essentially, a sump crock provides an outlet for water that may build up beneath the basement floor; of course, as that water rises, it is imperative that the water not flood the basement. Hence, in the crock we see the presence of a floating pump that sits on the surface of the water: this pump is activated by elevation, so when the water level reaches a particular height, the pump turns on and pumps a certain portion of the water out of the crock, hence relieving the water buildup beneath the foundation. One of the questions we'd like to answer is: how much work does a sump pump accomplish? To that end, let's suppose that we have a sump crock that has the shape of a frustum of a cone, as pictured in Figure 6.16. Assume that the crock has a diameter of 3 feet at its surface, a diameter of 1.5 feet at its base, and a depth of 4 feet. In addition, suppose that the sump pump is set up so that it pumps the water vertically up a pipe to a drain that is located at ground level just outside a basement window. To accomplish this, the pump must send the water to a location 9 feet above the surface of the sump crock.

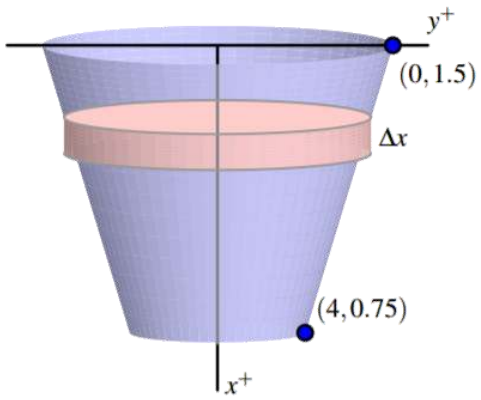


Figure 6.16: A sump crock with approximately cylindrical cross-sections that is 4 feet deep, 1.5 feet in diameter at its base, and 3 feet in diameter at its top.

It turns out to be advantageous to think of the depth below the surface of the crock as being the independent variable, so, in problems such as this one we typically let the positive x-axis point down, and the positive y-axis to the right, as pictured in the figure. As we think about the work that the pump does, we first realize that the pump sits on the surface of the water, so it makes sense to think about the pump moving the water one "slice" at a time, where it takes a thin slice from the surface, pumps it out of the tank, and then proceeds to pump the next slice below. For the sump crock described in this example, each slice of water is cylindrical in shape. We see that the radius of each approximately cylindrical slice varies according to the linear function $y = f(x)$ that passes through the points (0, 1.5) and (4, 0.75), where x is the depth of the particular slice in the tank; it is a straightforward exercise to find that $f(x) = 1.5 - 0.1875x$. Now we are prepared to think about the overall problem in several steps:

a. determining the volume of a typical slice;
b. finding the weight (We assume that the weight density of water is 62.4 pounds per cubic foot) of a typical slice (and thus the force that must be exerted on it)
c. deciding the distance that a typical slice moves; and
d. computing the work to move a representative slice. Once we know the work it takes to move one slice, we use a definite integral over an appropriate interval to find the total work.

Consider a representative cylindrical slice that sits on the surface of the water at a depth of x feet below the top of the crock. It follows that the approximate volume of that slice is given by

$$V_{\text{slice}} = \pi f(x)^2 \Delta x = \pi(1.5 - 0.1875x)^2 \Delta x$$

Since water weighs 62.4 lb/ft3 , it follows that the approximate weight of a representative slice, which is also the approximate force the pump must exert to move the slice, is

$$F_{\text{slice}} = 62.4 \cdot V_{\text{slice}} = 62.4\pi(1.5 - 0.1875x)^2 \Delta x.$$

Because the slice is located at a depth of x feet below the top of the crock, the slice being moved by the pump must move x feet to get to the level of the basement floor, and then, as stated in the problem description, be moved another 9 feet to reach the drain at ground level outside a basement window. Hence, the total distance a representative slice travels is

$$d_{\text{slice}} = x + 9\,.$$

Finally, we note that the work to move a representative slice is given by

$$W_{\text{slice}} = F_{\text{slice}} \cdot d_{\text{slice}} = 62.4\pi(1.5 - 0.1875x)^2 \Delta x \cdot (x + 9),$$

since the force to move a particular slice is constant. We sum the work required to move slices throughout the tank (from $x = 0$ to $x = 4$), let $\Delta x \to 0$, and hence

$$W = \int_0^4 62.4\pi(1.5 - 0.1875x)^2(x + 9)dx,$$

which, when evaluated using appropriate technology, shows that the total work is $W = 10970.5\pi$ foot-pounds.

The preceding example demonstrates the standard approach to finding the work required to empty a tank filled with liquid. The main task in each such problem is to determine the volume of a representative slice, followed by the force exerted on the slice, as well as the distance such a slice moves. In the case where the units are metric, there is one key difference: in the metric setting, rather than weight, we normally first find the mass of a slice. For instance, if distance is measured in meters, the mass density of water is 1000 kg/m^3 . In that setting, we can find the mass of a typical slice (in kg). To determine the force required to move it, we use F = ma, where m is the object's mass and a is the gravitational constant 9.81 N/kg^3 . That is, in metric units, the weight density of water is 9810 N/m^3 .

Activity 6.5.2

In each of the following problems, determine the total work required to accomplish the described task. In parts (b) and (c), a key step is to find a formula for a function that describes the curve that forms the side boundary of the tank.

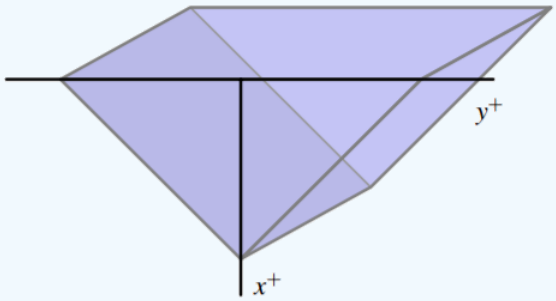


Figure 6.17: A trough with triangular ends, as described in Activity 6.11, part (c).

a. Consider a vertical cylindrical tank of radius 2 meters and depth 6 meters. Suppose the tank is filled with 4 meters of water of mass density 1000 kg/m^3 , and the top 1 meter of water is pumped over the top of the tank.
b. Consider a hemispherical tank with a radius of 10 feet. Suppose that the tank is full to a depth of 7 feet with water of weight density 62.4 pounds/ft^3, and the top 5 feet of water are pumped out of the tank to a tanker truck whose height is 5 feet above the top of the tank.
c. Consider a trough with triangular ends, as pictured in Figure 6.17, where the tank is 10 feet long, the top is 5 feet wide, and the tank is 4 feet deep. Say that the trough is full to within 1 foot of the top with water of weight density 62.4 pounds/ft^3, and a pump is used to empty the tank until the water remaining in the tank is 1 foot deep.

Force due to Hydrostatic Pressure

When a dam is built, it is imperative to for engineers to understand how much force water will exert against the face of the dam. The first thing we realize is the force exerted by the fluid is related to the natural concept of pressure. The pressure a force exerts on a region is measured in units of force per unit of area: for example, the air pressure in a tire is often measured in pounds per square inch (PSI). Hence, we see that the general relationship is given by

$$P = \frac{F}{A}, \text{ or } F = P \cdot A,$$

where P represents pressure, F represents force, and A the area of the region being considered. Of course, in the equation F = PA, we assume that the pressure is constant over the entire region A.

Most people know from experience that the deeper one dives underwater while swimming, the greater the pressure that is exerted by the water. This is due to the fact that the deeper one dives, the more water there is right on top of the swimmer: it is the force

that “column” of water exerts that determines the pressure the swimmer experiences. To get water pressure measured in its standard units (pounds per square foot), we say that the total water pressure is found by computing the total weight of the column of water that lies above a region of area 1 square foot at a fixed depth. Such a rectangular column with a 1 × 1 base and a depth of d feet has volume V = 1 · 1 · d ft^3, and thus the corresponding weight of the water overhead is 62.4d. Since this is also the amount of force being exerted on a 1 square foot region at a depth d feet underwater, we see that P = 62.4d (lbs/ft^2) is the pressure exerted by water at depth d.

The understanding that P = 62.4d will tell us the pressure exerted by water at a depth of d, along with the fact that F = PA, will now enable us to compute the total force that water exerts on a dam, as we see in the following example.

Example 6.5.3

Consider a trapezoid-shaped dam that is 60 feet wide at its base and 90 feet wide at its top, and assume the dam is 25 feet tall with water that rises to within 5 feet of the top of its face. Water weighs 62.5 pounds per cubic foot. How much force does the water exert against the dam?

Solution

First, we sketch a picture of the dam, as shown in Figure 6.18. Note that, as in problems involving the work to pump out a tank, we let the positive x-axis point down.

It is essential to use the fact that pressure is constant at a fixed depth. Hence, we consider a slice of water at constant depth on the face, such as the one shown in the figure. First, the approximate area of this slice is the area of the pictured rectangle. Since the width of that rectangle depends on the variable x (which represents the how far the slice lies from the top of the dam), we find a formula for the function $y = f(x)$ that determines one side of the face of the dam. Since f is linear, it is straightforward to find that $y = f(x) = 45 - \frac{3}{5}x$. Hence, the approximate area of a representative slice is

$$A_{\text{slice}} = 2f(x)\Delta x = 2(45 - \frac{3}{5}x)\Delta x.$$

At any point on this slice, the depth is approximately constant, and thus the pressure can be considered constant. In particular, we note that since x measures the distance to the top of the dam, and because the water rises to within 5 feet of the top of the dam, the depth of any point on the representative slice is approximately $(x - 5)$. Now, since pressure

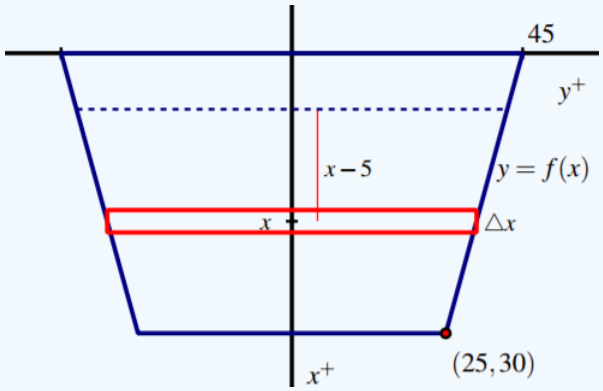


Figure 6.18: A trapezoidal dam that is 25 feet tall, 60 feet wide at its base, 90 feet wide at its top, with the water line 5 feet down from the top of its face.

is given by $P = 62.4d$, we have that at any point on the representative slice

$$P_{\text{slice}} = 62.4(x - 5).$$

Knowing both the pressure and area, we can find the force the water exerts on the slice. Using $F = PA$, it follows that

$$F_{\text{slice}} = P_{\text{slice}} \cdot A_{\text{slice}} = 62.4(x - 5) \cdot 2(45 - \frac{3}{5}x)\Delta x \ .$$

Finally, we use a definite integral to sum the forces over the appropriate range of x-values. Since the water rises to within 5 feet of the top of the dam, we start at $x = 5$ and slice all the way to the bottom of the dam, where $x = 30$. Hence,

$$F = \int_{x=5}^{x=30} 62.4(x-5) \cdot 2(45 - \frac{3}{5}x)dx\,.$$

Using technology to evaluate the integral, we find F ≈ 1.248 × 10^6 pounds.

Activity 6.5.4

In each of the following problems, determine the total force exerted by water against the surface that is described.

a. Consider a rectangular dam that is 100 feet wide and 50 feet tall, and suppose that water presses against the dam all the way to the top.
b. Consider a semicircular dam with a radius of 30 feet. Suppose that the water rises to within 10 feet of the top of the dam.
c. Consider a trough with triangular ends, as pictured in Figure 6.17, where the tank is 10 feet long, the top is 5 feet wide, and the tank is 4 feet deep. Say that the trough is full to within 1 foot of the top with water of weight density 62.4 pounds/ft^3. How much force does the water exert against one of the triangular ends?

While there are many different formulas that we use in solving problems involving work, force, and pressure, it is important to understand that the fundamental ideas behind these problems are similar to several others that we've encountered in applications of the definite integral. In particular, the basic idea is to take a difficult problem and somehow slice it into more manageable pieces that we understand, and then use a definite integral to add up these simpler pieces.

Summary

In this section, we encountered the following important ideas:

- To measure the work accomplished by a varying force that moves an object, we subdivide the problem into pieces on which we can use the formula W = F · d, and then use a definite integral to sum the work accomplished on each piece.
- To find the total force exerted by water against a dam, we use the formula F = P · A to measure the force exerted on a slice that lies at a fixed depth, and then use a definite integral to sum the forces across the appropriate range of depths.
- Because work is computed as the product of force and distance (provided force is constant), and the force water exerts on a dam can be computed as the product of pressure and area (provided pressure is constant), problems involving these concepts are similar to earlier problems we did using definite integrals to find distance (via "distance equals rate times time") and mass ("mass equals density times volume").

6.5: Work is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **6.4: Physics Applications - Work, Force, and Pressure** is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

6.6: Moments and Centers of Mass

Learning Objectives

- Find the center of mass of objects distributed along a line.
- Locate the center of mass of a thin plate.
- Use symmetry to help locate the centroid of a thin plate.
- Apply the theorem of Pappus for volume.

In this section, we consider centers of mass (also called *centroids*, under certain conditions) and moments. The basic idea of the center of mass is the notion of a balancing point. Many of us have seen performers who spin plates on the ends of sticks. The performers try to keep several of them spinning without allowing any of them to drop. If we look at a single plate (without spinning it), there is a sweet spot on the plate where it balances perfectly on the stick. If we put the stick anywhere other than that sweet spot, the plate does not balance and it falls to the ground. (That is why performers spin the plates; the spin helps keep the plates from falling even if the stick is not exactly in the right place.) Mathematically, that sweet spot is called the *center of mass of the plate*.

In this section, we first examine these concepts in a one-dimensional context, then expand our development to consider centers of mass of two-dimensional regions and symmetry. Last, we use centroids to find the volume of certain solids by applying the theorem of Pappus.

Center of Mass and Moments

Let's begin by looking at the center of mass in a one-dimensional context. Consider a long, thin wire or rod of negligible mass resting on a fulcrum, as shown in Figure $6.6.1a$. Now suppose we place objects having masses m_1 and m_2 at distances d_1 and d_2 from the fulcrum, respectively, as shown in Figure $6.6.1b$.

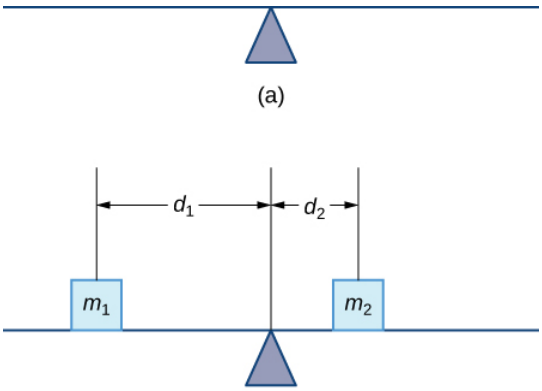


Figure 6.6.1: (a) A thin rod rests on a fulcrum. (b) Masses are placed on the rod.

The most common real-life example of a system like this is a playground seesaw, or teeter-totter, with children of different weights sitting at different distances from the center. On a seesaw, if one child sits at each end, the heavier child sinks down and the lighter child is lifted into the air. If the heavier child slides in toward the center, though, the seesaw balances. Applying this concept to the masses on the rod, we note that the masses balance each other if and only if

$$m_1 d_1 = m_2 d_2.$$

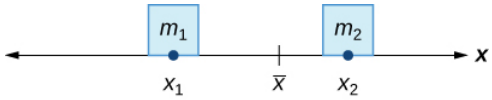


Figure 6.6.2: The center of mass $\bar{x}$ is the balance point of the system.

In the seesaw example, we balanced the system by moving the masses (children) with respect to the fulcrum. However, we are really interested in systems in which the masses are not allowed to move, and instead we balance the system by moving the

fulcrum. Suppose we have two point masses, m_1 and m_2, located on a number line at points x_1 and x_2, respectively (Figure 6.6.2). The center of mass, $\bar{x}$, is the point where the fulcrum should be placed to make the system balance.

Thus, we have

$$\begin{aligned} m_1|x_1-\bar{x}| &= m_2|x_2-\bar{x}| \\ m_1(\bar{x}-x_1) &= m_2(x_2-\bar{x}) \\ m_1\bar{x}-m_1x_1 &= m_2x_2-m_2\bar{x} \\ \bar{x}(m_1+m_2) &= m_1x_1+m_2x_2 \end{aligned}$$

or

$$\bar{x}=\frac{m_1x_1+m_2x_2}{m_1+m_2} \tag{6.6.1}$$

The expression in the numerator of Equation 6.6.1, $m_1x_1+m_2x_2$, is called the *first moment of the system* with respect to the origin. If the context is clear, we often drop the word first and just refer to this expression as the **moment** of the system. The expression in the denominator, m_1+m_2, is the total mass of the system. Thus, the **center of mass** of the system is the point at which the total mass of the system could be concentrated without changing the moment.

This idea is not limited just to two point masses. In general, if n masses, $m_1, m_2, \ldots, m_n$, are placed on a number line at points $x_1, x_2, \ldots, x_n$, respectively, then the center of mass of the system is given by

$$\bar{x}=\frac{\sum_{i=1}^{n} m_ix_i}{\sum_{i=1}^{n} m_i}$$

Center of Mass of Objects on a Line

Let $m_1, m_2, \ldots, m_n$ be point masses placed on a number line at points $x_1, x_2, \ldots, x_n$, respectively, and let $m=\sum_{i=1}^{n} m_i$ denote the total mass of the system. Then, the *moment of the system* with respect to the origin is given by

$$M=\sum_{i=1}^{n} m_ix_i \tag{6.6.2}$$

and the *center of mass of the system* is given by

$$\bar{x}=\frac{M}{m}. \tag{6.6.3}$$

We apply this theorem in the following example.

Example 6.6.1: Finding the Center of Mass of Objects along a Line

Suppose four point masses are placed on a number line as follows:

- $m_1=30\,kg,$ placed at $x_1=-2m$
- $m_2=5\,kg,$ placed at $x_2=3m$
- $m_3=10\,kg,$ placed at $x_3=6m$
- $m_4=15\,kg,$ placed at $x_4=-3m.$

Solution

Find the moment of the system with respect to the origin and find the center of mass of the system.

First, we need to calculate the moment of the system (Equation 6.6.2):

$$M = \sum_{i=1}^{4} m_i x_i$$
$$= -60 + 15 + 60 - 45$$
$$= -30.$$

Now, to find the center of mass, we need the total mass of the system:

$$m = \sum_{i=1}^{4} m_i$$
$$= 30 + 5 + 10 + 15$$
$$= 60\ kg$$

Then we have (from Equation 6.6.3)

$$\bar{x} - = \frac{M}{m} = -\frac{30}{60} = -\frac{1}{2}\ .$$

The center of mass is located 1/2 m to the left of the origin.

? Exercise 6.6.1

Suppose four point masses are placed on a number line as follows:

- $m_1 = 12\ kg$ placed at $x_1 = -4m$
- $m_2 = 12\ kg$ placed at $x_2 = 4m$
- $m_3 = 30\ kg$ placed at $x_3 = 2m$
- $m_4 = 6\ kg,$ placed at $x_4 = -6m.$

Find the moment of the system with respect to the origin and find the center of mass of the system.

Hint

Use the process from the previous example.

Answer

$$M = 24, \bar{x} = \frac{2}{5}m$$

We can generalize this concept to find the center of mass of a system of point masses in a plane. Let m_1 be a point mass located at point (x_1, y_1) in the plane. Then the moment M_x of the mass with respect to the x-axis is given by $M_x = m_1 y_1$. Similarly, the moment M_y with respect to the y-axis is given by

$$M_y = m_1 x_1.$$

Notice that the x-coordinate of the point is used to calculate the moment with respect to the y-axis, and vice versa. The reason is that the x-coordinate gives the distance from the point mass to the y-axis, and the y-coordinate gives the distance to the x-axis (see the following figure).

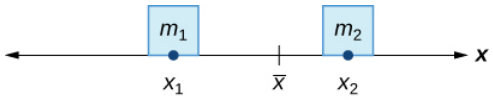


Figure 6.6.3: Point mass m_1 is located at point (x_1, y_1) in the plane.

If we have several point masses in the xy-plane, we can use the moments with respect to the x- and y-axes to calculate the x- and y-coordinates of the center of mass of the system.

Center of Mass of Objects in a Plane

Let m_1, m_2, ..., m_n be point masses located in the xy-plane at points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, respectively, and let $m = \sum_{i=1}^{n} m_i$ denote the total mass of the system. Then the moments M_x and M_y of the system with respect to the x- and y-axes, respectively, are given by

$$M_x = \sum_{i=1}^{n} m_i y_i \tag{6.6.4}$$

and

$$M_y = \sum_{i=1}^{n} m_i x_i. \tag{6.6.5}$$

Also, the coordinates of the center of mass $(\bar{x}, \bar{y})$ of the system are

$$\bar{x} = \frac{M_y}{m} \tag{6.6.6}$$

and

$$\bar{y} = \frac{M_x}{m}. \tag{6.6.7}$$

The next example demonstrates how to the center of mass formulas (Equations 6.6.4 - 6.6.7) may be applied.

Example 6.6.2: Finding the Center of Mass of Objects in a Plane

Suppose three point masses are placed in the xy-plane as follows (assume coordinates are given in meters):

- $m_1 = 2\,kg$ placed at $(-1, 3)$,
- $m_2 = 6\,kg$ placed at $(1, 1)$,
- $m_3 = 4\,kg$ placed at $(2, -2)$.

Find the center of mass of the system.

Solution

First we calculate the total mass of the system:

$$m = \sum_{i=1}^{3} m_i = 2 + 6 + 4 = 12\,kg.$$

Next we find the moments with respect to the x- and y-axes:

$$M_y = \sum_{i=1}^{3} m_i x_i = -2 + 6 + 8 = 12,$$

$$M_x = \sum_{i=1}^{3} m_i y_i = 6 + 6 - 8 = 4.$$

Then we have

$$\bar{x} = \frac{M_y}{m} = \frac{12}{12} = 1$$

and

$$\bar{y} = \frac{M_x}{m} = \frac{4}{12} = \frac{1}{3}.$$

The center of mass of the system is $(1, 1/3)$, in meters.

Exercise 6.6.2

Suppose three point masses are placed on a number line as follows (assume coordinates are given in meters):

- $m_1 = 5\,kg$, placed at $(-2, -3)$,
- $m_2 = 3\,kg$, placed at $(2, 3)$,
- $m_3 = 2\,kg$, placed at $(-3, -2)$.

Find the center of mass of the system.

Hint

Use the process from the previous example.

Answer

$(-1, -1)$ m

Center of Mass of Thin Plates

So far we have looked at systems of point masses on a line and in a plane. Now, instead of having the mass of a system concentrated at discrete points, we want to look at systems in which the mass of the system is distributed continuously across a thin sheet of material. For our purposes, we assume the sheet is thin enough that it can be treated as if it is two-dimensional. Such a sheet is called a **lamina**. Next we develop techniques to find the center of mass of a lamina. In this section, we also assume the density of the lamina is constant.

Laminas are often represented by a two-dimensional region in a plane. The geometric center of such a region is called its **centroid**. Since we have assumed the density of the lamina is constant, the center of mass of the lamina depends only on the shape of the corresponding region in the plane; it does not depend on the density. In this case, the center of mass of the lamina corresponds to the centroid of the delineated region in the plane. As with systems of point masses, we need to find the total mass of the lamina, as well as the moments of the lamina with respect to the x- and y-axes.

We first consider a lamina in the shape of a rectangle. Recall that the center of mass of a lamina is the point where the lamina balances. For a rectangle, that point is both the horizontal and vertical center of the rectangle. Based on this understanding, it is clear that the center of mass of a rectangular lamina is the point where the diagonals intersect, which is a result of the **symmetry principle**, and it is stated here without proof.

The Symmetry Principle

If a region R is symmetric about a line l, then the centroid of R lies on l.

Let's turn to more general laminas. Suppose we have a lamina bounded above by the graph of a continuous function $f(x)$, below by the x-axis, and on the left and right by the lines $x = a$ and $x = b$, respectively, as shown in the following figure.

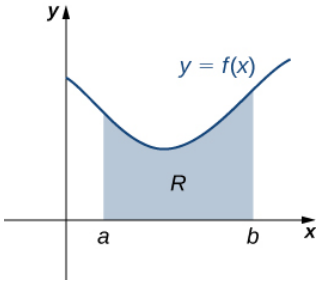


Figure 6.6.4: A region in the plane representing a lamina.

As with systems of point masses, to find the center of mass of the lamina, we need to find the total mass of the lamina, as well as the moments of the lamina with respect to the x- and y-axes. As we have done many times before, we approximate these quantities

by partitioning the interval $[a, b]$ and constructing rectangles.

For $i = 0, 1, 2, \ldots, n$, let $P = x_i$ be a regular partition of $[a, b]$. Recall that we can choose any point within the interval $[x_{i-1}, x_i]$ as our x_i^*. In this case, we want x_i^* to be the x-coordinate of the centroid of our rectangles. Thus, for $i = 1, 2, \ldots, n$, we select $x_i^* \in [x_{i-1}, x_i]$ such that x_i^* is the midpoint of the interval. That is, $x_i^* = (x_{i-1} + x_i)/2$. Now, for $i = 1, 2, \ldots, n$, construct a rectangle of height $f(x_i^*)$ on $[x_{i-1}, x_i]$. The center of mass of this rectangle is $(x_i^*, (f(x_i^*))/2)$, as shown in the following figure.

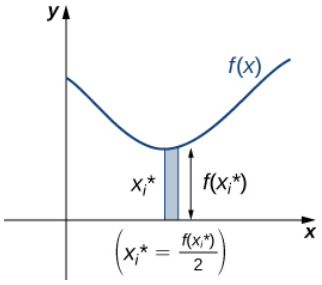


Figure 6.6.5: A representative rectangle of the lamina.

Next, we need to find the total mass of the rectangle. Let ρ represent the density of the lamina (note that ρ is a constant). In this case, ρ is expressed in terms of mass per unit area. Thus, to find the total mass of the rectangle, we multiply the area of the rectangle by ρ. Then, the mass of the rectangle is given by $\rho f(x_i^*)\Delta x$.

To get the approximate mass of the lamina, we add the masses of all the rectangles to get

$$m \approx \sum_{i=1}^{n} \rho f(x_i^*)\Delta x. \tag{6.6.8}$$

Equation 6.6.8 is a Riemann sum. Taking the limit as $n \to \infty$ gives the exact mass of the lamina:

$$\begin{aligned} m &= \lim_{n\to\infty} \sum_{i=1}^{n} \rho f(x_i^*)\Delta x \\ &= \rho \int_a^b f(x)dx. \end{aligned}$$

Next, we calculate the moment of the lamina with respect to the x-axis. Returning to the representative rectangle, recall its center of mass is $(x_i^*, (f(x_i^*))/2)$. Recall also that treating the rectangle as if it is a point mass located at the center of mass does not change the moment. Thus, the moment of the rectangle with respect to the x-axis is given by the mass of the rectangle, $\rho f(x_i^*)\Delta x$, multiplied by the distance from the center of mass to the x-axis: $(f(x_i^*))/2$. Therefore, the moment with respect to the x-axis of the rectangle is $\rho([f(x_i^*)]^2/2)\Delta x$. Adding the moments of the rectangles and taking the limit of the resulting Riemann sum, we see that the moment of the lamina with respect to the x-axis is

$$\begin{aligned} M_x &= \lim_{n\to\infty} \sum_{i=1}^{n} \rho \frac{[f(x_i^*)]^2}{2}\Delta x \\ &= \rho \int_a^b \frac{[f(x)]^2}{2}dx. \end{aligned}$$

We derive the moment with respect to the y-axis similarly, noting that the distance from the center of mass of the rectangle to the y-axis is x_i^*. Then the moment of the lamina with respect to the y-axis is given by

$$\begin{aligned} M_y &= \lim_{n\to\infty} \sum_{i=1}^{n} \rho x_i^* f(x^*)i)\Delta x \\ &= \rho \int_a^b x f(x)dx. \end{aligned}$$

We find the coordinates of the center of mass by dividing the moments by the total mass to give $\bar{x} = M_y/m$ and $\bar{y} = M_x/m$. If we look closely at the expressions for M_x, M_y, and m, we notice that the constant ρ cancels out when $\bar{x}$ and $\bar{y}$ are calculated.

We summarize these findings in the following theorem.

Center of Mass of a Thin Plate in the xy-Plane

Let R denote a region bounded above by the graph of a continuous function $f(x)$, below by the x-axis, and on the left and right by the lines $x = a$ and $x = b$, respectively. Let ρ denote the density of the associated lamina. Then we can make the following statements:

i. The mass of the lamina is

$$m = \rho \int_a^b f(x)dx. \tag{6.6.9}$$

ii. The moments M_x and M_y of the lamina with respect to the x- and y-axes, respectively, are

$$M_x = \rho \int_a^b \frac{[f(x)]^2}{2} dx \tag{6.6.10}$$

and

$$M_y = \rho \int_a^b x f(x)dx. \tag{6.6.11}$$

iii. The coordinates of the center of mass $(\bar{x}, \bar{y})$ are

$$\bar{x} = \frac{M_y}{m} \tag{6.6.12}$$

and

$$\bar{y} = \frac{M_x}{m}. \tag{6.6.13}$$

In the next example, we use this theorem to find the center of mass of a lamina.

Example 6.6.3: Finding the Center of Mass of a Lamina

Let R be the region bounded above by the graph of the function $f(x) = \sqrt{x}$ and below by the x-axis over the interval $[0, 4]$. Find the centroid of the region.

Solution

The region is depicted in the following figure.

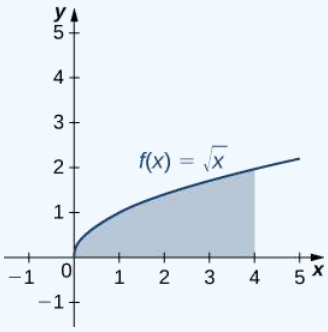


Figure 6.6.6: Finding the center of mass of a lamina.

Since we are only asked for the centroid of the region, rather than the mass or moments of the associated lamina, we know the density constant ρ cancels out of the calculations eventually. Therefore, for the sake of convenience, let's assume $\rho = 1$.

First, we need to calculate the total mass (Equation 6.6.9):

$$\begin{aligned} m &= \rho \int_a^b f(x)dx \\ &= \int_0^4 \sqrt{x}dx \\ &= \frac{2}{3}x^{3/2}\Big|_0^4 \\ &= \frac{2}{3}[8-0] \\ &= \frac{16}{3}. \end{aligned}$$

Next, we compute the moments (Equation 6.6.12):

$$\begin{aligned} M_x &= \rho \int_a^b \frac{[f(x)]^2}{2}dx \\ &= \int_0^4 \frac{x}{2}dx \\ &= \frac{1}{4}x^2\Big|_0^4 \\ &= 4 \end{aligned}$$

and (Equation 6.6.11):

$$\begin{aligned} M_y &= \rho \int_a^b xf(x)dx \\ &= \int_0^4 x\sqrt{x}dx \\ &= \int_0^4 x^{3/2}dx \\ &= \frac{2}{5}x^{5/2}\Big|_0^4 \\ &= \frac{2}{5}[32-0] \\ &= \frac{64}{5}. \end{aligned}$$

Thus, we have (Equation 6.6.12):

$$\begin{aligned} \bar{x} &= \frac{M_y}{m} \\ &= \frac{64/5}{16/3} \\ &= \frac{64}{5}\cdot\frac{3}{16} \\ &= \frac{12}{5} \end{aligned}$$

and (Equation 6.6.13):

$$\begin{aligned}\bar{y} &= \frac{M_x}{y} \\ &= \frac{4}{16/3} \\ &= 4 \cdot \frac{3}{16} \\ &= \frac{3}{4}.\end{aligned}$$

The centroid of the region is $(12/5, 3/4)$.

Exercise 6.6.3

Let R be the region bounded above by the graph of the function $f(x) = x^2$ and below by the x-axis over the interval $[0, 2]$. Find the centroid of the region.

Hint

Use the process from the previous example.

Answer

The centroid of the region is $(3/2, 6/5)$.

We can adapt this approach to find centroids of more complex regions as well. Suppose our region is bounded above by the graph of a continuous function $f(x)$, as before, but now, instead of having the lower bound for the region be the x-axis, suppose the region is bounded below by the graph of a second continuous function, $g(x)$, as shown in Figure 6.6.7.

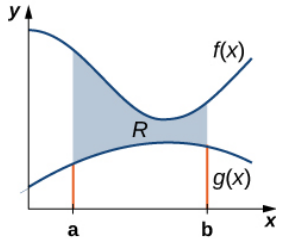


Figure 6.6.7: A region between two functions.

Again, we partition the interval $[a, b]$ and construct rectangles. A representative rectangle is shown in Figure 6.6.8.

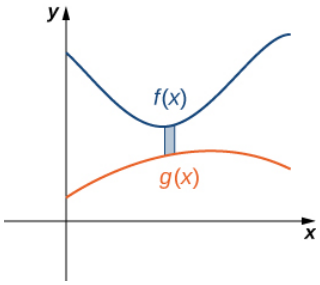


Figure 6.6.8: A representative rectangle of the region between two functions.

Note that the centroid of this rectangle is $(x_i^*, (f(x_i^*) + g(x_i^*))/2)$. We won't go through all the details of the Riemann sum development, but let's look at some of the key steps. In the development of the formulas for the mass of the lamina and the moment with respect to the y-axis, the height of each rectangle is given by $f(x_i^*) - g(x_i^*)$, which leads to the expression $f(x) - g(x)$ in the integrands.

In the development of the formula for the moment with respect to the x-axis, the moment of each rectangle is found by multiplying the area of the rectangle, $\rho[f(x_i^*) - g(x_i^*)]\Delta x$, by the distance of the centroid from the x-axis, $(f(x_i^*) + g(x_i^*))/2$, which gives $\rho(1/2)[f(x_i^*)]^2 - [g(x_i^*)]^2 \Delta x$. Summarizing these findings, we arrive at the following theorem.

Center of Mass of a Lamina Bounded by Two Functions

Let R denote a region bounded above by the graph of a continuous function $f(x)$, below by the graph of the continuous function $g(x)$, and on the left and right by the lines $x = a$ and $x = b$, respectively. Let ρ denote the density of the associated lamina. Then we can make the following statements:

i. The mass of the lamina is

$$m = \rho \int_a^b [f(x) - g(x)]dx.$$

ii. The moments M_x and M_y of the lamina with respect to the x- and y-axes, respectively, are

$$M_x = \rho \int_a^b \frac{1}{2}([f(x)]^2 - [g(x)]^2)dx$$

and

$$M_y = \rho \int_a^b x[f(x) - g(x)]dx.$$

iii. The coordinates of the center of mass $\bar{x}, \bar{y})$ are

$$\bar{x} = \frac{M_y}{m}$$

and

$$\bar{y} = \frac{M_x}{m}$$

We illustrate this theorem in the following example.

Example 6.6.4: Finding the Centroid of a Region Bounded by Two Functions

Let R be the region bounded above by the graph of the function $f(x) = 1 - x^2$ and below by the graph of the function $g(x) = x - 1$. Find the centroid of the region.

Solution

The region is depicted in the following figure.

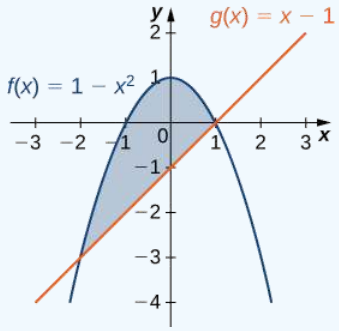


Figure 6.6.9: Finding the centroid of a region between two curves.

The graphs of the functions intersect at $(-2, -3)$ and $(1, 0)$, so we integrate from -2 to 1. Once again, for the sake of convenience, assume $\rho = 1$.

First, we need to calculate the total mass:

$$\begin{aligned} m &= \rho \int_a^b [f(x) - g(x)]dx \\ &= \int_{-2}^1 [1 - x^2 - (x-1)]dx \\ &= \int_{-2}^1 (2 - x^2 - x)dx \\ &= \left[2x - \frac{1}{3}x^3 - \frac{1}{2}x^2\right]\Big|_{-2}^1 \\ &= \left[2 - \frac{1}{3} - \frac{1}{2}\right] - \left[-4 + \frac{8}{3} - 2\right] \\ &= \frac{9}{2}. \end{aligned}$$

Next, we compute the moments:

$$\begin{aligned} M_x &= \rho \int_a^b \frac{1}{2}([f(x)]^2 - [g(x)]^2)dx \\ &= \frac{1}{2}\int_{-2}^1 ((1-x^2)^2 - (x-1)^2)dx \\ &= \frac{1}{2}\int_{-2}^1 (x^4 - 3x^2 + 2x)dx \\ &= \frac{1}{2}\left[\frac{x^5}{5} - x^3 + x^2\right]\Big|_{-2}^1 \\ &= -\frac{27}{10} \end{aligned}$$

and

$$\begin{aligned} M_y &= \rho \int_a^b x[f(x) - g(x)]dx \\ &= \int_{-2}^1 x[(1-x^2) - (x-1)]dx \\ &= \int_{-2}^1 x[2 - x^2 - x]dx \\ &= \int_{-2}^1 (2x - x^3 - x^2)dx \\ &= \left[x^2 - \frac{x^4}{4} - \frac{x^3}{3}\right]\Big|_{-2}^1 \\ &= -\frac{9}{4}. \end{aligned}$$

Therefore, we have

$$\begin{aligned}\bar{x} &= \frac{M_y}{m} \\ &= -\frac{9}{4} \cdot \frac{2}{9} \\ &= -\frac{1}{2}\end{aligned}$$

and

$$\begin{aligned}\bar{y} &= \frac{M_x}{y} \\ &= -\frac{27}{10} \cdot \frac{2}{9} \\ &= -\frac{3}{5}.\end{aligned}$$

The centroid of the region is $(-(1/2), -(3/5))$.

Exercise 6.6.4

Let R be the region bounded above by the graph of the function $f(x) = 6 - x^2$ and below by the graph of the function $g(x) = 3 - 2x$. Find the centroid of the region.

Hint

Use the process from the previous example.

Answer

The centroid of the region is $(1, 13/5)$.

The Symmetry Principle

We stated the symmetry principle earlier, when we were looking at the centroid of a rectangle. The symmetry principle can be a great help when finding centroids of regions that are symmetric. Consider the following example.

Example 6.6.5: Finding the Centroid of a Symmetric Region

Let R be the region bounded above by the graph of the function $f(x) = 4 - x^2$ and below by the x-axis. Find the centroid of the region.

Solution

The region is depicted in the following figure

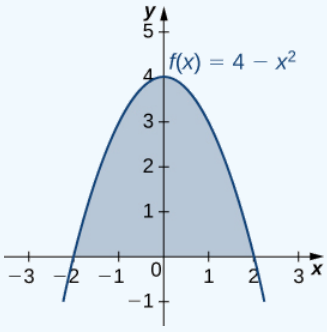


Figure 6.6.10: We can use the symmetry principle to help find the centroid of a symmetric region.

The region is symmetric with respect to the y-axis. Therefore, the x-coordinate of the centroid is zero. We need only calculate $\bar{y}$. Once again, for the sake of convenience, assume $\rho = 1$.

First, we calculate the total mass:

$$\begin{aligned} m &= \rho \int_a^b f(x)dx \\ &= \int_{-2}^{2} (4-x^2)dx \\ &= \left[4x - \frac{x^3}{3}\right]\Big|_{-2}^{2} \\ &= \frac{32}{3}. \end{aligned}$$

Next, we calculate the moments. We only need M_x:

$$\begin{aligned} M_x &= \rho \int_a^b \frac{[f(x)]^2}{2} dx \\ &= \frac{1}{2}\int_{-2}^{2} \left[4-x^2\right]^2 dx = \frac{1}{2}\int_{-2}^{2} (16 - 8x^2 + x^4)dx \\ &= \frac{1}{2}\left[\frac{x^5}{5} - \frac{8x^3}{3} + 16x\right] |_{-2}^{2} = \frac{256}{15} \end{aligned}$$

Then we have

$$\bar{y} = \frac{M_x}{y} = \frac{256}{15} \cdot \frac{3}{32} = \frac{8}{5}.$$

The centroid of the region is $(0, 8/5)$.

Exercise 6.6.5

Let R be the region bounded above by the graph of the function $f(x) = 1 - x^2$ and below by x-axis. Find the centroid of the region.

Hint

Use the process from the previous example.

Answer

The centroid of the region is $(0, 2/5)$.

The Grand Canyon Skywalk

The Grand Canyon Skywalk opened to the public on March 28, 2007. This engineering marvel is a horseshoe-shaped observation platform suspended 4000 ft above the Colorado River on the West Rim of the Grand Canyon. Its crystal-clear glass floor allows stunning views of the canyon below (see the following figure).



Figure 6.6.11: The Grand Canyon Skywalk offers magnificent views of the canyon. (credit: 10da_ralta, Wikimedia Commons)

The Skywalk is a cantilever design, meaning that the observation platform extends over the rim of the canyon, with no visible means of support below it. Despite the lack of visible support posts or struts, cantilever structures are engineered to be very stable and the Skywalk is no exception. The observation platform is attached firmly to support posts that extend 46 ft down into bedrock. The structure was built to withstand 100-mph winds and an 8.0-magnitude earthquake within 50 mi, and is capable of supporting more than 70,000,000 lb.

One factor affecting the stability of the Skywalk is the center of gravity of the structure. We are going to calculate the center of gravity of the Skywalk, and examine how the center of gravity changes when tourists walk out onto the observation platform.

The observation platform is U-shaped. The legs of the U are 10 ft wide and begin on land, under the visitors' center, 48 ft from the edge of the canyon. The platform extends 70 ft over the edge of the canyon.

To calculate the center of mass of the structure, we treat it as a lamina and use a two-dimensional region in the xy-plane to represent the platform. We begin by dividing the region into three subregions so we can consider each subregion separately. The first region, denoted R_1, consists of the curved part of the U. We model R_1 as a semicircular annulus, with inner radius 25 ft and outer radius 35 ft, centered at the origin (Figure 6.6.12).

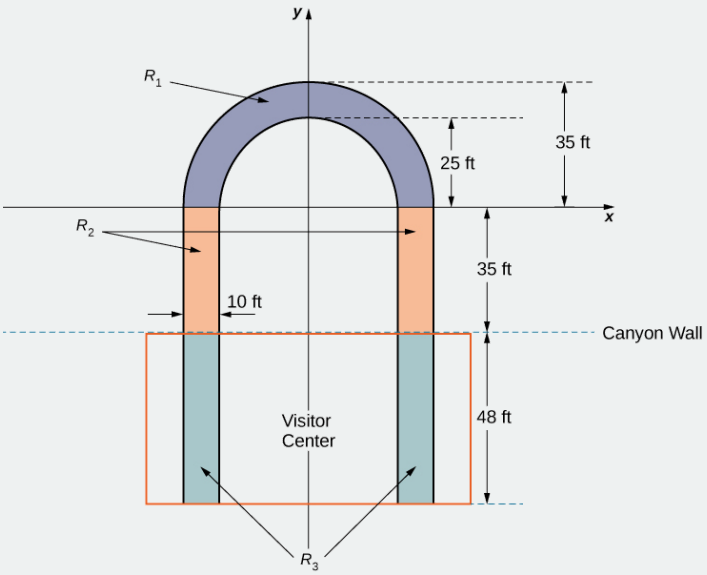


Figure 6.6.12: We model the Skywalk with three sub-regions.

The legs of the platform, extending 35 ft between R_1 and the canyon wall, comprise the second sub-region, R_2. Last, the ends of the legs, which extend 48 ft under the visitor center, comprise the third sub-region, R_3. Assume the density of the lamina is constant and assume the total weight of the platform is 1,200,000 lb (not including the weight of the visitor center; we will consider that later). Use $g = 32\ ft/sec^2$.

1. Compute the area of each of the three sub-regions. Note that the areas of regions R_2 and R_3 should include the areas of the legs only, not the open space between them. Round answers to the nearest square foot.

2. Determine the mass associated with each of the three sub-regions.
3. Calculate the center of mass of each of the three sub-regions.
4. Now, treat each of the three sub-regions as a point mass located at the center of mass of the corresponding sub-region. Using this representation, calculate the center of mass of the entire platform.
5. Assume the visitor center weighs 2,200,000 lb, with a center of mass corresponding to the center of mass of R_3.Treating the visitor center as a point mass, recalculate the center of mass of the system. How does the center of mass change?
6. Although the Skywalk was built to limit the number of people on the observation platform to 120, the platform is capable of supporting up to 800 people weighing 200 lb each. If all 800 people were allowed on the platform, and all of them went to the farthest end of the platform, how would the center of gravity of the system be affected? (Include the visitor center in the calculations and represent the people by a point mass located at the farthest edge of the platform, 70 ft from the canyon wall.)

Theorem of Pappus

This section ends with a discussion of the *theorem of Pappus for volume*, which allows us to find the volume of particular kinds of solids by using the centroid. (There is also a theorem of Pappus for surface area, but it is much less useful than the theorem for volume.)

Theorem of Pappus for Volume

Let R be a region in the plane and let l be a line in the plane that does not intersect R. Then the volume of the solid of revolution formed by revolving R around l is equal to the area of R multiplied by the distance d traveled by the centroid of R.

Proof

We can prove the case when the region is bounded above by the graph of a function $f(x)$ and below by the graph of a function $g(x)$ over an interval $[a,b]$, and for which the axis of revolution is the y-axis. In this case, the area of the region is $A=\int_a^b [f(x)-g(x)]\,dx$. Since the axis of rotation is the y-axis, the distance traveled by the centroid of the region depends only on the x-coordinate of the centroid, $\bar{x}$, which is

$$x=\frac{M_y}{m},$$

where

$$m=\rho\int_a^b [f(x)-g(x)]dx$$

and

$$M_y=\rho\int_a^b x[f(x)-g(x)]dx.$$

Then,

$$d=2\pi\frac{\rho\int_a^b x[f(x)-g(x)]dx}{\rho\int_a^b [f(x)-g(x)]dx}$$

and thus

$$d\cdot A=2\pi\int_a^b x[f(x)-g(x)]dx.$$

However, using the method of cylindrical shells, we have

$$V = 2\pi \int_a^b x[f(x) - g(x)]dx.$$

So,

$$V = d \cdot A$$

and the proof is complete.

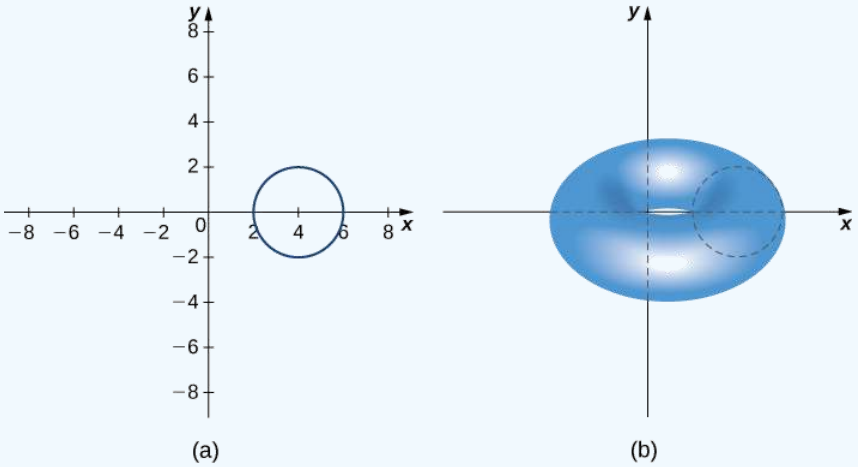
□

Example 6.6.6: Using the Theorem of Pappus for Volume

Let R be a circle of radius 2 centered at $(4, 0)$. Use the theorem of Pappus for volume to find the volume of the torus generated by revolving R around the y-axis.

Solution

The region and torus are depicted in the following figure.



(a) (b)

Figure 6.6.13: Determining the volume of a torus by using the theorem of Pappus. (a) A circular region R in the plane; (b) the torus generated by revolving R about the y-axis.

The region R is a circle of radius 2, so the area of R is $A = 4\pi$ units^2. By the symmetry principle, the centroid of R is the center of the circle. The centroid travels around the y-axis in a circular path of radius 4, so the centroid travels $d = 8\pi$ units. Then, the volume of the torus is $A \cdot d = 32\pi^2$ units^3.

Exercise 6.6.6

Let R be a circle of radius 1 centered at $(3, 0)$. Use the theorem of Pappus for volume to find the volume of the torus generated by revolving R around the y-axis.

Hint

Use the process from the previous example.

Answer

$6\pi^2$ units^3

Key Concepts

- Mathematically, the center of mass of a system is the point at which the total mass of the system could be concentrated without changing the moment. Loosely speaking, the center of mass can be thought of as the balancing point of the system.

- For point masses distributed along a number line, the moment of the system with respect to the origin is $M = \sum_{i=1}^{n} m_i x_i$. For point masses distributed in a plane, the moments of the system with respect to the x- and y-axes, respectively, are $M_x = \sum_{i=1}^{n} m_i y_i$ and $M_y = \sum_{i=}^{n} m_i x_i$, respectively.
- For a lamina bounded above by a function $f(x)$, the moments of the system with respect to the x- and y-axes, respectively, are $M_x = \rho \int_a^b \frac{[f(x)]^2}{2}\, dx$ and $M_y = \rho \int_a^b x f(x)\, dx$.
- The x- and y-coordinates of the center of mass can be found by dividing the moments around the y-axis and around the x-axis, respectively, by the total mass. The symmetry principle says that if a region is symmetric with respect to a line, then the centroid of the region lies on the line.
- The theorem of Pappus for volume says that if a region is revolved around an external axis, the volume of the resulting solid is equal to the area of the region multiplied by the distance traveled by the centroid of the region.

Key Equations

- **Mass of a lamina**

$$m = \rho \int_a^b f(x) dx$$

- **Moments of a lamina**

$$M_x = \rho \int_a^b \frac{[f(x)]^2}{2}\, dx \text{ and } M_y = \rho \int_a^b x f(x)\, dx$$

- **Center of mass of a lamina**

$$\bar{x} = \frac{M_y}{m} \text{ and } \bar{y} = \frac{M_x}{m}$$

Glossary

center of mass

the point at which the total mass of the system could be concentrated without changing the moment

centroid

the centroid of a region is the geometric center of the region; laminas are often represented by regions in the plane; if the lamina has a constant density, the center of mass of the lamina depends only on the shape of the corresponding planar region; in this case, the center of mass of the lamina corresponds to the centroid of the representative region

lamina

a thin sheet of material; laminas are thin enough that, for mathematical purposes, they can be treated as if they are two-dimensional

moment

if n masses are arranged on a number line, the moment of the system with respect to the origin is given by $M = \sum_{i=1}^{n} m_i x_i$; if, instead, we consider a region in the plane, bounded above by a function $f(x)$ over an interval $[a, b]$, then the moments of the region with respect to the x- and y-axes are given by $M_x = \rho \int_a^b \frac{[f(x)]^2}{2}\, dx$ and $M_y = \rho \int_a^b x f(x)\, dx$, respectively

symmetry principle

the symmetry principle states that if a region R is symmetric about a line I, then the centroid of R lies on I

theorem of Pappus for volume

this theorem states that the volume of a solid of revolution formed by revolving a region around an external axis is equal to the area of the region multiplied by the distance traveled by the centroid of the region

---

6.6: Moments and Centers of Mass is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **6.6: Moments and Centers of Mass** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

CHAPTER OVERVIEW

7: Integrals and Transcendental Functions

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

7.1: The Logarithm Defined as an Integral

7.2: Exponential Change and Separable Differential Equations

7.3: Hyperbolic Functions

7: Integrals and Transcendental Functions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.1: The Logarithm Defined as an Integral

Learning Objectives

- Write the definition of the natural logarithm as an integral.
- Recognize the derivative of the natural logarithm.
- Integrate functions involving the natural logarithmic function.
- Define the number e through an integral.
- Recognize the derivative and integral of the exponential function.
- Prove properties of logarithms and exponential functions using integrals.
- Express general logarithmic and exponential functions in terms of natural logarithms and exponentials.

We already examined exponential functions and logarithms in earlier chapters. However, we glossed over some key details in the previous discussions. For example, we did not study how to treat exponential functions with exponents that are irrational. The definition of the number e is another area where the previous development was somewhat incomplete. We now have the tools to deal with these concepts in a more mathematically rigorous way, and we do so in this section.

For purposes of this section, assume we have not yet defined the natural logarithm, the number e, or any of the integration and differentiation formulas associated with these functions. By the end of the section, we will have studied these concepts in a mathematically rigorous way (and we will see they are consistent with the concepts we learned earlier). We begin the section by defining the natural logarithm in terms of an integral. This definition forms the foundation for the section. From this definition, we derive differentiation formulas, define the number e, and expand these concepts to logarithms and exponential functions of any base.

The Natural Logarithm as an Integral

Recall the power rule for integrals:

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C, \quad n \neq -1.$$

Clearly, this does not work when $n = -1,$ as it would force us to divide by zero. So, what do we do with $\int \frac{1}{x}\,dx$? Recall from the Fundamental Theorem of Calculus that $\int_1^x \frac{1}{t}dt$ is an antiderivative of $\frac{1}{x}$. Therefore, we can make the following definition.

Definition: The Natural Logarithm

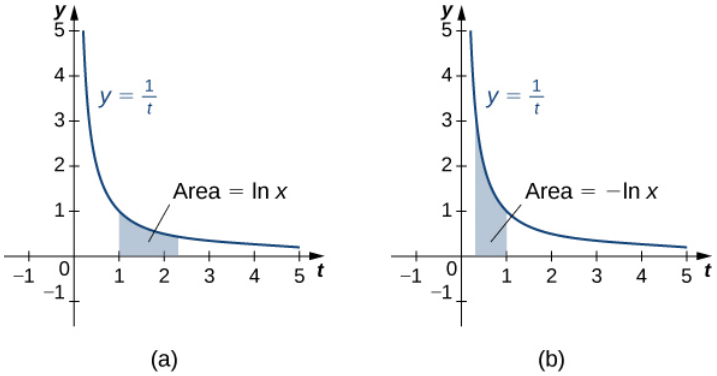
For $x > 0$, define the natural logarithm function by

$$\ln x = \int_1^x \frac{1}{t}\,dt.$$

For $x > 1$, this is just the area under the curve $y = \frac{1}{t}$ from 1 to x. For $x < 1$, we have

$$\int_1^x \frac{1}{t}\,dt = -\int_x^1 \frac{1}{t}\,dt,$$

so in this case it is the negative of the area under the curve from x to 1 (see the following figure).



(a) (b)

Figure 7.1.1: (a) When $x > 1$, the natural logarithm is the area under the curve $y = 1/t$ from 1 to x. (b) When $x < 1$, the natural logarithm is the negative of the area under the curve from x to 1.

Notice that $\ln 1 = 0$. Furthermore, the function $y = \dfrac{1}{t} > 0$ for $x > 0$. Therefore, by the properties of integrals, it is clear that $\ln x$ is increasing for $x > 0$.

Properties of the Natural Logarithm

Because of the way we defined the natural logarithm, the following differentiation formula falls out immediately as a result of to the Fundamental Theorem of Calculus.

Definition: Derivative of the Natural Logarithm

For $x > 0$, the derivative of the natural logarithm is given by

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

Corollary to the Derivative of the Natural Logarithm

The function $\ln x$ is differentiable; therefore, it is continuous.

A graph of $\ln x$ is shown in Figure. Notice that it is continuous throughout its domain of $(0, \infty)$.

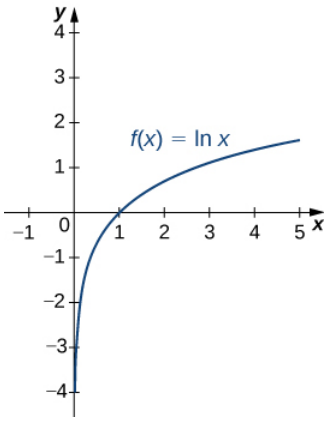


Figure 7.1.2: The graph of $f(x) = \ln x$ shows that it is a continuous function.

Example 7.1.1: Calculating Derivatives of Natural Logarithms

Calculate the following derivatives:

a. $\dfrac{d}{dx}\Big(\ln(5x^3 - 2)\Big)$

b. $\dfrac{d}{dx}\Big((\ln(3x))^2\Big)$

Solution

We need to apply the chain rule in both cases.

a. $\dfrac{d}{dx}\Big(\ln(5x^3-2)\Big)=\dfrac{15x^2}{5x^3-2}$

b. $\dfrac{d}{dx}\Big((\ln(3x))^2\Big)=\dfrac{2(\ln(3x))\cdot 3}{3x}=\dfrac{2(\ln(3x))}{x}$

? Exercise 7.1.1

Calculate the following derivatives:

a. $\dfrac{d}{dx}\Big(\ln(2x^2+x)\Big)$

b. $\dfrac{d}{dx}\Big((\ln(x^3))^2\Big)$

Hint

Apply the differentiation formula just provided and use the chain rule as necessary.

Answer

a. $\dfrac{d}{dx}\Big(\ln(2x^2+x)\Big)=\dfrac{4x+1}{2x^2+x}$

b. $\dfrac{d}{dx}\Big((\ln(x^3))^2\Big)=\dfrac{6\ln(x^3)}{x}$

Note that if we use the absolute value function and create a new function $\ln|x|$, we can extend the domain of the natural logarithm to include $x<0$. Then $\dfrac{d}{dx}\Big(\ln x\Big)=\dfrac{1}{x}$. This gives rise to the familiar integration formula.

Integral of $\frac{1}{u}\,du$

The natural logarithm is the antiderivative of the function $f(u)=\dfrac{1}{u}$:

$$\int \frac{1}{u}\,du=\ln|u|+C.$$

✓ Example 7.1.2: Calculating Integrals Involving Natural Logarithms

Calculate the integral $\displaystyle\int \frac{x}{x^2+4}\,dx$.

Solution

Using u-substitution, let $u=x^2+4$. Then $du=2x\,dx$ and we have

$$\int \frac{x}{x^2+4}\,dx=\frac{1}{2}\int \frac{1}{u}\,du=\frac{1}{2}\ln|u|+C=\frac{1}{2}\ln|x^2+4|+C=\frac{1}{2}\ln(x^2+4)+C.$$

? Exercise 7.1.2

Calculate the integral $\displaystyle\int \frac{x^2}{x^3+6}\,dx$.

Hint

Apply the integration formula provided earlier and use u-substitution as necessary.

Answer

$$\int \frac{x^2}{x^3+6}\,dx = \frac{1}{3}\ln|\,x^3+6\,|+C$$

Although we have called our function a "logarithm," we have not actually proved that any of the properties of logarithms hold for this function. We do so here.

Properties of the Natural Logarithm

If $a,\ b>0$ and r is a rational number, then

i. $\ln 1 = 0$
ii. $\ln(ab) = \ln a + \ln b$
iii. $\ln\left(\frac{a}{b}\right) = \ln a - \ln b$
iv. $\ln(a^r) = r\ln a$

Proof

i. By definition, $\ln 1 = \int_1^1 \frac{1}{t}\,dt = 0.$

ii. We have

$$\ln(ab) = \int_1^{ab} \frac{1}{t}\,dt = \int_1^a \frac{1}{t}\,dt + \int_a^{ab} \frac{1}{t}\,dt.$$

Use $u-substitution$ on the last integral in this expression. Let $u = t/a$. Then $du = (1/a)dt$. Furthermore, when $t=a,\ u=1$, and when $t=ab,\ u=b$. So we get

$$\ln(ab) = \int_1^a \frac{1}{t}\,dt + \int_a^{ab} \frac{1}{t}\,dt = \int_1^a \frac{1}{t}\,dt + \int_1^{ab} \frac{a}{t}\cdot\frac{1}{a}\,dt = \int_1^a \frac{1}{t}\,dt + \int_1^b \frac{1}{u}\,du = \ln a + \ln b.$$

iii. Note that

$$\frac{d}{dx}\left(\ln(x^r)\right) = \frac{rx^{r-1}}{x^r} = \frac{r}{x}\ .$$

Furthermore,

$$\frac{d}{dx}\Big((r\ln x)\Big) = \frac{r}{x}.$$

Since the derivatives of these two functions are the same, by the Fundamental Theorem of Calculus, they must differ by a constant. So we have

$$\ln(x^r) = r\ln x + C$$

for some constant C. Taking $x=1$, we get

$$\ln(1^r) = r\ln(1) + C$$
$$0 = r(0) + C$$
$$C = 0.$$

Thus $\ln(x^r) = r\ln x$ and the proof is complete. Note that we can extend this property to irrational values of r later in this section.

Part iii. follows from parts ii. and iv. and the proof is left to you.

□

Example 7.1.3: Using Properties of Logarithms

Use properties of logarithms to simplify the following expression into a single logarithm:

$$\ln 9 - 2\ln 3 + \ln\left(\tfrac{1}{3}\right).$$

Solution

We have

$$\ln 9 - 2\ln 3 + \ln\left(\tfrac{1}{3}\right) = \ln(3^2) - 2\ln 3 + \ln(3^{-1}) = 2\ln 3 - 2\ln 3 - \ln 3 = -\ln 3.$$

Exercise 7.1.3

Use properties of logarithms to simplify the following expression into a single logarithm:

$$\ln 8 - \ln 2 - \ln\left(\tfrac{1}{4}\right)$$

Hint

Apply the properties of logarithms.

Answer

$4\ln 2$

Defining the Number e

Now that we have the natural logarithm defined, we can use that function to define the number e.

Definition: e

The number e is defined to be the real number such that

$$\ln e = 1$$

To put it another way, the area under the curve $y = 1/t$ between $t = 1$ and $t = e$ is 1 (Figure). The proof that such a number exists and is unique is left to you. (Hint: Use the Intermediate Value Theorem to prove existence and the fact that $\ln x$ is increasing to prove uniqueness.)

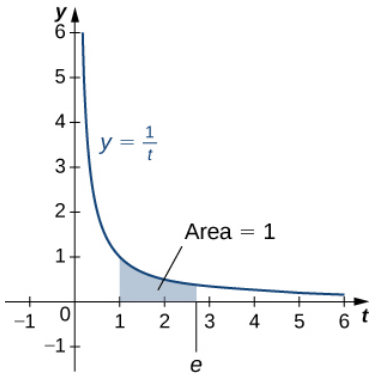


Figure 7.1.3:The area under the curve from 1 to e is equal to one.

The number e can be shown to be *irrational*, although we won't do so here (see the Student Project in Taylor and Maclaurin Series). Its approximate value is given by

$$e \approx 2.71828182846.$$

The Exponential Function

We now turn our attention to the function e^x. Note that the natural logarithm is one-to-one and therefore has an inverse function. For now, we denote this inverse function by $\exp x$. Then,

$$\exp(\ln x) = x$$

for $x > 0$ and

$$\ln(\exp x) = x$$

for all x.

The following figure shows the graphs of $\exp x$ and $\ln x$.

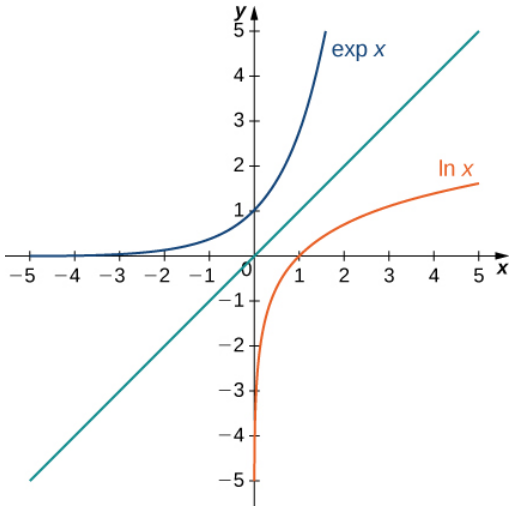


Figure 7.1.4: The graphs of $\ln x$ and $\exp x$.

We hypothesize that $\exp x = e^x$. For rational values of x, this is easy to show. If x is rational, then we have $\ln(e^x) = x \ln e = x$. Thus, when x is rational, $e^x = \exp x$. For irrational values of x, we simply define e^x as the inverse function of $\ln x$.

Definition

For any real number x, define $y = e^x$ to be the number for which

$$\ln y = \ln(e^x) = x.$$

Then we have $e^x = \exp x$ for all x, and thus

$$e^{\ln x} = x \text{ for } x > 0 \text{ and } \ln(e^x) = x$$

for all x.

Properties of the Exponential Function

Since the exponential function was defined in terms of an inverse function, and not in terms of a power of e we must verify that the usual laws of exponents hold for the function e^x.

Properties of the Exponential Function

If p and q are any real numbers and r is a rational number, then

i. $e^p e^q = e^{p+q}$
ii. $\dfrac{e^p}{e^q} = e^{p-q}$
iii. $(e^p)^r = e^{pr}$

Proof

Note that if p and q are rational, the properties hold. However, if p or q are irrational, we must apply the inverse function definition of e^x and verify the properties. Only the first property is verified here; the other two are left to you. We have

$$\ln(e^p e^q) = \ln(e^p) + \ln(e^q) = p + q = \ln(e^{p+q}).$$

Since $\ln x$ is one-to-one, then

$$e^p e^q = e^{p+q}.$$

□

As with part iv. of the logarithm properties, we can extend property iii. to irrational values of r, and we do so by the end of the section.

We also want to verify the differentiation formula for the function $y = e^x$. To do this, we need to use implicit differentiation. Let $y = e^x$. Then

$$\begin{aligned} \ln y &= x \\ \frac{d}{dx}\left(\ln y\right) &= \frac{d}{dx}\left(x\right) \\ \frac{1}{y}\frac{dy}{dx} &= 1 \\ \frac{dy}{dx} &= y. \end{aligned}$$

Thus, we see

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

as desired, which leads immediately to the integration formula

$$\int e^x \, dx = e^x + C.$$

We apply these formulas in the following examples.

Example 7.1.4: Using Properties of Exponential Functions

Evaluate the following derivatives:

a. $\frac{d}{dt}\left(e^{3t}e^{t^2}\right)$

b. $\frac{d}{dx}\left(e^{3x^2}\right)$

Solution

We apply the chain rule as necessary.

a. $\frac{d}{dt}\left(e^{3t}e^{t^2}\right) = \frac{d}{dt}\left(e^{3t+t^2}\right) = e^{3t+t^2}(3+2t)$

b. $\frac{d}{dx}\left(e^{3x^2}\right) = e^{3x^2}6x$

Exercise 7.1.4

Evaluate the following derivatives:

a. $\frac{d}{dx}\left(\frac{e^{x^2}}{e^{5x}}\right)$

b. $\dfrac{d}{dt}\left((e^{2t})^3\right)$

Hint

Use the properties of exponential functions and the chain rule as necessary.

Answer

a. $\dfrac{d}{dx}\left(\dfrac{e^{x^2}}{e^{5x}}\right) = e^{x^{2-5x}}(2x-5)$

b. $\dfrac{d}{dt}\left((e^{2t})^3\right) = 6e^{6t}$

Example 7.1.5: Using Properties of Exponential Functions

Evaluate the following integral: $\displaystyle\int 2xe^{-x^2}\,dx$.

Solution

Using u-substitution, let $u = -x^2$. Then $du = -2x\,dx$, and we have

$$\int 2xe^{-x^2}\,dx = -\int e^u\,du = -e^u + C = -e^{-x^2} + C.$$

Exercise 7.1.5

Evaluate the following integral: $\displaystyle\int \frac{4}{e^{3x}}\,dx$.

Hint

Use the properties of exponential functions and $u-substitution$ as necessary.

Answer

$$\int \frac{4}{e^{3x}}\,dx = -\frac{4}{3}e^{-3x} + C$$

General Logarithmic and Exponential Functions

We close this section by looking at exponential functions and logarithms with bases other than e. Exponential functions are functions of the form $f(x) = a^x$. Note that unless $a = e$, we still do not have a mathematically rigorous definition of these functions for irrational exponents. Let's rectify that here by defining the function $f(x) = a^x$ in terms of the exponential function e^x. We then examine logarithms with bases other than e as inverse functions of exponential functions.

Definition: Exponential Function

For any $a > 0$, and for any real number x, define $y = a^x$ as follows:

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

Now a^x is defined rigorously for all values of x. This definition also allows us to generalize property iv. of logarithms and property iii. of exponential functions to apply to both rational and irrational values of r. It is straightforward to show that properties of exponents hold for general exponential functions defined in this way.

Let's now apply this definition to calculate a differentiation formula for a^x. We have

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

The corresponding integration formula follows immediately.

Derivatives and Integrals Involving General Exponential Functions

Let $a > 0$. Then,

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

and

$$\int a^x \, dx = \frac{1}{\ln a} a^x + C.$$

If $a \neq 1$, then the function a^x is one-to-one and has a well-defined inverse. Its inverse is denoted by $\log_a x$. Then,

$$y = \log_a x \text{ if and only if } x = a^y.$$

Note that general logarithm functions can be written in terms of the natural logarithm. Let $y = \log_a x$. Then, $x = a^y$. Taking the natural logarithm of both sides of this second equation, we get

$$\begin{aligned} \ln x &= \ln(a^y) \\ \ln x &= y \ln a \\ y &= \frac{\ln x}{\ln a} \\ \log_a x &= \frac{\ln x}{\ln a}. \end{aligned}$$

Thus, we see that all logarithmic functions are constant multiples of one another. Next, we use this formula to find a differentiation formula for a logarithm with base a. Again, let $y = \log_a x$. Then,

$$\begin{aligned} \frac{dy}{dx} &= \frac{d}{dx}\left(\log_a x\right) \\ &= \frac{d}{dx}\left(\frac{\ln x}{\ln a}\right) \\ &= (\frac{1}{\ln a})\frac{d}{dx}\left(\ln x\right) \\ &= \frac{1}{\ln a} \cdot \frac{1}{x} = \frac{1}{x \ln a} \end{aligned}$$

Derivatives of General Logarithm Functions

Let $a > 0$. Then,

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

Example 7.1.6: Calculating Derivatives of General Exponential and Logarithm Functions

Evaluate the following derivatives:

a. $\frac{d}{dt}\left(4^t \cdot 2^{t^2}\right)$

b. $\frac{d}{dx}\left(\log_8(7x^2+4)\right)$

Solution: We need to apply the chain rule as necessary.

a. $\frac{d}{dt}\left(4^t \cdot 2^{t^2}\right) = \frac{d}{dt}\left(2^{2t} \cdot 2^{t^2}\right) = \frac{d}{dt}\left(2^{2t+t^2}\right) = 2^{2t+t^2} \ln(2)(2+2t)$

b. $\frac{d}{dx}\left(\log_8(7x^2+4)\right) = \frac{1}{(7x^2+4)(\ln 8)}(14x)$

? Exercise 7.1.6

Evaluate the following derivatives:

a. $\frac{d}{dt}\left(4^{t^4}\right)$

b. $\frac{d}{dx}\left(\log_3(\sqrt{x^2+1})\right)$

Hint

Use the formulas and apply the chain rule as necessary.

Answer

a. $\frac{d}{dt}\left(4^{t^4}\right) = 4^{t^4}(\ln 4)(4t^3)$

b. $\frac{d}{dx}\left(\log_3(\sqrt{x^2+1})\right) = \frac{x}{(\ln 3)(x^2+1)}$

✓ Example 7.1.7: Integrating General Exponential Functions

Evaluate the following integral: $\int \frac{3}{2^{3x}}\,dx$.

Solution

Use $u-substitution$ and let $u=-3x$. Then $du=-3\,dx$ and we have

$$\int \frac{3}{2^{3x}}\,dx = \int 3\cdot 2^{-3x}\,dx = -\int 2^u\,du = -\frac{1}{\ln 2}2^u + C = -\frac{1}{\ln 2}2^{-3x} + C.$$

? Exercise 7.1.7

Evaluate the following integral: $\int x^2 2^{x^3}\,dx$.

Hint

Use the properties of exponential functions and u-substitution

Answer

$$\int x^2 2^{x^3}\,dx = \frac{1}{3\ln 2}2^{x^3} + C$$

Key Concepts

- The earlier treatment of logarithms and exponential functions did not define the functions precisely and formally. This section develops the concepts in a mathematically rigorous way.
- The cornerstone of the development is the definition of the natural logarithm in terms of an integral.
- The function e^x is then defined as the inverse of the natural logarithm. General exponential functions are defined in terms of e^x, and the corresponding inverse functions are general logarithms.
- Familiar properties of logarithms and exponents still hold in this more rigorous context.

Key Equations

- **Natural logarithm function**
- $\ln x = \int_1^x \frac{1}{t}\, dt$
- **Exponential function** $y = e^x$
- $\ln y = \ln(e^x) = x$

7.1: The Logarithm Defined as an Integral is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **6.7: Integrals, Exponential Functions, and Logarithms** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

7.2: Exponential Change and Separable Differential Equations

7.2: Exponential Change and Separable Differential Equations is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.3: Hyperbolic Functions

The hyperbolic functions appear with some frequency in applications, and are quite similar in many respects to the trigonometric functions. This is a bit surprising given our initial definitions.

Definition 4.11.1: Hyperbolic Cosines and Sines

The **hyperbolic cosine** is the function

$$\cosh x = \frac{e^x + e^{-x}}{2}, \tag{7.3.1}$$

and the **hyperbolic sine** is the function

$$\sinh x = \frac{e^x - e^{-x}}{2}. \tag{7.3.2}$$

Notice that cosh is even (that is, $\cosh(-x) = \cosh(x)$) while sinh is odd ($\sinh(-x) = -\sinh(x)$), and $\cosh x + \sinh x = e^x$. Also, for all x, $\cosh x > 0$, while $\sinh x = 0$ if and only if $e^x - e^{-x} = 0$, which is true precisely when $x = 0$.

Lemma 4.11.2

The range of $\cosh x$ is $[1, \infty)$.

Proof

Let $y = \cosh x$. We solve for x:

$$\begin{aligned} y &= \frac{e^x + e^{-x}}{2} \\ 2y &= e^x + e^{-x} \\ 2ye^x &= e^{2x} + 1 \\ 0 &= e^{2x} - 2ye^x + 1 \\ e^x &= \frac{2y \pm \sqrt{4y^2 - 4}}{2} \\ e^x &= y \pm \sqrt{y^2 - 1} \end{aligned} \tag{7.3.3}$$

From the last equation, we see $y^2 \geq 1$, and since $y \geq 0$, it follows that $y \geq 1$.

Now suppose $y \geq 1$, so $y \pm \sqrt{y^2 - 1} > 0$. Then $x = \ln(y \pm \sqrt{y^2 - 1})$ is a real number, and $y = \cosh x$, so y is in the range of $\cosh(x)$.

□

Definition 4.11.3: Hyperbolic Tangent and Cotangent

The other hyperbolic functions are

$$\begin{aligned} \tanh x &= \frac{\sinh x}{\cosh x} \\ \coth x &= \frac{\cosh x}{\sinh x} \\ \operatorname{sech} x &= \frac{1}{\cosh x} \\ \operatorname{csch} x &= \frac{1}{\sinh x} \end{aligned} \tag{7.3.4}$$

The domain of coth and csch is $x \neq 0$ while the domain of the other hyperbolic functions is all real numbers. Graphs are shown in Figure 7.3.1

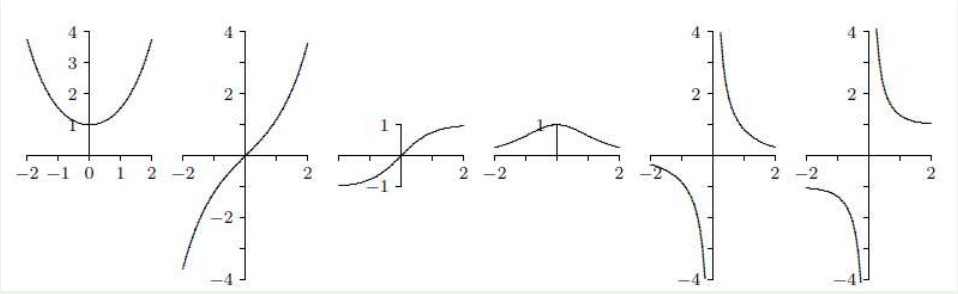


Figure 7.3.1: *The hyperbolic functions.*

Certainly the hyperbolic functions do not closely resemble the trigonometric functions graphically. But they do have analogous properties, beginning with the following identity.

Theorem 4.11.4

For all x in $\mathbb{R}$, $\cosh^2 x - \sinh^2 x = 1$.

Proof

The proof is a straightforward computation:

$$\cosh^2 x - \sinh^2 x = \frac{(e^x + e^{-x})^2}{4} - \frac{(e^x - e^{-x})^2}{4} = \frac{e^{2x} + 2 + e^{-2x} - e^{2x} + 2 - e^{-2x}}{4} = \frac{4}{4} = 1. \qquad (7.3.5)$$

□

This immediately gives two additional identities:

$$1 - \tanh^2 x = \operatorname{sech}^2 x \qquad \text{and} \qquad \coth^2 x - 1 = \operatorname{csch}^2 x. \qquad (7.3.6)$$

The identity of the theorem also helps to provide a geometric motivation. Recall that the graph of $x^2 - y^2 = 1$ is a hyperbola with asymptotes $x = \pm y$ whose x-intercepts are ± 1. If (x, y) is a point on the right half of the hyperbola, and if we let $x = \cosh t$, then $y = \pm\sqrt{x^2 - 1} = \pm\sqrt{\cosh^2 x - 1} = \pm \sinh t$. So for some suitable t, $\cosh t$ and $\sinh t$ are the coordinates of a typical point on the hyperbola. In fact, it turns out that t is twice the area shown in the first graph of Figure 7.3.2. Even this is analogous to trigonometry; $\cos t$ and $\sin t$ are the coordinates of a typical point on the unit circle, and t is twice the area shown in the second graph of Figure 7.3.2.

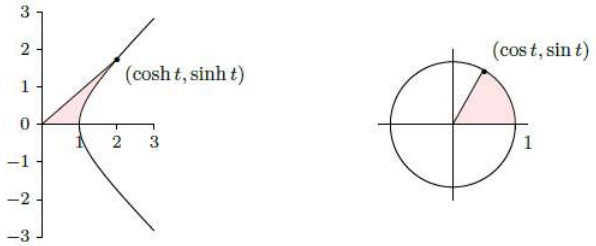


Figure 7.3.2: *Geometric definitions of sin, cos, sinh, cosh: t is twice the shaded area in each figure.*

Given the definitions of the hyperbolic functions, finding their derivatives is straightforward. Here again we see similarities to the trigonometric functions.

Theorem 4.11.5

$\frac{d}{dx}\cosh x = \sinh x$ and \thmrdef{thm:hyperbolic derivatives} $\frac{d}{dx}\sinh x = \cosh x$.

Proof

$$\frac{d}{dx}\cosh x = \frac{d}{dx}\frac{e^x + e^{-x}}{2} = \frac{e^x - e^{-x}}{2} = \sinh x, \tag{7.3.7}$$

and

$$\frac{d}{dx}\sinh x = \frac{d}{dx}\frac{e^x - e^{-x}}{2} = \frac{e^x + e^{-x}}{2} = \cosh x. \tag{7.3.8}$$

□

Since $\cosh x > 0$, $\sinh x$ is increasing and hence injective, so $\sinh x$ has an inverse, $\operatorname{arcsinh} x$. Also, $\sinh x > 0$ when $x > 0$, so $\cosh x$ is injective on $[0, \infty)$ and has a (partial) inverse, $\operatorname{arccosh} x$. The other hyperbolic functions have inverses as well, though $\operatorname{arcsech} x$ is only a partial inverse. We may compute the derivatives of these functions as we have other inverse functions.

Theorem 4.11.6

$\frac{d}{dx}\operatorname{arcsinh} x = \frac{1}{\sqrt{1+x^2}}$.

Proof

Let $y = \operatorname{arcsinh} x$, so $\sinh y = x$. Then

$$\frac{d}{dx}\sinh y = \cosh(y)\cdot y' = 1, \tag{7.3.9}$$

and so

$$y' = \frac{1}{\cosh y} = \frac{1}{\sqrt{1+\sinh^2 y}} = \frac{1}{\sqrt{1+x^2}}. \tag{7.3.10}$$

□

The other derivatives are left to the exercises.

Contributors

- David Guichard (Whitman College)
- Integrated by Justin Marshall.

7.3: Hyperbolic Functions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **4.11: Hyperbolic Functions** by David Guichard is licensed CC BY-NC-SA 4.0. Original source: https://www.whitman.edu/mathematics/multivariable/.

CHAPTER OVERVIEW

8: Techniques of Integration

8.1: Integration by Parts
8.2: Trigonometric Integrals
8.3: Trigonometric Substitutions
8.4: Integration of Rational Functions by Partial Fractions
8.5: Integral Tables and Computer Algebra Systems
8.6: Numerical Integration
8.7: Improper Integrals

8: Techniques of Integration is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.1: Integration by Parts

8.1: Integration by Parts is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.2: Trigonometric Integrals

8.2: Trigonometric Integrals is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.3: Trigonometric Substitutions

8.3: Trigonometric Substitutions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.4: Integration of Rational Functions by Partial Fractions

8.4: Integration of Rational Functions by Partial Fractions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.5: Integral Tables and Computer Algebra Systems

8.5: Integral Tables and Computer Algebra Systems is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.6: Numerical Integration

8.6: Numerical Integration is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.7: Improper Integrals

8.7: Improper Integrals is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

9: Infinite Sequence and Series

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

9.1: The Binomial Series and Applications of Taylor Series
9.2: Sequences
9.3: Infinite Series
9.4: The Integral Test
9.5: Comparison Tests
9.6: The Ratio and Root Tests
9.7: Alternating Series, Absolute and Conditional Convergence
9.8: Power Series
9.9: Taylor and Maclaurin Series
9.10: Convergence of Taylor Series

9: Infinite Sequence and Series is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.1: The Binomial Series and Applications of Taylor Series

9.1: The Binomial Series and Applications of Taylor Series is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.2: Sequences

9.2: Sequences is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.3: Infinite Series

9.3: Infinite Series is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.4: The Integral Test

9.4: The Integral Test is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.5: Comparison Tests

9.5: Comparison Tests is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.6: The Ratio and Root Tests

9.6: The Ratio and Root Tests is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.7: Alternating Series, Absolute and Conditional Convergence

9.7: Alternating Series, Absolute and Conditional Convergence is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.8: Power Series

9.8: Power Series is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.9: Taylor and Maclaurin Series

9.9: Taylor and Maclaurin Series is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.10: Convergence of Taylor Series

9.10: Convergence of Taylor Series is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

10: Parametric Equations and Polar Coordinates

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

10.1: Parametrizations of Plane Curves
10.2: Calculus with Parametric Curves
10.3: Polar Coordinates
10.4: Graphing in Polar Coordinates
10.5: Areas and Lengths in Polar Coordinates
10.6: Conics in Polar Coordinates

10: Parametric Equations and Polar Coordinates is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.1: Parametrizations of Plane Curves

Learning Objectives

- Plot a curve described by parametric equations.
- Convert the parametric equations of a curve into the form $y = f(x)$.
- Recognize the parametric equations of basic curves, such as a line and a circle.
- Recognize the parametric equations of a cycloid.

In this section we examine parametric equations and their graphs. In the two-dimensional coordinate system, parametric equations are useful for describing curves that are not necessarily functions. The parameter is an independent variable that both x and y depend on, and as the parameter increases, the values of x and y trace out a path along a plane curve. For example, if the parameter is t (a common choice), then t might represent time. Then x and y are defined as functions of time, and $(x(t), y(t))$ can describe the position in the plane of a given object as it moves along a curved path.

Parametric Equations and Their Graphs

Consider the orbit of Earth around the Sun. Our year lasts approximately 365.25 days, but for this discussion we will use 365 days. On January 1 of each year, the physical location of Earth with respect to the Sun is nearly the same, except for leap years, when the lag introduced by the extra $\frac{1}{4}$ day of orbiting time is built into the calendar. We call January 1 "day 1" of the year. Then, for example, day 31 is January 31, day 59 is February 28, and so on.

The number of the day in a year can be considered a variable that determines Earth's position in its orbit. As Earth revolves around the Sun, its physical location changes relative to the Sun. After one full year, we are back where we started, and a new year begins. According to Kepler's laws of planetary motion, the shape of the orbit is elliptical, with the Sun at one focus of the ellipse. We study this idea in more detail in Conic Sections.

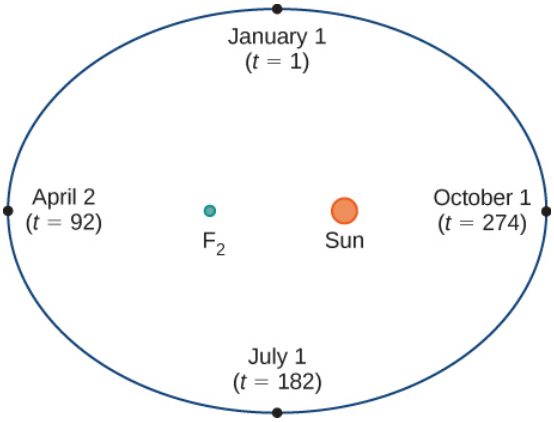


Figure 10.1.1: Earth's orbit around the Sun in one year.

Figure 10.1.1 depicts **Earth's orbit** around the Sun during one year. The point labeled F_2 is one of the foci of the ellipse; the other focus is occupied by the Sun. If we superimpose coordinate axes over this graph, then we can assign ordered pairs to each point on the ellipse (Figure 10.1.2). Then each x value on the graph is a value of position as a function of time, and each y value is also a value of position as a function of time. Therefore, each point on the graph corresponds to a value of Earth's position as a function of time.

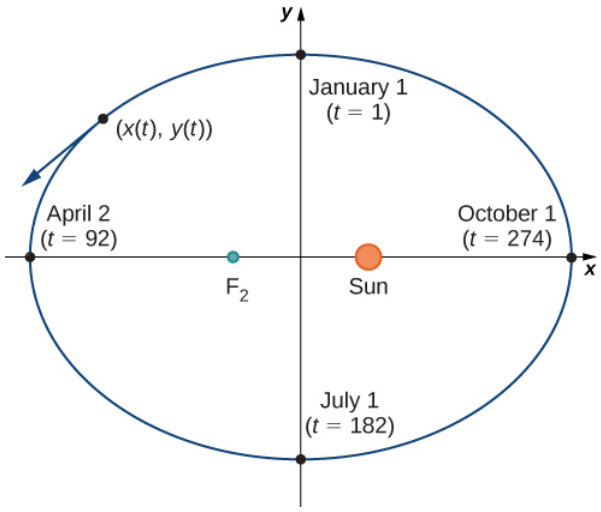


Figure 10.1.2: Coordinate axes superimposed on the orbit of Earth.

We can determine the functions for $x(t)$ and $y(t)$, thereby parameterizing the orbit of Earth around the Sun. The variable t is called an independent parameter and, in this context, represents time relative to the beginning of each year.

A curve in the (x, y) plane can be represented parametrically. The equations that are used to define the curve are called *parametric equations.*

Definition: Parametric Equations

If x and y are continuous functions of t on an interval I, then the equations

$$x = x(t)$$

and

$$y = y(t)$$

are called parametric equations and t is called the parameter. The set of points (x, y) obtained as t varies over the interval I is called the graph of the parametric equations. The graph of parametric equations is called a *parametric curve* or plane curve, and is denoted by C.

Notice in this definition that x and y are used in two ways. The first is as functions of the independent variable t. As t varies over the interval I, the functions $x(t)$ and $y(t)$ generate a set of ordered pairs (x, y). This set of ordered pairs generates the graph of the parametric equations. In this second usage, to designate the ordered pairs, x and y are variables. It is important to distinguish the variables x and y from the functions $x(t)$ and $y(t)$.

Example 10.1.1: Graphing a Parametrically Defined Curve

Sketch the curves described by the following parametric equations:

a. $x(t) = t - 1, \quad y(t) = 2t + 4, \quad \text{for } -3 \le t \le 2$
b. $x(t) = t^2 - 3, \quad y(t) = 2t + 1, \quad \text{for } -2 \le t \le 3$
c. $x(t) = 4\cos t, \quad y(t) = 4\sin t, \quad \text{for } 0 \le t \le 2\pi$

Solution a

To create a graph of this curve, first set up a table of values. Since the independent variable in both $x(t)$ and $y(t)$ is t, let t appear in the first column. Then $x(t)$ and $y(t)$ will appear in the second and third columns of the table.

t	$x(t)$	$y(t)$
−3	−4	−2
−2	−3	0

t	$x(t)$	$y(t)$
−1	−2	2
0	−1	4
1	0	6
2	1	8

The second and third columns in this table provide a set of points to be plotted. The graph of these points appears in Figure 10.1.3 The arrows on the graph indicate the **orientation** of the graph, that is, the direction that a point moves on the graph as t varies from −3 to 2.

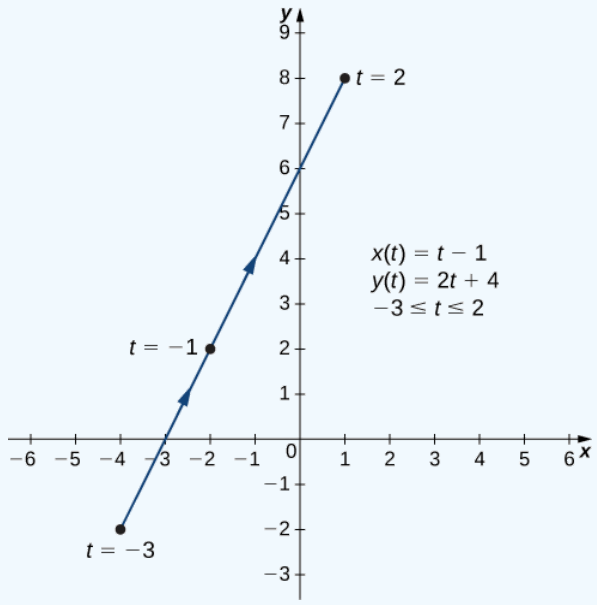


Figure 10.1.3: Graph of the plane curve described by the parametric equations in part a.

Solution b

To create a graph of this curve, again set up a table of values.

t	$x(t)$	$y(t)$
−2	1	−3
−1	−2	−1
0	−3	1
1	−2	3
2	1	5
3	6	7

The second and third columns in this table give a set of points to be plotted (Figure 10.1.4). The first point on the graph (corresponding to $t = -2$) has coordinates $(1, -3)$, and the last point (corresponding to $t = 3$) has coordinates $(6, 7)$. As t progresses from -2 to 3, the point on the curve travels along a parabola. The direction the point moves is again called the orientation and is indicated on the graph.

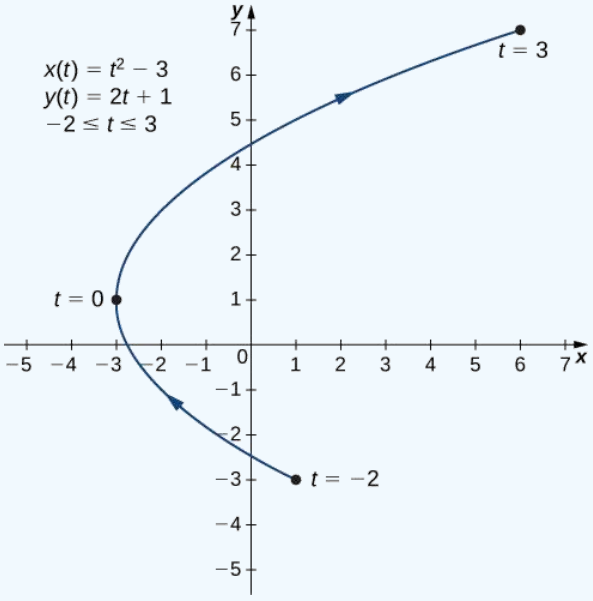


Figure 10.1.4: Graph of the plane curve described by the parametric equations in part b.

Solution c

In this case, use multiples of $\pi/6$ for t and create another table of values:

t	$x(t)$	$y(t)$	t	$x(t)$	$y(t)$
0	4	0	$\frac{7\pi}{6}$	$-2\sqrt{3} \approx -3.5$	-2
$\frac{\pi}{6}$	$2\sqrt{3} \approx 3.5$	2	$\frac{4\pi}{3}$	-2	$-2\sqrt{3} \approx -3.5$
$\frac{\pi}{3}$	2	$2\sqrt{3} \approx 3.5$	$\frac{3\pi}{2}$	0	-4
$\frac{\pi}{2}$	0	4	$\frac{5\pi}{3}$	2	$-2\sqrt{3} \approx -3.5$
$\frac{2\pi}{3}$	-2	$2\sqrt{3} \approx 3.5$	$\frac{11\pi}{6}$	$2\sqrt{3} \approx 3.5$	-2
$\frac{5\pi}{6}$	$-2\sqrt{3} \approx -3.5$	2	2π	4	0
π	-4	0			

The graph of this plane curve appears in the following graph.

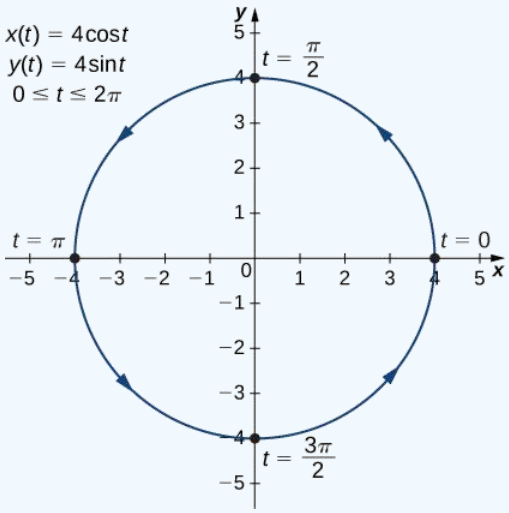


Figure 10.1.5: Graph of the plane curve described by the parametric equations in part c.

This is the graph of a circle with radius 4 centered at the origin, with a counterclockwise orientation. The starting point and ending points of the curve both have coordinates $(4, 0)$.

? Exercise 10.1.1

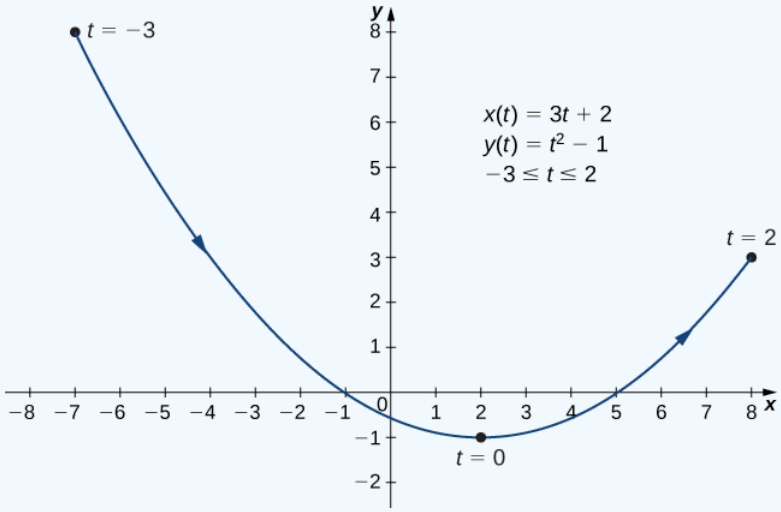
Sketch the curve described by the parametric equations

$$x(t) = 3t + 2, \quad y(t) = t^2 - 1, \quad \text{for } -3 \le t \le 2.$$

Hint

Make a table of values for $x(t)$ and $y(t)$ using t values from -3 to 2.

Answer



Eliminating the Parameter

To better understand the graph of a curve represented parametrically, it is useful to rewrite the two equations as a single equation relating the variables x and y. Then we can apply any previous knowledge of equations of curves in the plane to identify the curve. For example, the equations describing the plane curve in Example $10.1.1b$ are

$$x(t) = t^2 - 3 \tag{10.1.1}$$

$$y(t) = 2t + 1 \tag{10.1.2}$$

over the region $-2 \le t \le 3$.

Solving Equation 10.1.2 for t gives

$$t = \frac{y-1}{2}.$$

This can be substituted into Equation 10.1.1:

$$x = \left(\frac{y-1}{2}\right)^2 - 3 \tag{10.1.3}$$

$$= \frac{y^2 - 2y + 1}{4} - 3 \tag{10.1.4}$$

$$= \frac{y^2 - 2y - 11}{4}. \tag{10.1.5}$$

Equation 10.1.5 describes x as a function of y. These steps give an example of eliminating the parameter. The graph of this function is a parabola opening to the right (Figure 10.1.4). Recall that the plane curve started at $(1, -3)$ and ended at $(6, 7)$. These terminations were due to the restriction on the parameter t.

Example 10.1.2: Eliminating the Parameter

Eliminate the parameter for each of the plane curves described by the following parametric equations and describe the resulting graph.

a. $x(t) = \sqrt{2t+4}, \quad y(t) = 2t+1, \quad \text{for } -2 \le t \le 6$
b. $x(t) = 4\cos t, \quad y(t) = 3\sin t, \quad \text{for } 0 \le t \le 2\pi$

Solution a

To eliminate the parameter, we can solve either of the equations for t. For example, solving the first equation for t gives

$$\begin{aligned} x &= \sqrt{2t+4} \\ x^2 &= 2t+4 \\ x^2 - 4 &= 2t \\ t &= \frac{x^2-4}{2}. \end{aligned}$$

Note that when we square both sides it is important to observe that $x \ge 0$. Substituting $t = \dfrac{x^2-4}{2}$ into $y(t)$ yields

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

This is the equation of a parabola opening upward. There is, however, a domain restriction because of the limits on the parameter t. When $t = -2$, $x = \sqrt{2(-2)+4} = 0$, and when $t = 6$, $x = \sqrt{2(6)+4} = 4$. The graph of this plane curve follows.

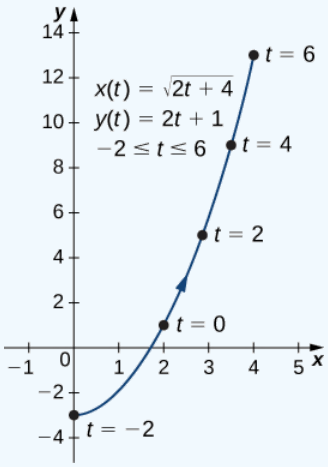


Figure 10.1.6: Graph of the plane curve described by the parametric equations in part a.

Solution b

Sometimes it is necessary to be a bit creative in eliminating the parameter. The parametric equations for this example are

$$x(t) = 4\cos t$$

and

$$y(t) = 3\sin t$$

Solving either equation for t directly is not advisable because sine and cosine are not one-to-one functions. However, dividing the first equation by 4 and the second equation by 3 (and suppressing the t) gives us

$$\cos t = \frac{x}{4}$$

and

$$\sin t = \frac{y}{3}.$$

Now use the Pythagorean identity $\cos^2 t + \sin^2 t = 1$ and replace the expressions for $\sin t$ and $\cos t$ with the equivalent expressions in terms of x and y. This gives

$$\left(\frac{x}{4}\right)^2 + \left(\frac{y}{3}\right)^2 = 1$$

$$\frac{x^2}{16} + \frac{y^2}{9} = 1.$$

This is the equation of a horizontal ellipse centered at the origin, with semi-major axis 4 and semi-minor axis 3 as shown in the following graph.

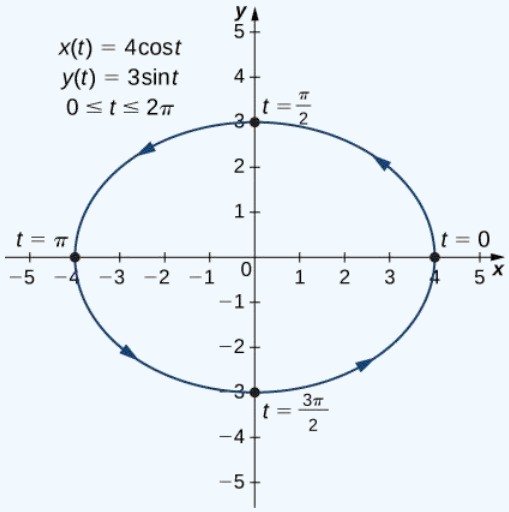


Figure 10.1.7: Graph of the plane curve described by the parametric equations in part b.

As t progresses from 0 to 2π, a point on the curve traverses the ellipse once, in a counterclockwise direction. Recall from the section opener that the orbit of Earth around the Sun is also elliptical. This is a perfect example of using parameterized curves to model a real-world phenomenon.

? Exercise 10.1.2

Eliminate the parameter for the plane curve defined by the following parametric equations and describe the resulting graph.

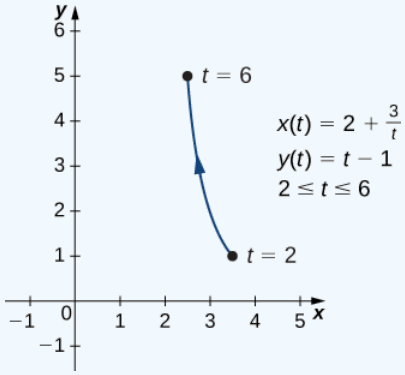
$$x(t) = 2 + \frac{3}{t}, \quad y(t) = t - 1, \quad \text{for } 2 \le t \le 6$$

Hint

Solve one of the equations for t and substitute into the other equation.

Answer

$x = 2 + \frac{3}{y+1}$, or $y = -1 + \frac{3}{x-2}$. This equation describes a portion of a rectangular hyperbola centered at $(2, -1)$.



So far we have seen the method of eliminating the parameter, assuming we know a set of parametric equations that describe a plane curve. What if we would like to start with the equation of a curve and determine a pair of parametric equations for that curve? This is certainly possible, and in fact it is possible to do so in many different ways for a given curve. The process is known as **parameterization of a curve.**

Example 10.1.3: Parameterizing a Curve

Find two different pairs of parametric equations to represent the graph of $y = 2x^2 - 3$.

Solution

First, it is always possible to parameterize a curve by defining $x(t) = t$, then replacing x with t in the equation for $y(t)$. This gives the parameterization

$$x(t) = t, \quad y(t) = 2t^2 - 3.$$

Since there is no restriction on the domain in the original graph, there is no restriction on the values of t.

We have complete freedom in the choice for the second parameterization. For example, we can choose $x(t) = 3t - 2$. The only thing we need to check is that there are no restrictions imposed on x; that is, the range of $x(t)$ is all real numbers. This is the case for $x(t) = 3t - 2$. Now since $y = 2x^2 - 3$, we can substitute $x(t) = 3t - 2$ for x. This gives

$$y(t) = 2(3t-2)^2 - 2 = 2(9t^2 - 12t + 4) - 2 = 18t^2 - 24t + 8 - 2 = 18t^2 - 24t + 6.$$

Therefore, a second parameterization of the curve can be written as

$$x(t) = 3t - 2 \text{ and } y(t) = 18t^2 - 24t + 6.$$

Exercise 10.1.3

Find two different sets of parametric equations to represent the graph of $y = x^2 + 2x$.

Hint

Follow the steps in Example 10.1.3 Remember we have freedom in choosing the parameterization for $x(t)$.

Answer

One possibility is $x(t) = t, \quad y(t) = t^2 + 2t.$ Another possibility is $x(t) = 2t - 3, \quad y(t) = (2t-3)^2 + 2(2t-3) = 4t^2 - 8t + 3.$ There are, in fact, an infinite number of possibilities.

Cycloids and Other Parametric Curves

Imagine going on a bicycle ride through the country. The tires stay in contact with the road and rotate in a predictable pattern. Now suppose a very determined ant is tired after a long day and wants to get home. So he hangs onto the side of the tire and gets a free ride. The path that this ant travels down a straight road is called a *cycloid* (Figure 10.1.8). A cycloid generated by a circle (or bicycle wheel) of radius a is given by the parametric equations

$$x(t) = a(t - \sin t), \quad y(t) = a(1 - \cos t).$$

To see why this is true, consider the path that the center of the wheel takes. The center moves along the x-axis at a constant height equal to the radius of the wheel. If the radius is a, then the coordinates of the center can be given by the equations

$$x(t) = at, \quad y(t) = a$$

for any value of t. Next, consider the ant, which rotates around the center along a circular path. If the bicycle is moving from left to right then the wheels are rotating in a clockwise direction. A possible parameterization of the circular motion of the ant (relative to the center of the wheel) is given by

$$x(t) = -a \sin t$$
$$y(t) = -a \cos t.$$

(The negative sign is needed to reverse the orientation of the curve. If the negative sign were not there, we would have to imagine the wheel rotating counterclockwise.) Adding these equations together gives the equations for the cycloid.

$$x(t) = a(t - \sin t)$$
$$y(t) = a(1 - \cos t)$$

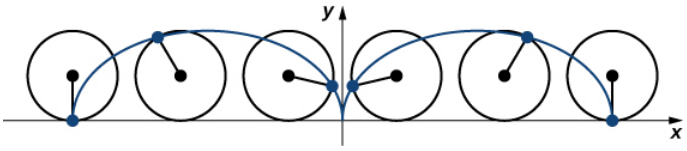


Figure 10.1.8: A wheel traveling along a road without slipping; the point on the edge of the wheel traces out a cycloid.

Now suppose that the bicycle wheel doesn't travel along a straight road but instead moves along the inside of a larger wheel, as in Figure 10.1.9. In this graph, the green circle is traveling around the blue circle in a counterclockwise direction. A point on the edge of the green circle traces out the red graph, which is called a *hypocycloid*.

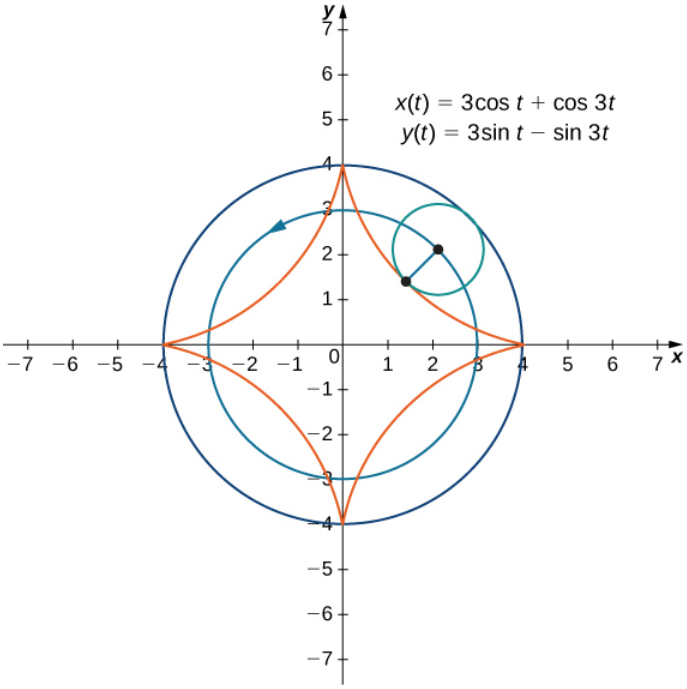


Figure 10.1.9: Graph of the hypocycloid described by the parametric equations shown.

The general parametric equations for a hypocycloid are

$$x(t) = (a - b)\cos t + b\cos\left(\frac{a - b}{b}\right)t$$

$$y(t) = (a-b)\sin t - b\sin(\frac{a-b}{b})t.$$

These equations are a bit more complicated, but the derivation is somewhat similar to the equations for the cycloid. In this case we assume the radius of the larger circle is a and the radius of the smaller circle is b. Then the center of the wheel travels along a circle of radius $a-b$. This fact explains the first term in each equation above. The period of the second trigonometric function in both $x(t)$ and $y(t)$ is equal to $\dfrac{2\pi b}{a-b}$.

The ratio $\dfrac{a}{b}$ is related to the number of **cusps** on the graph (cusps are the corners or pointed ends of the graph), as illustrated in Figure 10.1.10 This ratio can lead to some very interesting graphs, depending on whether or not the ratio is rational. Figure 10.1.9 corresponds to $a = 4$ and $b = 1$. The result is a hypocycloid with four cusps. Figure 10.1.10 shows some other possibilities. The last two hypocycloids have irrational values for $\dfrac{a}{b}$. In these cases the hypocycloids have an infinite number of cusps, so they never return to their starting point. These are examples of what are known as **space-filling curves.**

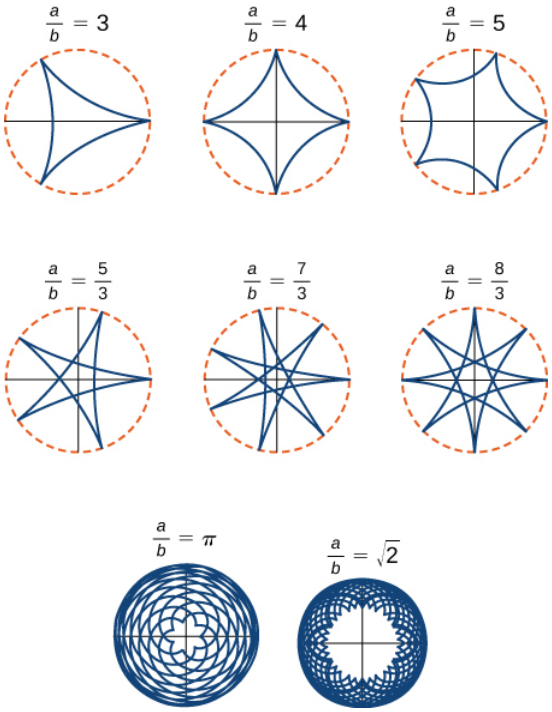


Figure 10.1.10: Graph of various hypocycloids corresponding to different values of a/b.

The Witch of Agnesi

Many plane curves in mathematics are named after the people who first investigated them, like the folium of Descartes or the spiral of Archimedes. However, perhaps the strangest name for a curve is the **witch of Agnesi.** Why a witch?

Maria Gaetana Agnesi (1718–1799) was one of the few recognized women mathematicians of eighteenth-century Italy. She wrote a popular book on analytic geometry, published in 1748, which included an interesting curve that had been studied by Fermat in 1630. The mathematician Guido Grandi showed in 1703 how to construct this curve, which he later called the "versoria," a Latin term for a rope used in sailing. Agnesi used the Italian term for this rope, "versiera," but in Latin, this same word means a "female goblin." When Agnesi's book was translated into English in 1801, the translator used the term "witch" for the curve, instead of rope. The name "witch of Agnesi" has stuck ever since.

The witch of Agnesi is a curve defined as follows: Start with a circle of radius a so that the points $(0,0)$ and $(0,2a)$ are points on the circle (Figure 10.1.11). Let O denote the origin. Choose any other point A on the circle, and draw the secant line OA. Let B denote the point at which the line OA intersects the horizontal line through $(0,2a)$. The vertical line through B intersects the horizontal line through A at the point P. As the point A varies, the path that the point P travels is the witch of Agnesi curve for the given circle.

Witch of Agnesi curves have applications in physics, including modeling water waves and distributions of spectral lines. In probability theory, the curve describes the probability density function of the *Cauchy distribution*. In this project you will parameterize these curves.

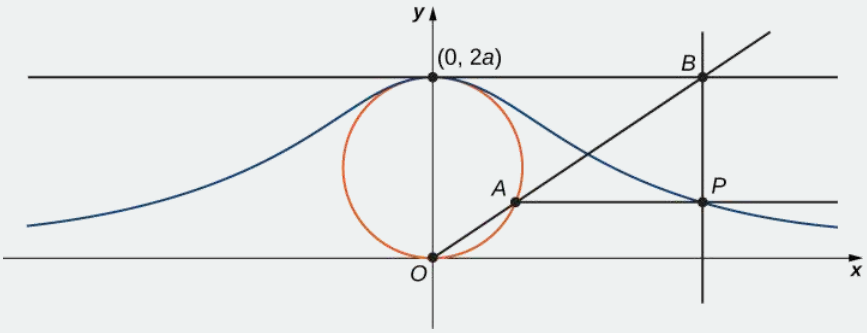


Figure 10.1.11: As the point A moves around the circle, the point P traces out the witch of Agnesi curve for the given circle.

1. On the figure, label the following points, lengths, and angle:
   a. C is the point on the x-axis with the same x-coordinate as A.
   b. x is the x-coordinate of P, and y is the y-coordinate of P.
   c. E is the point $(0, a)$.
   d. F is the point on the line segment OA such that the line segment EF is perpendicular to the line segment OA.
   e. b is the distance from O to F.
   f. c is the distance from F to A.
   g. d is the distance from O to B.
   h. θ is the measure of angle $\angle COA$.

The goal of this project is to parameterize the witch using θ as a parameter. To do this, write equations for x and y in terms of only θ.

2. Show that $d = \dfrac{2a}{\sin\theta}$.
3. Note that $x = d\cos\theta$. Show that $x = 2a\cot\theta$. When you do this, you will have parameterized the x-coordinate of the curve with respect to θ. If you can get a similar equation for y, you will have parameterized the curve.
4. In terms of θ, what is the angle $\angle EOA$?
5. Show that $b + c = 2a\cos\left(\frac{\pi}{2} - \theta\right)$.
6. Show that $y = 2a\cos\left(\frac{\pi}{2} - \theta\right)\sin\theta$.
7. Show that $y = 2a\sin^2\theta$. You have now parameterized the y-coordinate of the curve with respect to θ.
8. Conclude that a parameterization of the given witch curve is

$$x = 2a\cot\theta, \quad y = 2a\sin^2\theta, \quad \text{for } -\infty < \theta < \infty.$$

9. Use your parameterization to show that the given witch curve is the graph of the function $f(x) = \dfrac{8a^3}{x^2 + 4a^2}$.

Travels with My Ant: The Curtate and Prolate Cycloids

Earlier in this section, we looked at the parametric equations for a cycloid, which is the path a point on the edge of a wheel traces as the wheel rolls along a straight path. In this project we look at two different variations of the cycloid, called the curtate and prolate cycloids.

First, let's revisit the derivation of the parametric equations for a cycloid. Recall that we considered a tenacious ant trying to get home by hanging onto the edge of a bicycle tire. We have assumed the ant climbed onto the tire at the very edge, where the tire touches the ground. As the wheel rolls, the ant moves with the edge of the tire (Figure 10.1.12).

As we have discussed, we have a lot of flexibility when parameterizing a curve. In this case we let our parameter t represent the angle the tire has rotated through. Looking at Figure 10.1.12, we see that after the tire has rotated through an angle of t, the position of the center of the wheel, $C = (x_C, y_C)$, is given by

$$x_C = at \text{ and } y_C = a.$$

Furthermore, letting $A = (x_A, y_A)$ denote the position of the ant, we note that

$$x_C - x_A = a\sin t \text{ and } y_C - y_A = a\cos t$$

Then

$$x_A = x_C - a\sin t = at - a\sin t = a(t - \sin t)$$

$$y_A = y_C - a\cos t = a - a\cos t = a(1 - \cos t).$$

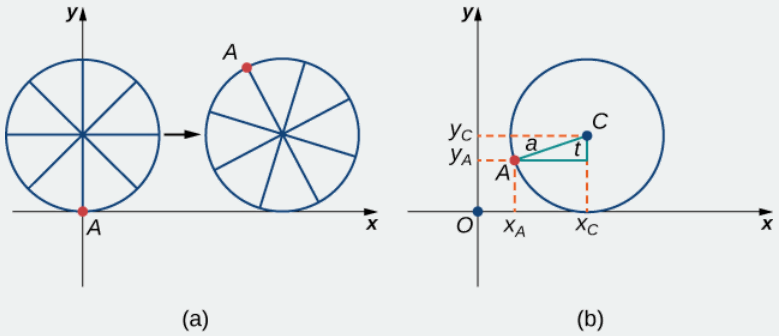


Figure 10.1.12: (a) The ant clings to the edge of the bicycle tire as the tire rolls along the ground. (b) Using geometry to determine the position of the ant after the tire has rotated through an angle of t.

Note that these are the same parametric representations we had before, but we have now assigned a physical meaning to the parametric variable t.

After a while the ant is getting dizzy from going round and round on the edge of the tire. So he climbs up one of the spokes toward the center of the wheel. By climbing toward the center of the wheel, the ant has changed his path of motion. The new path has less up-and-down motion and is called a **curtate cycloid** (Figure 10.1.13). As shown in the figure, we let b denote the distance along the spoke from the center of the wheel to the ant. As before, we let t represent the angle the tire has rotated through. Additionally, we let $C = (x_C, y_C)$ represent the position of the center of the wheel and $A = (x_A, y_A)$ represent the position of the ant.

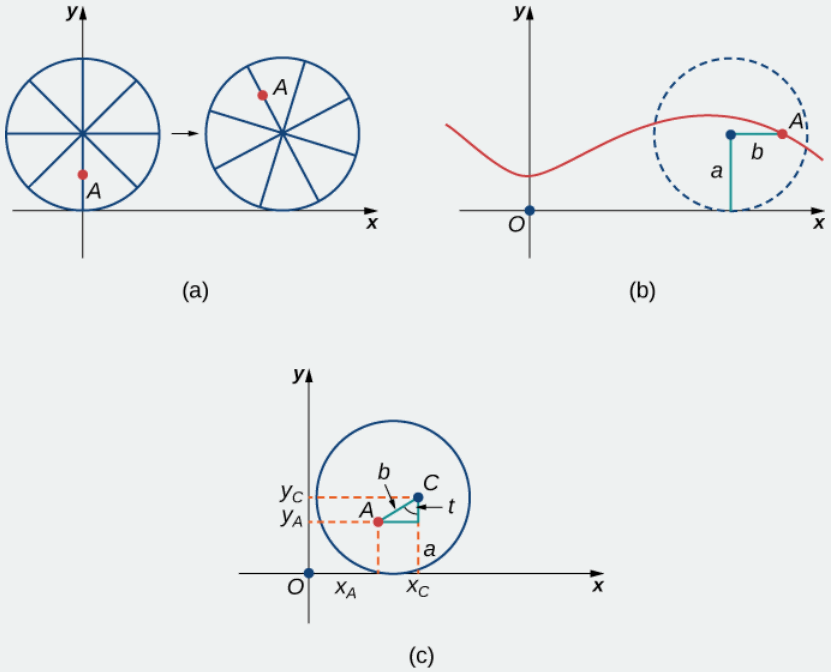


Figure 10.1.13: (a) The ant climbs up one of the spokes toward the center of the wheel. (b) The ant's path of motion after he climbs closer to the center of the wheel. This is called a curtate cycloid. (c) The new setup, now that the ant has moved closer to the center of the wheel.

1. What is the position of the center of the wheel after the tire has rotated through an angle of t?

2. Use geometry to find expressions for $x_C - x_A$ and for $y_C - y_A$.
3. On the basis of your answers to parts 1 and 2, what are the parametric equations representing the curtate cycloid?

Once the ant's head clears, he realizes that the bicyclist has made a turn, and is now traveling away from his home. So he drops off the bicycle tire and looks around. Fortunately, there is a set of train tracks nearby, headed back in the right direction. So the ant heads over to the train tracks to wait. After a while, a train goes by, heading in the right direction, and he manages to jump up and just catch the edge of the train wheel (without getting squished!).

The ant is still worried about getting dizzy, but the train wheel is slippery and has no spokes to climb, so he decides to just hang on to the edge of the wheel and hope for the best. Now, train wheels have a flange to keep the wheel running on the tracks. So, in this case, since the ant is hanging on to the very edge of the flange, the distance from the center of the wheel to the ant is actually greater than the radius of the wheel (Figure 10.1.14).

The setup here is essentially the same as when the ant climbed up the spoke on the bicycle wheel. We let b denote the distance from the center of the wheel to the ant, and we let t represent the angle the tire has rotated through. Additionally, we let $C = (x_C, y_C)$ represent the position of the center of the wheel and $A = (x_A, y_A)$ represent the position of the ant (Figure 10.1.14).

When the distance from the center of the wheel to the ant is greater than the radius of the wheel, his path of motion is called a prolate cycloid. A graph of a **prolate cycloid** is shown in the figure.

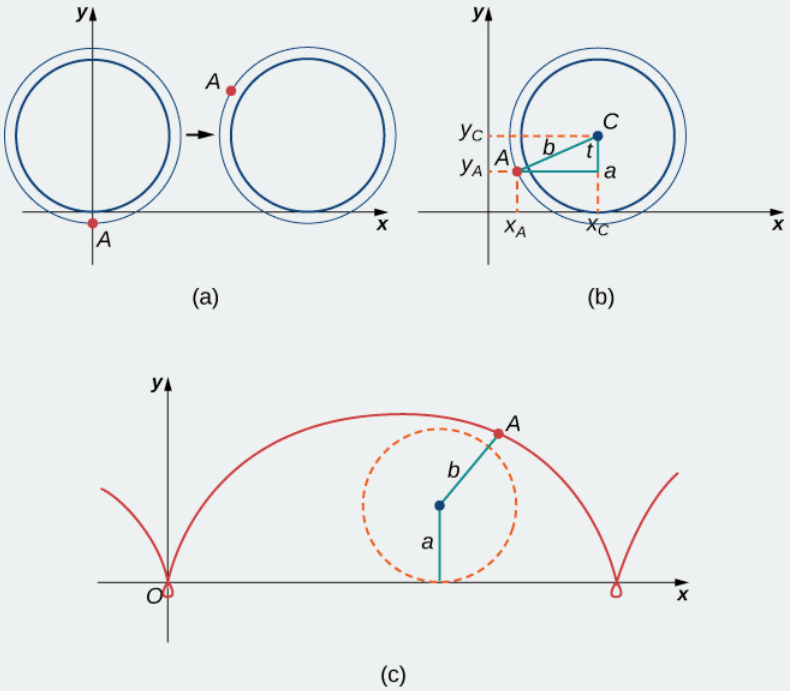


Figure 10.1.14: (a) The ant is hanging onto the flange of the train wheel. (b) The new setup, now that the ant has jumped onto the train wheel. (c) The ant travels along a prolate cycloid.

4. Using the same approach you used in parts 1– 3, find the parametric equations for the path of motion of the ant.
5. What do you notice about your answer to part 3 and your answer to part 4?

Notice that the ant is actually traveling backward at times (the "loops" in the graph), even though the train continues to move forward. He is probably going to be really dizzy by the time he gets home!

Key Concepts

- Parametric equations provide a convenient way to describe a curve. A parameter can represent time or some other meaningful quantity.
- It is often possible to eliminate the parameter in a parameterized curve to obtain a function or relation describing that curve.
- There is always more than one way to parameterize a curve.

- Parametric equations can describe complicated curves that are difficult or perhaps impossible to describe using rectangular coordinates.

Glossary

cycloid
the curve traced by a point on the rim of a circular wheel as the wheel rolls along a straight line without slippage

cusp
a pointed end or part where two curves meet

orientation
the direction that a point moves on a graph as the parameter increases

parameter
an independent variable that both x and y depend on in a parametric curve; usually represented by the variable t

parametric curve
the graph of the parametric equations $x(t)$ and $y(t)$ over an interval $a \leq t \leq b$ combined with the equations

parametric equations
the equations $x = x(t)$ and $y = y(t)$ that define a parametric curve

parameterization of a curve
rewriting the equation of a curve defined by a function $y = f(x)$ as parametric equations

---

10.1: Parametrizations of Plane Curves is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **11.1: Parametric Equations** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

10.2: Calculus with Parametric Curves

Learning Objectives

- Determine derivatives and equations of tangents for parametric curves.
- Find the area under a parametric curve.
- Use the equation for arc length of a parametric curve.
- Apply the formula for surface area to a volume generated by a parametric curve.

Now that we have introduced the concept of a parameterized curve, our next step is to learn how to work with this concept in the context of calculus. For example, if we know a parameterization of a given curve, is it possible to calculate the slope of a tangent line to the curve? How about the arc length of the curve? Or the area under the curve?

Another scenario: Suppose we would like to represent the location of a baseball after the ball leaves a pitcher's hand. If the position of the baseball is represented by the plane curve $(x(t), y(t))$ then we should be able to use calculus to find the speed of the ball at any given time. Furthermore, we should be able to calculate just how far that ball has traveled as a function of time.

Derivatives of Parametric Equations

We start by asking how to calculate the slope of a line tangent to a parametric curve at a point. Consider the plane curve defined by the parametric equations

$$x(t) = 2t + 3 \tag{10.2.1}$$
$$y(t) = 3t - 4 \tag{10.2.2}$$

within $-2 \leq t \leq 3$.

The graph of this curve appears in Figure 10.2.1. It is a line segment starting at $(-1, -10)$ and ending at $(9, 5)$.

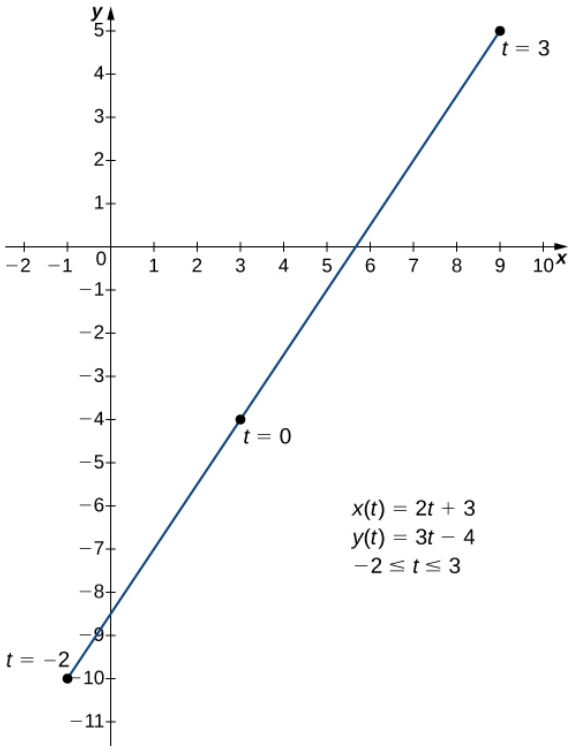


Figure 10.2.1: Graph of the line segment described by the given parametric equations.

We can eliminate the parameter by first solving Equation 10.2.1 for t:

$$x(t) = 2t + 3$$
$$x - 3 = 2t$$

$$t = \frac{x-3}{2}.$$

Substituting this into $y(t)$ (Equation 10.2.2), we obtain

$$y(t) = 3t - 4$$
$$y = 3\left(\frac{x-3}{2}\right) - 4$$
$$y = \frac{3x}{2} - \frac{9}{2} - 4$$
$$y = \frac{3x}{2} - \frac{17}{2}.$$

The slope of this line is given by $\frac{dy}{dx} = \frac{3}{2}$. Next we calculate $x'(t)$ and $y'(t)$. This gives $x'(t) = 2$ and $y'(t) = 3$. Notice that

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{3}{2}.$$

This is no coincidence, as outlined in the following theorem.

Derivative of Parametric Equations

Consider the plane curve defined by the parametric equations $x = x(t)$ and $y = y(t)$. Suppose that $x'(t)$ and $y'(t)$ exist, and assume that $x'(t) \neq 0$. Then the derivative $\frac{dy}{dx}$ is given by

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{y'(t)}{x'(t)}. \tag{10.2.3}$$

Proof

This theorem can be proven using the Chain Rule. In particular, assume that the parameter t can be eliminated, yielding a differentiable function $y = F(x)$. Then $y(t) = F(x(t))$. Differentiating both sides of this equation using the Chain Rule yields

$$y'(t) = F'\big(x(t)\big)x'(t),$$

so

$$F'\big(x(t)\big) = \frac{y'(t)}{x'(t)}.$$

But $F'\big(x(t)\big) = \frac{dy}{dx}$, which proves the theorem.

□

Equation 10.2.3 can be used to calculate derivatives of plane curves, as well as critical points. Recall that a critical point of a differentiable function $y = f(x)$ is any point $x = x_0$ such that either $f'(x_0) = 0$ or $f'(x_0)$ does not exist. Equation 10.2.3 gives a formula for the slope of a tangent line to a curve defined parametrically regardless of whether the curve can be described by a function $y = f(x)$ or not.

Example 10.2.1: Finding the Derivative of a Parametric Curve

Calculate the derivative $\frac{dy}{dx}$ for each of the following parametrically defined plane curves, and locate any critical points on their respective graphs.

a. $x(t) = t^2 - 3, \quad y(t) = 2t - 1, \quad \text{for } -3 \leq t \leq 4$
b. $x(t) = 2t + 1, \quad y(t) = t^3 - 3t + 4, \quad \text{for } -2 \leq t \leq 2$
c. $x(t) = 5\cos t, \quad y(t) = 5\sin t, \quad \text{for } 0 \leq t \leq 2\pi$

Solution

a. To apply Equation 10.2.3, first calculate $x'(t)$ and $y'(t)$:

$$x'(t) = 2t$$
$$y'(t) = 2.$$

Next substitute these into the equation:

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt}$$
$$\frac{dy}{dx} = \frac{2}{2t}$$
$$\frac{dy}{dx} = \frac{1}{t}.$$

This derivative is undefined when $t = 0$. Calculating $x(0)$ and $y(0)$ gives $x(0) = (0)^2 - 3 = -3$ and $y(0) = 2(0) - 1 = -1$, which corresponds to the point $(-3, -1)$ on the graph. The graph of this curve is a parabola opening to the right, and the point $(-3, -1)$ is its vertex as shown.

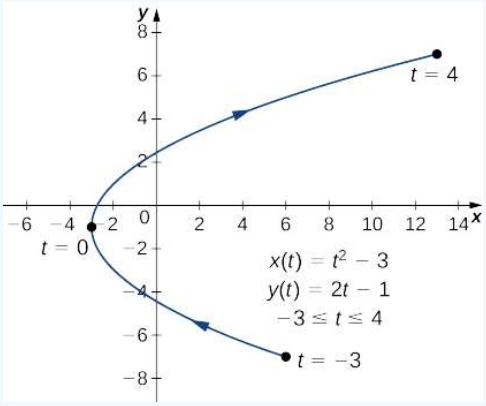


Figure 10.2.2: Graph of the parabola described by parametric equations in part a.

b. To apply Equation 10.2.3, first calculate $x'(t)$ and $y'(t)$:

$$x'(t) = 2$$
$$y'(t) = 3t^2 - 3\,.$$

Next substitute these into the equation:

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt}$$
$$\frac{dy}{dx} = \frac{3t^2 - 3}{2}.$$

This derivative is zero when $t = \pm 1$. When $t = -1$ we have

$$x(-1) = 2(-1) + 1 = -1 \text{ and } y(-1) = (-1)^3 - 3(-1) + 4 = -1 + 3 + 4 = 6\,,$$

which corresponds to the point $(-1, 6)$ on the graph. When $t = 1$ we have

$$x(1) = 2(1) + 1 = 3 \text{ and } y(1) = (1)^3 - 3(1) + 4 = 1 - 3 + 4 = 2,$$

which corresponds to the point $(3, 2)$ on the graph. The point $(3, 2)$ is a relative minimum and the point $(-1, 6)$ is a relative maximum, as seen in the following graph.

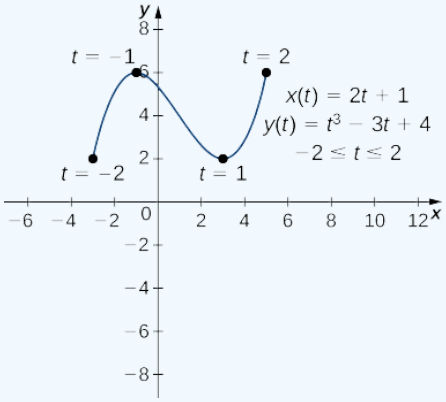


Figure 10.2.3: Graph of the curve described by parametric equations in part b.

c. To apply Equation 10.2.3, first calculate $x'(t)$ and $y'(t)$:

$$x'(t) = -5\sin t$$
$$y'(t) = 5\cos t.$$

Next substitute these into the equation:

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt}$$
$$\frac{dy}{dx} = \frac{5\cos t}{-5\sin t}$$
$$\frac{dy}{dx} = -\cot t.$$

This derivative is zero when $\cos t = 0$ and is undefined when $\sin t = 0$. This gives $t = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$, and 2π as critical points for t. Substituting each of these into $x(t)$ and $y(t)$, we obtain

t	$x(t)$	$y(t)$
0	5	0
$\frac{\pi}{2}$	0	5
π	−5	0
$\frac{3\pi}{2}$	0	−5
2π	5	0

These points correspond to the sides, top, and bottom of the circle that is represented by the parametric equations (Figure 10.2.4). On the left and right edges of the circle, the derivative is undefined, and on the top and bottom, the derivative equals zero.

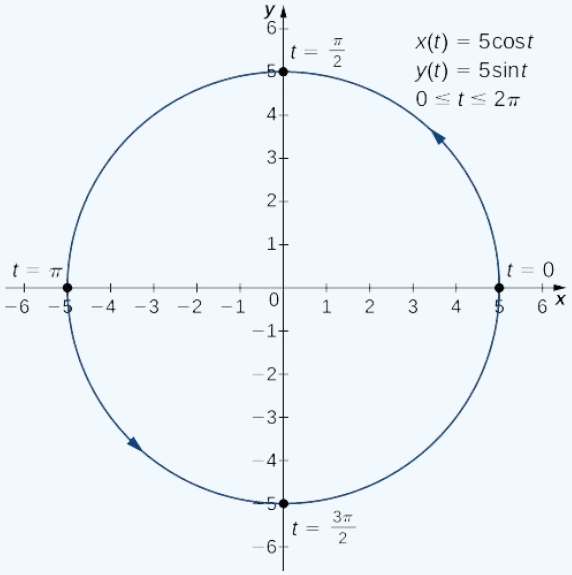


Figure 10.2.4: Graph of the curve described by parametric equations in part c.

Exercise 10.2.1

Calculate the derivative dy/dx for the plane curve defined by the equations

$$x(t) = t^2 - 4t, \quad y(t) = 2t^3 - 6t, \quad \text{for } -2 \le t \le 3$$

and locate any critical points on its graph.

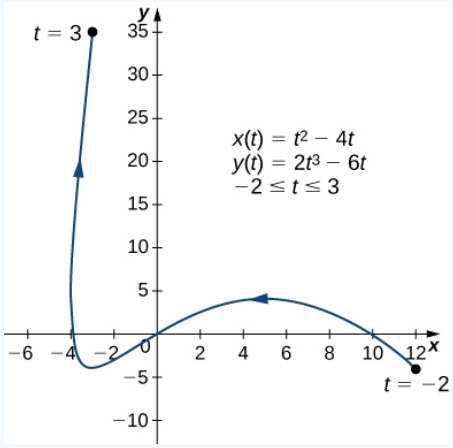
Hint

Calculate $x'(t)$ and $y'(t)$ and use Equation 10.2.3

Answer

$x'(t) = 2t - 4$ and $y'(t) = 6t^2 - 6$, so $\dfrac{dy}{dx} = \dfrac{6t^2 - 6}{2t - 4} = \dfrac{3t^2 - 3}{t - 2}$.

This expression is undefined when $t = 2$ and equal to zero when $t = \pm 1$.



Example 10.2.2: Finding a Tangent Line

Find the equation of the tangent line to the curve defined by the equations

$$x(t) = t^2 - 3, \quad y(t) = 2t - 1, \quad \text{for } -3 \le t \le 4$$

when $t = 2$.

Solution

First find the slope of the tangent line using Equation 10.2.3, which means calculating $x'(t)$ and $y'(t)$:

$$x'(t) = 2t$$
$$y'(t) = 2.$$

Next substitute these into the equation:

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt}$$
$$\frac{dy}{dx} = \frac{2}{2t}$$
$$\frac{dy}{dx} = \frac{1}{t}.$$

When $t = 2$, $\frac{dy}{dx} = \frac{1}{2}$, so this is the slope of the tangent line. Calculating $x(2)$ and $y(2)$ gives

$$x(2) = (2)^2 - 3 = 1 \text{ and } y(2) = 2(2) - 1 = 3,$$

which corresponds to the point $(1, 3)$ on the graph (Figure 10.2.5). Now use the point-slope form of the equation of a line to find the equation of the tangent line:

$$y - y_0 = m(x - x_0)$$
$$y - 3 = \frac{1}{2}(x - 1)$$
$$y - 3 = \frac{1}{2}x - \frac{1}{2}$$
$$y = \frac{1}{2}x + \frac{5}{2}.$$

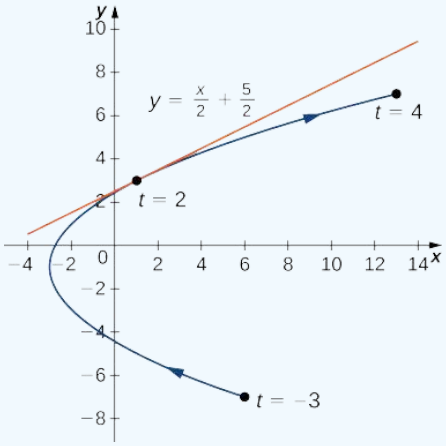


Figure 10.2.5: Tangent line to the parabola described by the given parametric equations when $t = 2$.

? Exercise 10.2.2

Find the equation of the tangent line to the curve defined by the equations

$$x(t) = t^2 - 4t, \quad y(t) = 2t^3 - 6t, \quad \text{for } -2 \le t \le 6 \quad \text{when } t = 5.$$

Hint

Calculate $x'(t)$ and $y'(t)$ and use Equation 10.2.3

Answer

The equation of the tangent line is $y = 24x + 100$.

Second-Order Derivatives

Our next goal is to see how to take the second derivative of a function defined parametrically. The second derivative of a function $y = f(x)$ is defined to be the derivative of the first derivative; that is,

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left[\frac{dy}{dx}\right]. \tag{10.2.4}$$

Since

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt},$$

we can replace the y on both sides of Equation 10.2.4with $\frac{dy}{dx}$. This gives us

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{(d/dt)(dy/dx)}{dx/dt}. \tag{10.2.5}$$

If we know dy/dx as a function of t, then this formula is straightforward to apply

Example 10.2.3: Finding a Second Derivative

Calculate the second derivative d^2y/dx^2 for the plane curve defined by the parametric equations $x(t) = t^2 - 3, \quad y(t) = 2t - 1, \quad \text{for } -3 \le t \le 4.$

Solution

From Example 10.2.1we know that $\frac{dy}{dx} = \frac{2}{2t} = \frac{1}{t}$. Using Equation 10.2.5, we obtain

$$\frac{d^2y}{dx^2} = \frac{(d/dt)(dy/dx)}{dx/dt} = \frac{(d/dt)(1/t)}{2t} = \frac{-t^{-2}}{2t} = -\frac{1}{2t^3}.$$

Exercise 10.2.3

Calculate the second derivative d^2y/dx^2 for the plane curve defined by the equations

$$x(t) = t^2 - 4t, \quad y(t) = 2t^3 - 6t, \quad \text{for } -2 \le t \le 3$$

and locate any critical points on its graph.

Hint

Start with the solution from the previous exercise, and use Equation 10.2.5

Answer

$\frac{d^2y}{dx^2} = \frac{3t^2 - 12t + 3}{2(t-2)^3}$. Critical points $(5, 4)$, $(-3, -4)$,and $(-4, 6)$.

Integrals Involving Parametric Equations

Now that we have seen how to calculate the derivative of a plane curve, the next question is this: How do we find the area under a curve defined parametrically? Recall the cycloid defined by these parametric equations

$$\begin{aligned} x(t) &= t - \sin t \\ y(t) &= 1 - \cos t. \end{aligned}$$

Suppose we want to find the area of the shaded region in the following graph.

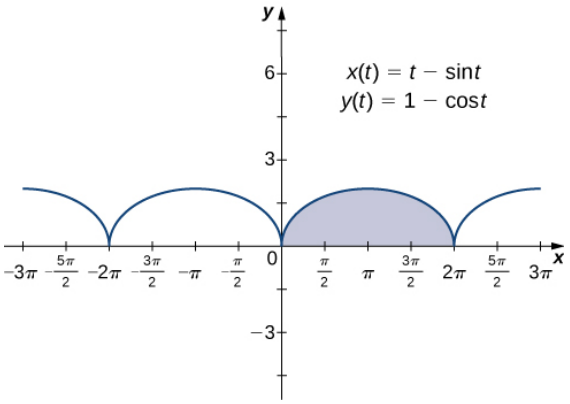


Figure 10.2.6: Graph of a cycloid with the arch over $[0, 2\pi]$ highlighted.

To derive a formula for the area under the curve defined by the functions

$$x = x(t)$$
$$y = y(t)$$

where $a \le t \le b$.

We assume that $x(t)$ is differentiable and start with an equal partition of the interval $a \le t \le b$. Suppose $t_0 = a < t_1 < t_2 < \cdots < t_n = b$ and consider the following graph.

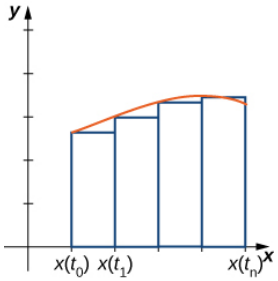


Figure 10.2.7: Approximating the area under a parametrically defined curve.

We use rectangles to approximate the area under the curve. The height of a typical rectangle in this parametrization is $y(x(\bar{t}_i))$ for some value $\bar{t}_i$ in the ith subinterval, and the width can be calculated as $x(t_i) - x(t_{i-1})$. Thus the area of the ith rectangle is given by

$$A_i = y(x(\bar{t}_i))(x(t_i) - x(t_{i-1})).$$

Then a Riemann sum for the area is

$$A_n = \sum_{i=1}^{n} y(x(\bar{t}_i))(x(t_i) - x(t_{i-1})).$$

Multiplying and dividing each area by $t_i - t_{i-1}$ gives

$$\begin{aligned} A_n &= \sum_{i=1}^{n} y(x(\bar{t}_i)) \left(\frac{x(t_i) - x(t_{i-1})}{t_i - t_{i-1}} \right) (t_i - t_{i-1}) \\ &= \sum_{i=1}^{n} y(x(\bar{t}_i)) \left(\frac{x(t_i) - x(t_{i-1})}{\Delta t} \right) \Delta t. \end{aligned}$$

Taking the limit as n approaches infinity gives

$$A = \lim_{n \to \infty} A_n = \int_a^b y(t) x'(t)\, dt.$$

This leads to the following theorem.

Area under a Parametric Curve

Consider the non-self-intersecting plane curve defined by the parametric equations

$$x = x(t), \quad y = y(t), \quad \text{for } a \le t \le b$$

and assume that $x(t)$ is differentiable. The area under this curve is given by

$$A = \int_a^b y(t)x'(t)\,dt. \tag{10.2.6}$$

Example 10.2.4: Finding the Area under a Parametric Curve

Find the area under the curve of the cycloid defined by the equations

$$x(t) = t - \sin t, \quad y(t) = 1 - \cos t, \quad \text{for } 0 \le t \le 2\pi.$$

Solution

Using Equation 10.2.6, we have

$$\begin{aligned} A &= \int_a^b y(t)x'(t)\,dt \\ &= \int_0^{2\pi} (1-\cos t)(1-\cos t)\,dt \\ &= \int_0^{2\pi} (1 - 2\cos t + \cos^2 t)\,dt \\ &= \int_0^{2\pi} \left(1 - 2\cos t + \frac{1+\cos(2t)}{2}\right) dt \\ &= \int_0^{2\pi} \left(\frac{3}{2} - 2\cos t + \frac{\cos(2t)}{2}\right) dt \\ &= \frac{3t}{2} - 2\sin t + \frac{\sin(2t)}{4}\Big|_0^{2\pi} \\ &= 3\pi \end{aligned}$$

Exercise 10.2.4

Find the area under the curve of the hypocycloid defined by the equations

$$x(t) = 3\cos t + \cos(3t), \quad y(t) = 3\sin t - \sin(3t), \quad \text{for } 0 \le t \le \pi.$$

Hint

Use Equation 10.2.6 along with the identities $\sin\alpha\sin\beta = \frac{1}{2}[\cos(\alpha-\beta) - \cos(\alpha+\beta)]$ and $\sin^2 t = \frac{1-\cos(2t)}{2}$.

Answer

$A = 3\pi$ (Note that the integral formula actually yields a negative answer. This is due to the fact that $x(t)$ is a decreasing function over the interval $[0, \pi]$; that is, the curve is traced from right to left.)

Arc Length of a Parametric Curve

In addition to finding the area under a parametric curve, we sometimes need to find the arc length of a parametric curve. In the case of a line segment, arc length is the same as the distance between the endpoints. If a particle travels from point A to point B along a curve, then the distance that particle travels is the arc length. To develop a formula for arc length, we start with an approximation by line segments as shown in the following graph.

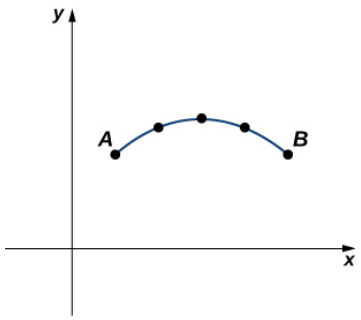


Figure 10.2.7: Approximation of a curve by line segments.

Given a plane curve defined by the functions $x = x(t), \quad y = y(t), \quad$ for $a \le t \le b$, we start by partitioning the interval $[a, b]$ into n equal subintervals: $t_0 = a < t_1 < t_2 < \cdots < t_n = b$. The width of each subinterval is given by $\Delta t = (b-a)/n$. We can calculate the length of each line segment:

$$d_1 = \sqrt{(x(t_1) - x(t_0))^2 + (y(t_1) - y(t_0))^2}$$

$$d_2 = \sqrt{(x(t_2) - x(t_1))^2 + (y(t_2) - y(t_1))^2}$$

etc.

Then add these up. We let s denote the exact arc length and s_n denote the approximation by n line segments:

$$s \approx \sum_{k=1}^{n} s_k = \sum_{k=1}^{n} \sqrt{(x(t_k) - x(t_{k-1}))^2 + (y(t_k) - y(t_{k-1}))^2} \, . \tag{10.2.7}$$

If we assume that $x(t)$ and $y(t)$ are differentiable functions of t, then the Mean Value Theorem applies, so in each subinterval $[t_{k-1}, t_k]$ there exist $\hat{t_k}$ and $\tilde{t_k}$ such that

$$x(t_k) - x(t_{k-1}) = x'(\hat{t_k})(t_k - t_{k-1}) = x'(\hat{t_k})\,\Delta t$$

$$y(t_k) - y(t_{k-1}) = y'(\tilde{t_k})(t_k - t_{k-1}) = y'(\tilde{t_k})\,\Delta t.$$

Therefore Equation 10.2.7 becomes

$$\begin{aligned} s \approx \sum_{k=1}^{n} s_k &= \sum_{k=1}^{n} \sqrt{(x'(\hat{t_k})\Delta t)^2 + (y'(\tilde{t_k})\Delta t)^2} \\ &= \sum_{k=1}^{n} \sqrt{(x'(\hat{t_k}))^2(\Delta t)^2 + (y'(\tilde{t_k}))^2(\Delta t)^2} \\ &= \sum_{k=1}^{n} \sqrt{(x'(\hat{t_k}))^2 + (y'(\tilde{t_k}))^2}\,\Delta t. \end{aligned}$$

This is a Riemann sum that approximates the arc length over a partition of the interval $[a, b]$. If we further assume that the derivatives are continuous and let the number of points in the partition increase without bound, the approximation approaches the exact arc length. This gives

$$\begin{aligned} s &= \lim_{n\to\infty} \sum_{k=1}^{n} s_k \\ &= \lim_{n\to\infty} \sum_{k=1}^{n} \sqrt{(x'(\hat{t_k}))^2 + (y'(\tilde{t_k}))^2}\,\Delta t \\ &= \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2}\,dt. \end{aligned}$$

When taking the limit, the values of $\hat{t_k}$ and $\tilde{t_k}$ are both contained within the same ever-shrinking interval of width Δt, so they must converge to the same value.

We can summarize this method in the following theorem.

Arc Length of a Parametric Curve

Consider the plane curve defined by the parametric equations

$$x = x(t), \quad y = y(t), \quad \text{for } t_1 \le t \le t_2$$

and assume that $x(t)$ and $y(t)$ are differentiable functions of t. Then the arc length of this curve is given by

$$s = \int_{t_1}^{t_2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt. \tag{10.2.8}$$

At this point a side derivation leads to a previous formula for arc length. In particular, suppose the parameter can be eliminated, leading to a function $y = F(x)$. Then $y(t) = F(x(t))$ and the Chain Rule gives

$$y'(t) = F'\big(x(t)\big)x'(t).$$

Substituting this into Equation 10.2.8 gives

$$\begin{aligned} s &= \int_{t_1}^{t_2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(F'(x)\frac{dx}{dt}\right)^2}\, dt \\ &= \int_{t_1}^{t_2} \sqrt{\left(\frac{dx}{dt}\right)^2 \left(1 + (F'(x))^2\right)}\, dt \\ &= \int_{t_1}^{t_2} x'(t)\sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dt. \end{aligned}$$

Here we have assumed that $x'(t) > 0$, which is a reasonable assumption. The Chain Rule gives $dx = x'(t)\, dt$, and letting $a = x(t_1)$ and $b = x(t_2)$ we obtain the formula

$$s = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx,$$

which is the formula for arc length obtained in the Introduction to the Applications of Integration.

Example 10.2.5: Finding the Arc Length of a Parametric Curve

Find the arc length of the semicircle defined by the equations

$$x(t) = 3\cos t, \quad y(t) = 3\sin t, \quad \text{for } 0 \le t \le \pi.$$

Solution

The values $t = 0$ to $t = \pi$ trace out the blue curve in Figure 10.2.8. To determine its length, use Equation 10.2.8:

$$\begin{aligned} s &= \int_{t_1}^{t_2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt \\ &= \int_0^{\pi} \sqrt{(-3\sin t)^2 + (3\cos t)^2}\, dt \\ &= \int_0^{\pi} \sqrt{9\sin^2 t + 9\cos^2 t}\, dt \\ &= \int_0^{\pi} \sqrt{9(\sin^2 t + \cos^2 t)}\, dt \\ &= \int_0^{\pi} 3\, dt = 3t\Big|_0^{\pi} \\ &= 3\pi \text{ units}. \end{aligned}$$

Note that the formula for the arc length of a semicircle is πr and the radius of this circle is 3. This is a great example of using calculus to derive a known formula of a geometric quantity.

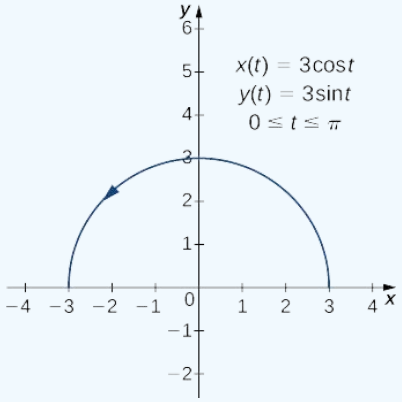


Figure 10.2.8: The arc length of the semicircle is equal to its radius times π.

? Exercise 10.2.5

Find the arc length of the curve defined by the equations

$$x(t) = 3t^2, \quad y(t) = 2t^3, \quad \text{for } 1 \le t \le 3.$$

Hint

Use Equation 10.2.8

Answer

$s = 2(10^{3/2} - 2^{3/2}) \approx 57.589$ units

We now return to the problem posed at the beginning of the section about a baseball leaving a pitcher's hand. Ignoring the effect of air resistance (unless it is a curve ball!), the ball travels a parabolic path. Assuming the pitcher's hand is at the origin and the ball travels left to right in the direction of the positive x-axis, the parametric equations for this curve can be written as

$$x(t) = 140t, \quad y(t) = -16t^2 + 2t$$

where t represents time. We first calculate the distance the ball travels as a function of time. This distance is represented by the arc length. We can modify the arc length formula slightly. First rewrite the functions $x(t)$ and $y(t)$ using v as an independent variable, so as to eliminate any confusion with the parameter t:

$$x(v) = 140v, \quad y(v) = -16v^2 + 2v.$$

Then we write the arc length formula as follows:

$$\begin{aligned} s(t) &= \int_0^t \sqrt{\left(\frac{dx}{dv}\right)^2 + \left(\frac{dy}{dv}\right)^2}\, dv \\ &= \int_0^t \sqrt{140^2 + (-32v+2)^2}\, dv \end{aligned}$$

The variable v acts as a dummy variable that disappears after integration, leaving the arc length as a function of time t. To integrate this expression we can use a formula from Appendix A,

$$\int \sqrt{a^2+u^2}\, du = \frac{u}{2}\sqrt{a^2+u^2} + \frac{a^2}{2}\ln\left|\, u + \sqrt{a^2+u^2}\,\right| + C.$$

We set $a = 140$ and $u = -32v + 2$. This gives $du = -32\, dv$, so $dv = -\dfrac{1}{32}\, du$. Therefore

$$\begin{aligned} \int \sqrt{140^2 + (-32v+2)^2}\, dv &= -\frac{1}{32}\int \sqrt{a^2+u^2}\, du \\ &= -\frac{1}{32}\left[\frac{(-32v+2)}{2}\sqrt{140^2 + (-32v+2)^2} + \frac{140^2}{2}\ln\left|\,(-32v+2) + \sqrt{140^2 + (-32v+2)^2}\,\right| + C\right] \end{aligned}$$

and

$$\begin{aligned} s(t) &= -\frac{1}{32}\left[\frac{(-32t+2)}{2}\sqrt{140^2+(-32t+2)^2}+\frac{140^2}{2}\ln\left|(-32t+2)+\sqrt{140^2+(-32t+2)^2}\right|\right] \\ &\quad +\frac{1}{32}\left[\sqrt{140^2+2^2}+\frac{140^2}{2}\ln\left|2+\sqrt{140^2+2^2}\right|\right] \\ &= \left(\frac{t}{2}-\frac{1}{32}\right)\sqrt{1024t^2-128t+19604}-\frac{1225}{4}\ln\left|(-32t+2)+\sqrt{1024t^2-128t+19604}\right|+\frac{\sqrt{19604}}{32}+\frac{1225}{4}\ln \\ &\quad (2+\sqrt{19604}) \end{aligned}$$

This function represents the distance traveled by the ball as a function of time. To calculate the speed, take the derivative of this function with respect to t. While this may seem like a daunting task, it is possible to obtain the answer directly from the Fundamental Theorem of Calculus:

$$\frac{d}{dx}\int_a^x f(u)\,du = f(x).$$

Therefore

$$\begin{aligned} s'(t) &= \frac{d}{dt}\left[s(t)\right] \\ &= \frac{d}{dt}\left[\int_0^t \sqrt{140^2+(-32v+2)^2}\,dv\right] \\ &= \sqrt{140^2+(-32t+2)^2} \\ &= \sqrt{1024t^2-128t+19604} \\ &= 2\sqrt{256t^2-32t+4901}. \end{aligned}$$

One third of a second after the ball leaves the pitcher's hand, the distance it travels is equal to

$$\begin{aligned} s\left(\frac{1}{3}\right) &= \left(\frac{1/3}{2}-\frac{1}{32}\right)\sqrt{1024\left(\frac{1}{3}\right)^2-128\left(\frac{1}{3}\right)+19604} \\ &\quad -\frac{1225}{4}\ln\left|\left(-32\left(\frac{1}{3}\right)+2\right)+\sqrt{1024\left(\frac{1}{3}\right)^2-128\left(\frac{1}{3}\right)+19604}\right| \\ &\quad +\frac{\sqrt{19604}}{32}+\frac{1225}{4}\ln(2+\sqrt{19604}) \\ &\approx 46.69 \text{ feet}. \end{aligned}$$

This value is just over three quarters of the way to home plate. The speed of the ball is

$$s'\left(\tfrac{1}{3}\right) = 2\sqrt{256\left(\tfrac{1}{3}\right)^2-32\left(\tfrac{1}{3}\right)+4901} \approx 140.27 \text{ ft/s}.$$

This speed translates to approximately 95 mph—a major-league fastball.

Surface Area Generated by a Parametric Curve

Recall the problem of finding the surface area of a volume of revolution. In Curve Length and Surface Area, we derived a formula for finding the surface area of a volume generated by a function $y=f(x)$ from $x=a$ to $x=b$, revolved around the x-axis:

$$S = 2\pi\int_a^b f(x)\sqrt{1+(f'(x))^2}\,dx.$$

We now consider a volume of revolution generated by revolving a parametrically defined curve $x=x(t), \quad y=y(t), \quad$ for $a\le t\le b$ around the x-axis as shown in Figure 10.2.9

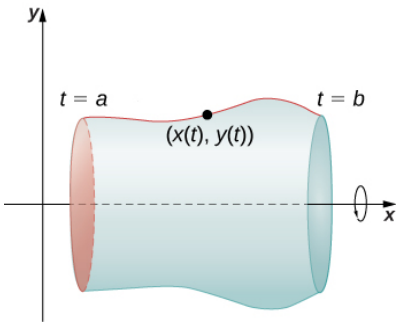


Figure 10.2.9: A surface of revolution generated by a parametrically defined curve.

The analogous formula for a parametrically defined curve is

$$S = 2\pi \int_a^b y(t)\sqrt{(x'(t))^2 + (y'(t))^2}\, dt \tag{10.2.9}$$

provided that $y(t)$ is not negative on $[a, b]$.

Example 10.2.6: Finding Surface Area

Find the surface area of a sphere of radius r centered at the origin.

Solution

We start with the curve defined by the equations

$$x(t) = r\cos t, \quad y(t) = r\sin t, \quad \text{for } 0 \le t \le \pi.$$

This generates an upper semicircle of radius r centered at the origin as shown in the following graph.

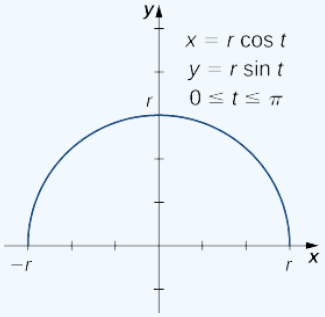


Figure 10.2.10: A semicircle generated by parametric equations.

When this curve is revolved around the x-axis, it generates a sphere of radius r. To calculate the surface area of the sphere, we use Equation 10.2.9:

$$
\begin{aligned}
S &= 2\pi \int_a^b y(t)\sqrt{(x'(t))^2 + (y'(t))^2}\, dt \\
&= 2\pi \int_0^\pi r\sin t \sqrt{(-r\sin t)^2 + (r\cos t)^2}\, dt \\
&= 2\pi \int_0^\pi r\sin t \sqrt{r^2\sin^2 t + r^2\cos^2 t}\, dt \\
&= 2\pi \int_0^\pi r\sin t \sqrt{r^2(\sin^2 t + \cos^2 t)}\, dt \\
&= 2\pi \int_0^\pi r^2 \sin t\, dt \\
&= 2\pi r^2 \left(-\cos t\Big|_0^\pi\right) \\
&= 2\pi r^2(-\cos\pi + \cos 0) \\
&= 4\pi r^2 \text{ units}^2.
\end{aligned}
$$

This is, in fact, the formula for the surface area of a sphere.

? Exercise 10.2.6

Find the surface area generated when the plane curve defined by the equations

$$x(t) = t^3, \quad y(t) = t^2, \quad \text{for } 0 \le t \le 1$$

is revolved around the x-axis.

Hint

Use Equation 10.2.9 When evaluating the integral, use a u-substitution.

Answer

$$A = \frac{\pi(494\sqrt{13} + 128)}{1215} \text{ units}^2$$

Key Concepts

- The derivative of the parametrically defined curve $x = x(t)$ and $y = y(t)$ can be calculated using the formula $\frac{dy}{dx} = \frac{y'(t)}{x'(t)}$. Using the derivative, we can find the equation of a tangent line to a parametric curve.
- The area between a parametric curve and the x-axis can be determined by using the formula $A = \int_{t_1}^{t_2} y(t)x'(t)\, dt$.
- The arc length of a parametric curve can be calculated by using the formula

$$s = \int_{t_1}^{t_2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt.$$

- The surface area of a volume of revolution revolved around the x-axis is given by

$$S = 2\pi \int_a^b y(t)\sqrt{(x'(t))^2 + (y'(t))^2}\, dt.$$

- If the curve is revolved around the y-axis, then the formula is

$$S = 2\pi \int_a^b x(t)\sqrt{(x'(t))^2 + (y'(t))^2}\, dt.$$

Key Equations

- **Derivative of parametric equations**

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{y'(t)}{x'(t)}$$

- **Second-order derivative of parametric equations**

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{(d/dt)(dy/dx)}{dx/dt}$$

- **Area under a parametric curve**

$$A = \int_a^b y(t)x'(t)\,dt$$

- **Arc length of a parametric curve**

$$s = \int_{t_1}^{t_2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\,dt$$

- **Surface area generated by a parametric curve**

$$S = 2\pi \int_a^b y(t)\sqrt{(x'(t))^2 + (y'(t))^2}\,dt$$

10.2: Calculus with Parametric Curves is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **11.2: Calculus of Parametric Curves** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

10.3: Polar Coordinates

Learning Objectives

- Locate points in a plane by using polar coordinates.
- Convert points between rectangular and polar coordinates.
- Sketch polar curves from given equations.
- Convert equations between rectangular and polar coordinates.
- Identify symmetry in polar curves and equations.

The rectangular coordinate system (or Cartesian plane) provides a means of mapping points to ordered pairs and ordered pairs to points. This is called a *one-to-one mapping* from points in the plane to ordered pairs. The polar coordinate system provides an alternative method of mapping points to ordered pairs. In this section we see that in some circumstances, polar coordinates can be more useful than rectangular coordinates.

Defining Polar Coordinates

To find the coordinates of a point in the polar coordinate system, consider Figure 10.3.1. The point P has Cartesian coordinates (x, y). The line segment connecting the origin to the point P measures the distance from the origin to P and has length r. The angle between the positive x-axis and the line segment has measure θ. This observation suggests a natural correspondence between the coordinate pair (x, y) and the values r and θ. This correspondence is the basis of the **polar coordinate system**. Note that every point in the Cartesian plane has two values (hence the term ordered pair) associated with it. In the polar coordinate system, each point also has two values associated with it: r and θ.

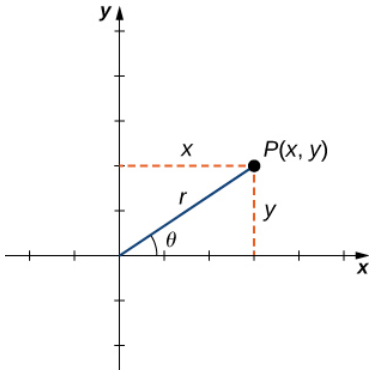


Figure 10.3.1: An arbitrary point in the Cartesian plane.

Using right-triangle trigonometry, the following equations are true for the point P:

$$\cos\theta = \frac{x}{r} \text{ so } x = r\cos\theta$$

$$\sin\theta = \frac{y}{r} \text{ so } y = r\sin\theta.$$

Furthermore,

$$r^2 = x^2 + y^2$$

and

$$\tan\theta = \frac{y}{x}.$$

Each point (x, y) in the Cartesian coordinate system can therefore be represented as an ordered pair (r, θ) in the polar coordinate system. The first coordinate is called the **radial coordinate** and the second coordinate is called the **angular coordinate**. Every point in the plane can be represented in this form.

Note that the equation $\tan\theta = y/x$ has an infinite number of solutions for any ordered pair (x, y). However, if we restrict the solutions to values between 0 and 2π then we can assign a unique solution to the quadrant in which the original point (x, y) is

located. Then the corresponding value of r is positive, so $r^2 = x^2 + y^2$.

Converting Points between Coordinate Systems

Given a point P in the plane with Cartesian coordinates (x, y) and polar coordinates (r, θ), the following conversion formulas hold true:

$$x = r\cos\theta \tag{10.3.1}$$

$$y = r\sin\theta \tag{10.3.2}$$

and

$$r^2 = x^2 + y^2 \tag{10.3.3}$$

$$\tan\theta = \frac{y}{x} \tag{10.3.4}$$

These formulas can be used to convert from rectangular to polar or from polar to rectangular coordinates. Notice that Equation 10.3.3 is the *Pythagorean theorem*. (Figure 10.3.1).

Example 10.3.1: Converting between Rectangular and Polar Coordinates

Convert each of the following points into polar coordinates.

a. $(1, 1)$
b. $(-3, 4)$
c. $(0, 3)$
d. $(5\sqrt{3}, -5)$

Convert each of the following points into rectangular coordinates.

e. $(3, \pi/3)$
f. $(2, 3\pi/2)$
g. $(6, -5\pi/6)$

Solution

a. Use $x = 1$ and $y = 1$ in Equation 10.3.3:

$$\begin{aligned} r^2 &= x^2 + y^2 \\ &= 1^2 + 1^2 \\ r &= \sqrt{2} \end{aligned}$$

and via Equation 10.3.4

$$\begin{aligned} \tan\theta &= \frac{y}{x} = \frac{1}{1} = 1 \\ \theta &= \frac{\pi}{4}. \end{aligned}$$

Therefore this point can be represented as $(\sqrt{2}, \frac{\pi}{4})$ in polar coordinates.

b. Use $x = -3$ and $y = 4$ in Equation 10.3.3:

$$\begin{aligned} r^2 &= x^2 + y^2 = (-3)^2 + (4)^2 \\ r &= 5 \end{aligned}$$

and via Equation 10.3.4

$$\tan\theta = \frac{y}{x} = -\frac{4}{3}$$

$$\theta = \arctan(-\frac{4}{3}) + \pi \approx 2.21.$$

Therefore this point can be represented as $(5, 2.21)$ in polar coordinates.

c. Use $x = 0$ and $y = 3$ in Equation 10.3.3:

$$r^2 = x^2 + y^2 = (3)^2 + (0)^2 = 9 + 0 \quad r = 3$$

and via Equation 10.3.4

$$\tan\theta = \frac{y}{x} = \frac{3}{0}\ .$$

Direct application of the second equation leads to division by zero. Graphing the point $(0, 3)$ on the rectangular coordinate system reveals that the point is located on the positive y-axis. The angle between the positive x-axis and the positive y-axis is $\frac{\pi}{2}$. Therefore this point can be represented as $(3, \frac{\pi}{2})$ in polar coordinates.

d. Use $x = 5\sqrt{3}$ and $y = -5$ in Equation 10.3.3:

$$r^2 = x^2 + y^2 = (5\sqrt{3})^2 + (-5)^2 = 75 + 25$$

$$r = 10$$

and via Equation 10.3.4

$$\tan\theta = \frac{y}{x} = \frac{-5}{5\sqrt{3}} = -\frac{\sqrt{3}}{3}$$

$$\theta = -\frac{\pi}{6}\ .$$

Therefore this point can be represented as $(10, -\frac{\pi}{6})$ in polar coordinates.

e. Use $r = 3$ and $\theta = \frac{\pi}{3}$ in Equation 10.3.1:

$$x = r\cos\theta = 3\cos(\frac{\pi}{3}) = 3(\frac{1}{2}) = \frac{3}{2}$$

and

$$y = r\sin\theta = 3\sin(\frac{\pi}{3}) = 3(\frac{\sqrt{3}}{2}) = \frac{3\sqrt{3}}{2}\ .$$

Therefore this point can be represented as $(\frac{3}{2}, \frac{3\sqrt{3}}{2})$ in rectangular coordinates.

f. Use $r = 2$ and $\theta = \frac{3\pi}{2}$ in Equation 10.3.1:

$$x = r\cos\theta = 2\cos(\frac{3\pi}{2}) = 2(0) = 0$$

and

$$y = r\sin\theta = 2\sin(\frac{3\pi}{2}) = 2(-1) = -2.$$

Therefore this point can be represented as $(0, -2)$ in rectangular coordinates.

g. Use $r = 6$ and $\theta = -\frac{5\pi}{6}$ in Equation 10.3.1:

$$x = r\cos\theta = 6\cos(-\frac{5\pi}{6}) = 6(-\frac{\sqrt{3}}{2}) = -3\sqrt{3}$$

and

$$y = r\sin\theta = 6\sin(-\frac{5\pi}{6}) = 6(-\frac{1}{2}) = -3\ .$$

Therefore this point can be represented as $(-3\sqrt{3}, -3)$ in rectangular coordinates.

Exercise 10.3.1

Convert $(-8, -8)$ into polar coordinates and $(4, \frac{2\pi}{3})$ into rectangular coordinates.

Hint

Use Equation 10.3.3and Equation 10.3.1. Make sure to check the quadrant when calculating θ.

Answer

$(8\sqrt{2}, \frac{5\pi}{4})$ and $(-2, 2\sqrt{3})$

The polar representation of a point is not unique. For example, the polar coordinates $(2, \frac{\pi}{3})$ and $(2, \frac{7\pi}{3})$ both represent the point $(1, \sqrt{3})$ in the rectangular system. Also, the value of r can be negative. Therefore, the point with polar coordinates $(-2, \frac{4\pi}{3})$ also represents the point $(1, \sqrt{3})$ in the rectangular system, as we can see by using Equation 10.3.1:

$$x = r\cos\theta = -2\cos(\frac{4\pi}{3}) = -2(-\frac{1}{2}) = 1$$

and

$$y = r\sin\theta = -2\sin(\frac{4\pi}{3}) = -2(-\frac{\sqrt{3}}{2}) = \sqrt{3}.$$

Every point in the plane has an infinite number of representations in polar coordinates. However, each point in the plane has only one representation in the rectangular coordinate system.

Note that the polar representation of a point in the plane also has a visual interpretation. In particular, r is the directed distance that the point lies from the origin, and θ measures the angle that the line segment from the origin to the point makes with the positive x-axis. Positive angles are measured in a counterclockwise direction and negative angles are measured in a clockwise direction. The polar coordinate system appears in Figure 10.3.2

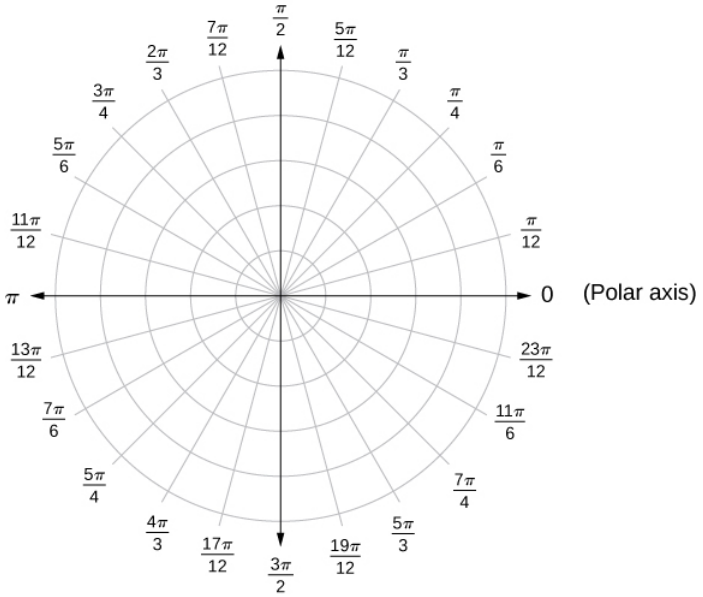


Figure 10.3.2: The polar coordinate system.

The line segment starting from the center of the graph going to the right (called the positive x-axis in the Cartesian system) is the **polar axis.** The center point is the **pole**, or origin, of the coordinate system, and corresponds to $r = 0$. The innermost circle shown in Figure 10.3.2contains all points a distance of 1 unit from the pole, and is represented by the equation $r = 1$. Then $r = 2$ is the

set of points 2 units from the pole, and so on. The line segments emanating from the pole correspond to fixed angles. To plot a point in the polar coordinate system, start with the angle. If the angle is positive, then measure the angle from the polar axis in a counterclockwise direction. If it is negative, then measure it clockwise. If the value of r is positive, move that distance along the terminal ray of the angle. If it is negative, move along the ray that is opposite the terminal ray of the given angle.

Example 10.3.2: Plotting Points in the Polar Plane

Plot each of the following points on the polar plane.

a. $(2, \frac{\pi}{4})$

b. $(-3, \frac{2\pi}{3})$

c. $(4, \frac{5\pi}{4})$

Solution

The three points are plotted in Figure $10.3.3$

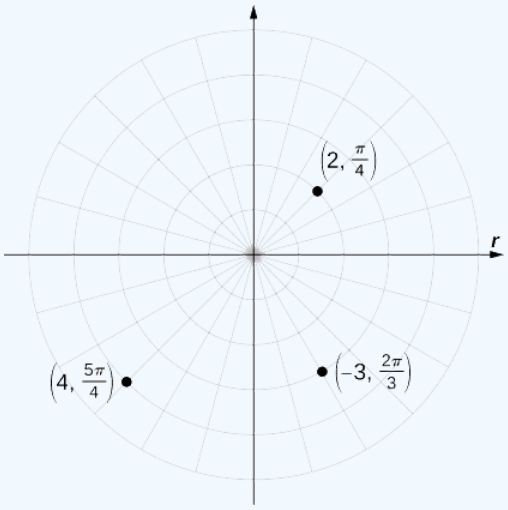


Figure 10.3.3: Three points plotted in the polar coordinate system.

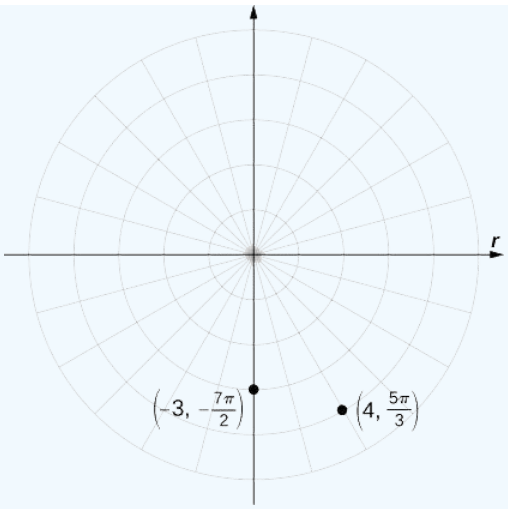
Exercise 10.3.2

Plot $(4, \frac{5\pi}{3})$ and $(-3, -\frac{7\pi}{2})$ on the polar plane.

Hint

Start with θ, then use r.

Answer



Polar Curves

Now that we know how to plot points in the polar coordinate system, we can discuss how to plot curves. In the rectangular coordinate system, we can graph a function $y = f(x)$ and create a curve in the Cartesian plane. In a similar fashion, we can graph a curve that is generated by a function $r = f(\theta)$.

The general idea behind graphing a function in polar coordinates is the same as graphing a function in rectangular coordinates. Start with a list of values for the independent variable (θ in this case) and calculate the corresponding values of the dependent variable r. This process generates a list of ordered pairs, which can be plotted in the polar coordinate system. Finally, connect the points, and take advantage of any patterns that may appear. The function may be periodic, for example, which indicates that only a limited number of values for the independent variable are needed.

Problem-Solving Strategy: Plotting a Curve in Polar Coordinates

1. Create a table with two columns. The first column is for θ, and the second column is for r.
2. Create a list of values for θ.
3. Calculate the corresponding r values for each θ.
4. Plot each ordered pair (r, θ) on the coordinate axes.
5. Connect the points and look for a pattern.

Example 10.3.3: Graphing a Function in Polar Coordinates

Graph the curve defined by the function $r = 4\sin\theta$. Identify the curve and rewrite the equation in rectangular coordinates.

Solution

Because the function is a multiple of a sine function, it is periodic with period 2π, so use values for θ between 0 and 2π. The result of steps 1–3 appear in the following table. Figure $10.3.4$shows the graph based on this table.

Table 10.3.1: Table of values for this polar function.

θ	$r = 4\sin\theta$	θ	$r = 4\sin\theta$
0	0	π	0
$\frac{\pi}{6}$	2	$\frac{7\pi}{6}$	-2
$\frac{\pi}{4}$	$2\sqrt{2} \approx 2.8$	$\frac{5\pi}{4}$	$-2\sqrt{2} \approx -2.8$
$\frac{\pi}{3}$	$2\sqrt{3} \approx 3.4$	$\frac{4\pi}{3}$	$-2\sqrt{3} \approx -3.4$

θ	$r=4\sin\theta$	θ	$r=4\sin\theta$
$\frac{\pi}{2}$	4	$\frac{3\pi}{2}$	-4
$\frac{2\pi}{3}$	$2\sqrt{3}\approx 3.4$	$\frac{5\pi}{3}$	$-2\sqrt{3}\approx -3.4$
$\frac{3\pi}{4}$	$2\sqrt{2}\approx 2.8$	$\frac{7\pi}{4}$	$-2\sqrt{2}\approx -2.8$
$\frac{5\pi}{6}$	2	$\frac{11\pi}{6}$	-2
		2π	0

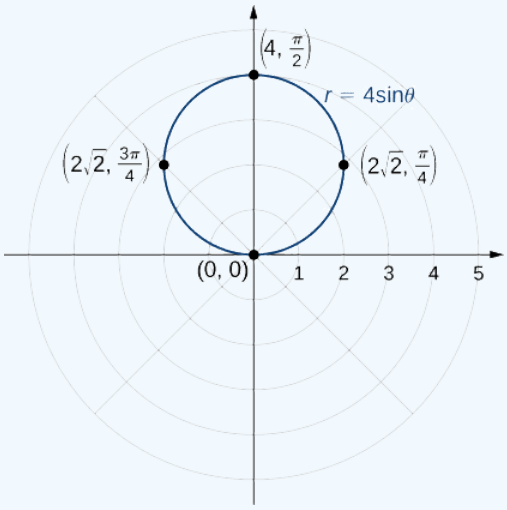


Figure 10.3.4: The graph of the function $r=4\sin\theta$ is a circle.

This is the graph of a circle. The equation $r=4\sin\theta$ can be converted into rectangular coordinates by first multiplying both sides by r. This gives the equation $r^2=4r\sin\theta$. Next use the facts that $r^2=x^2+y^2$ and $y=r\sin\theta$. This gives $x^2+y^2=4y$. To put this equation into standard form, subtract $4y$ from both sides of the equation and complete the square:

$$\begin{aligned} x^2+y^2-4y &= 0 \\ x^2+(y^2-4y) &= 0 \\ x^2+(y^2-4y+4) &= 0+4 \\ x^2+(y-2)^2 &= 4 \end{aligned}$$

This is the equation of a circle with radius 2 and center $(0,2)$ in the rectangular coordinate system.

? Exercise 10.3.3

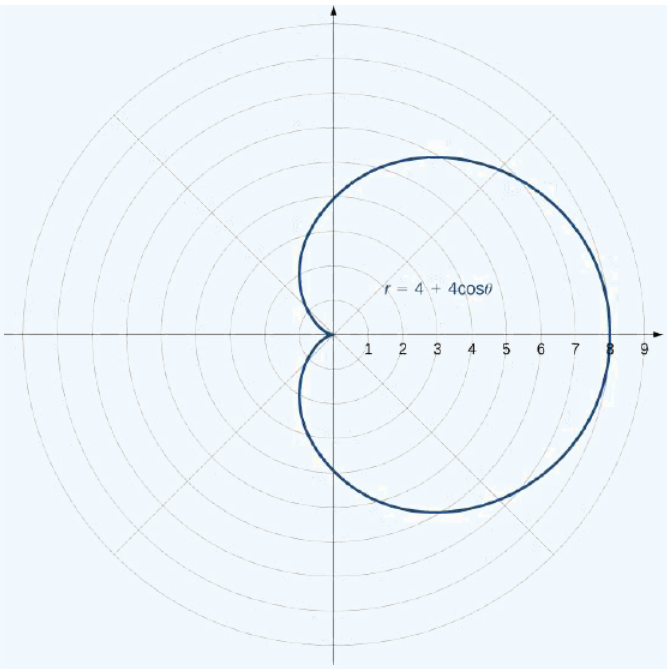
Create a graph of the curve defined by the function $r=4+4\cos\theta$.

Hint

Follow the problem-solving strategy for creating a graph in polar coordinates.

Answer

The name of this shape is a cardioid, which we will study further later in this section.



The graph in Example $10.3.3$ was that of a circle. The equation of the circle can be transformed into rectangular coordinates using the coordinate transformation formulas in Equation 10.3.1. Example $10.3.4$ gives some more examples of functions for transforming from polar to rectangular coordinates.

Example 10.3.4: Transforming Polar Equations to Rectangular Coordinates

Rewrite each of the following equations in rectangular coordinates and identify the graph.

a. $\theta = \dfrac{\pi}{3}$
b. $r = 3$
c. $r = 6\cos\theta - 8\sin\theta$

Solution:

a. Take the tangent of both sides. This gives $\tan\theta = \tan(\pi/3) = \sqrt{3}$. Since $\tan\theta = y/x$ we can replace the left-hand side of this equation by y/x. This gives $y/x = \sqrt{3}$, which can be rewritten as $y = x\sqrt{3}$. This is the equation of a straight line passing through the origin with slope $\sqrt{3}$. In general, any polar equation of the form $\theta = K$ represents a straight line through the pole with slope equal to $\tan K$.

b. First, square both sides of the equation. This gives $r^2 = 9$. Next replace r^2 with $x^2 + y^2$. This gives the equation $x^2 + y^2 = 9$, which is the equation of a circle centered at the origin with radius 3. In general, any polar equation of the form $r = k$ where *k* is a positive constant represents a circle of radius *k* centered at the origin. (Note: when squaring both sides of an equation it is possible to introduce new points unintentionally. This should always be taken into consideration. However, in this case we do not introduce new points. For example, $(-3, \dfrac{\pi}{3})$ is the same point as $(3, \dfrac{4\pi}{3})$.)

c. Multiply both sides of the equation by r. This leads to $r^2 = 6r\cos\theta - 8r\sin\theta$. Next use the formulas

$$r^2 = x^2 + y^2,\ x = r\cos\theta,\ y = r\sin\theta.$$

This gives

$$r^2 = 6(r\cos\theta) - 8(r\sin\theta)$$
$$x^2 + y^2 = 6x - 8y.$$

To put this equation into standard form, first move the variables from the right-hand side of the equation to the left-hand side, then complete the square.

$$x^2+y^2=6x-8y$$
$$x^2-6x+y^2+8y=0$$
$$(x^2-6x)+(y^2+8y)=0$$
$$(x^2-6x+9)+(y^2+8y+16)=9+16$$
$$(x-3)^2+(y+4)^2=25.$$

This is the equation of a circle with center at $(3,-4)$ and radius 5. Notice that the circle passes through the origin since the center is 5 units away.

Exercise 10.3.4

Rewrite the equation $r=\sec\theta\tan\theta$ in rectangular coordinates and identify its graph.

Hint

Convert to sine and cosine, then multiply both sides by cosine.

Answer

$y=x^2$, which is the equation of a parabola opening upward.

We have now seen several examples of drawing graphs of curves defined by **polar equations.** A summary of some common curves is given in the tables below. In each equation, *a* and *b* are arbitrary constants.

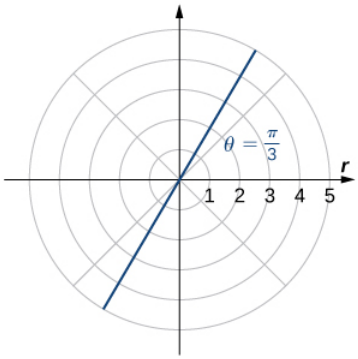
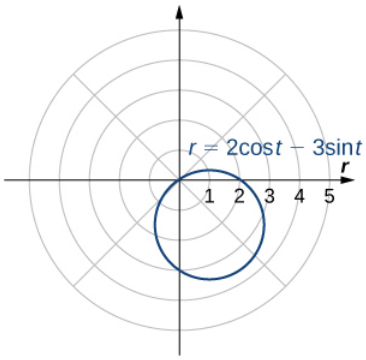
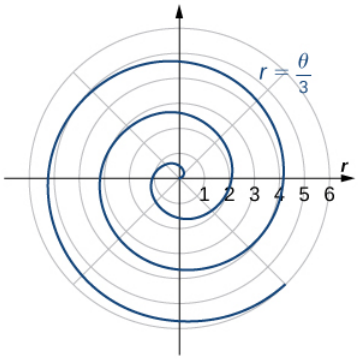
Name	Equation	Example
Line passing through the pole with slope tan K	$\theta = K$	
Circle	$r = a\cos\theta + b\sin\theta$	
Spiral	$r = a + b\theta$	

Figure 10.3.5

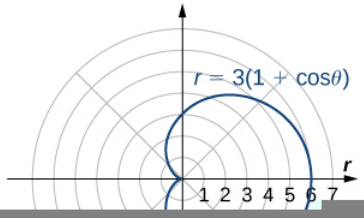
Name	Equation	Example
Cardioid	$r = a(1 + \cos\theta)$ $r = a(1 - \cos\theta)$ $r = a(1 + \sin\theta)$ $r = a(1 - \sin\theta)$	

Figure 10.3.6

A **cardioid** is a special case of a **limaçon** (pronounced “lee-mah-son”), in which $a = b$ or $a = -b$. The **rose i**s a very interesting curve. Notice that the graph of $r = 3\sin 2\theta$ has four petals. However, the graph of $r = 3\sin 3\theta$ has three petals as shown.

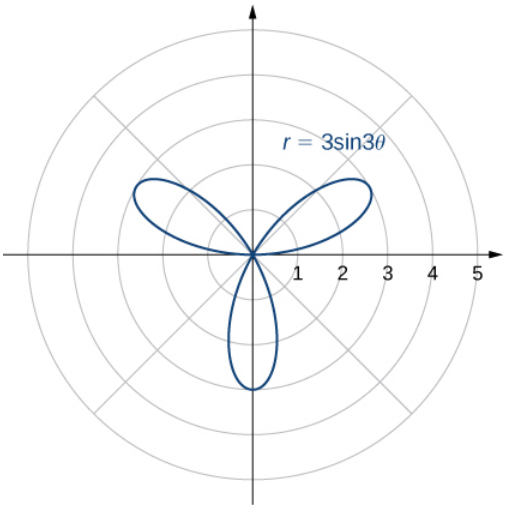


Figure 10.3.7: Graph of $r = 3\sin 3\theta$.

If the coefficient of θ is even, the graph has twice as many petals as the coefficient. If the coefficient of θ is odd, then the number of petals equals the coefficient. You are encouraged to explore why this happens. Even more interesting graphs emerge when the coefficient of θ is not an integer. For example, if it is rational, then the curve is closed; that is, it eventually ends where it started (Figure 10.3.8*a*). However, if the coefficient is irrational, then the curve never closes (Figure 10.3.8*b*). Although it may appear that the curve is closed, a closer examination reveals that the petals just above the positive x axis are slightly thicker. This is because the petal does not quite match up with the starting point.

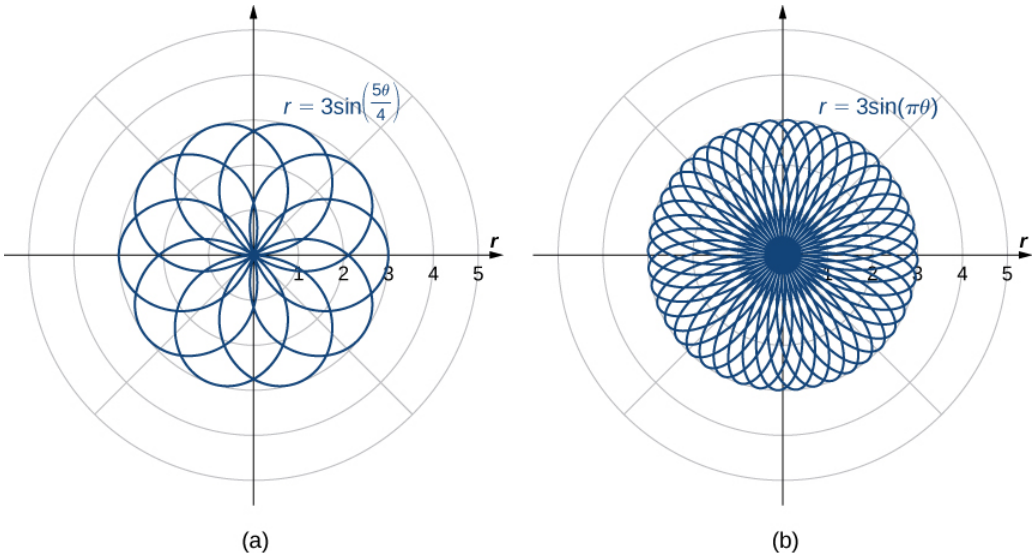


Figure 10.3.8: Polar rose graphs of functions with (a) rational coefficient and (b) irrational coefficient. Note that the rose in part (b) would actually fill the entire circle if plotted in full.

Since the curve defined by the graph of $r = 3\sin(\pi\theta)$ never closes, the curve depicted in Figure 10.3.8*b* is only a partial depiction. In fact, this is an example of a **space-filling curve.** A space-filling curve is one that in fact occupies a two-dimensional subset of the real plane. In this case the curve occupies the circle of radius 3 centered at the origin.

Example 10.3.5: Describing a Spiral

Recall the **chambered nautilus** introduced in the chapter prelude. This creature displays a spiral when half the outer shell is cut away. It is possible to describe a spiral using rectangular coordinates. Figure 10.3.9 shows a spiral in rectangular coordinates. How can we describe this curve mathematically?

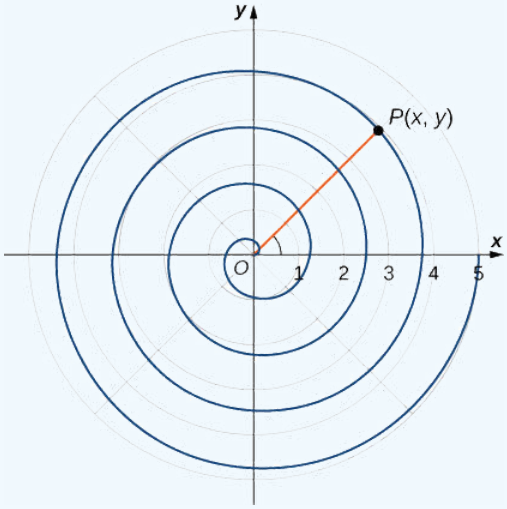


Figure 10.3.9: How can we describe a spiral graph mathematically?

Solution

As the point P travels around the spiral in a counterclockwise direction, its distance d from the origin increases. Assume that the distance d is a constant multiple k of the angle θ that the line segment OP makes with the positive x-axis. Therefore $d(P,O)=k\theta$, where O is the origin. Now use the distance formula and some trigonometry:

$$d(P,O)=k\theta$$

$$\sqrt{(x-0)^2+(y-0)^2}=k\arctan(\frac{y}{x})$$

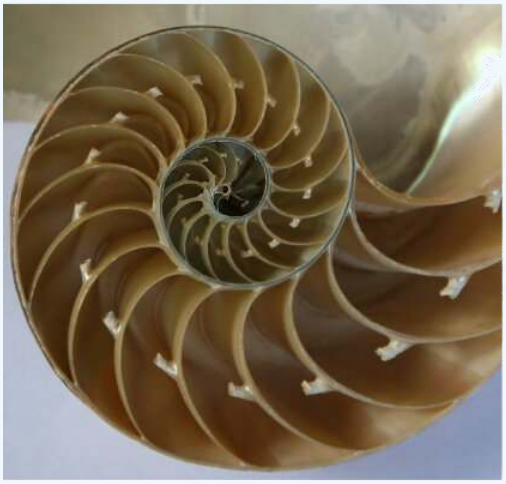
$$\sqrt{x^2+y^2}=k\arctan(\frac{y}{x})$$

$$\arctan(\frac{y}{x})=\frac{\sqrt{x^2+y^2}}{k}$$

$$y=x\tan(\frac{\sqrt{x^2+y^2}}{k}).$$

Although this equation describes the spiral, it is not possible to solve it directly for either x or y. However, if we use polar coordinates, the equation becomes much simpler. In particular, $d(P,O)=r$, and θ is the second coordinate. Therefore the equation for the spiral becomes $r=k\theta$. Note that when $\theta=0$ we also have $r=0$, so the spiral emanates from the origin. We can remove this restriction by adding a constant to the equation. Then the equation for the spiral becomes $r=a+k\theta$ for arbitrary constants a and k. This is referred to as an **Archimedean spiral**, after the Greek mathematician Archimedes.

Another type of spiral is the logarithmic spiral, described by the function $r=a\cdot b^{\theta}$. A graph of the function $r=1.2(1.25^{\theta})$ is given in Figure 10.3.10 This spiral describes the shell shape of the chambered nautilus.



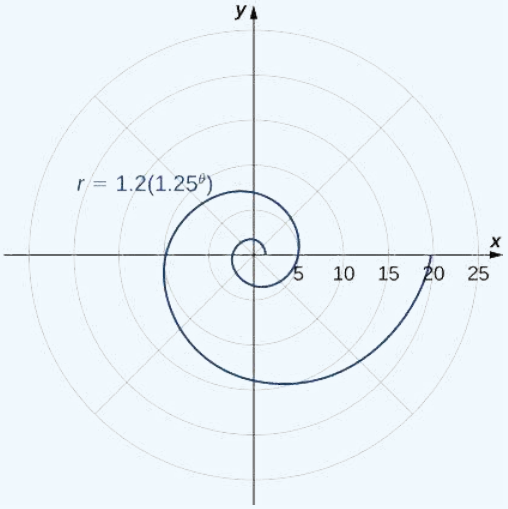


Figure 10.3.10: A logarithmic spiral is similar to the shape of the chambered nautilus shell. (credit: modification of work by Jitze Couperus, Flickr)

Suppose a curve is described in the polar coordinate system via the function $r = f(\theta)$. Since we have conversion formulas from polar to rectangular coordinates given by

$$x = r\cos\theta$$

$$y = r\sin\theta$$

,

it is possible to rewrite these formulas using the function

$$x = f(\theta)\cos\theta$$

$$y = f(\theta)\sin\theta.$$

This step gives a parameterization of the curve in rectangular coordinates using θ as the parameter. For example, the spiral formula $r = a + b\theta$ from Figure becomes

$$x = (a + b\theta)\cos\theta$$

$$y = (a + b\theta)\sin\theta.$$

Letting θ range from $-\infty$ to ∞ generates the entire spiral.

Symmetry in Polar Coordinates

When studying **symmetry** of functions in rectangular coordinates (i.e., in the form $y = f(x)$), we talk about symmetry with respect to the *y*-axis and symmetry with respect to the origin. In particular, if $f(-x) = f(x)$ for all x in the domain of f, then f is an even function and its graph is symmetric with respect to the *y*-axis. If $f(-x) = -f(x)$ for all *x* in the domain of f, then f is an odd function and its graph is symmetric with respect to the origin. By determining which types of symmetry a graph exhibits, we can learn more about the shape and appearance of the graph. Symmetry can also reveal other properties of the function that generates the graph. Symmetry in polar curves works in a similar fashion.

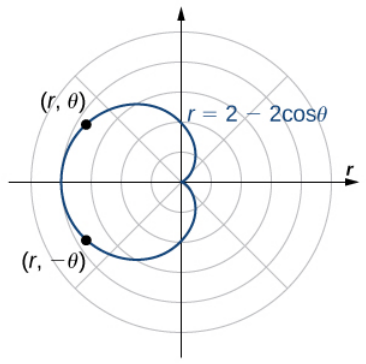
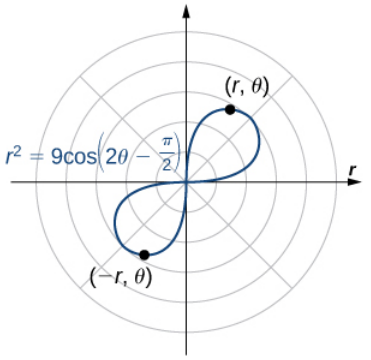
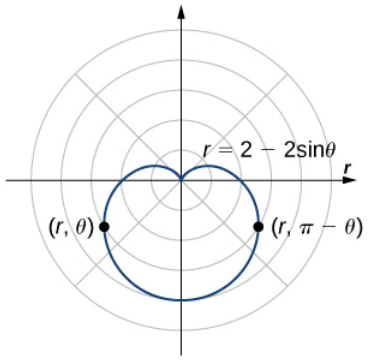
Symmetry in Polar Curves and Equations

Consider a curve generated by the function $r = f(\theta)$ in polar coordinates.

i. The curve is symmetric about the polar axis if for every point (r, θ) on the graph, the point $(r, -\theta)$ is also on the graph. Similarly, the equation $r = f(\theta)$ is unchanged by replacing θ with $-\theta$.
ii. The curve is symmetric about the pole if for every point (r, θ) on the graph, the point $(r, \pi + \theta)$ is also on the graph. Similarly, the equation $r = f(\theta)$ is unchanged when replacing r with $-r$, or θ with $\pi + \theta$.

iii. The curve is symmetric about the vertical line $\theta = \frac{\pi}{2}$ if for every point (r, θ) on the graph, the point $(r, \pi - \theta)$ is also on the graph. Similarly, the equation $r = f(\theta)$ is unchanged when θ is replaced by $\pi - \theta$.

The following table shows examples of each type of symmetry.

Symmetry with respect to the polar axis: For every point (r, θ) on the graph, there is also a point reflected directly across the horizontal (polar) axis.	
Symmetry with respect to the pole: For every point (r, θ) on the graph, there is also a point on the graph that is reflected through the pole as well.	
Symmetry with respect to the vertical line $\theta = \frac{\pi}{2}$: For every point (r, θ) on the graph, there is also a point reflected directly across the vertical axis.	

Example 10.3.6: Using Symmetry to Graph a Polar Equation

Find the symmetry of the rose defined by the equation $r = 3\sin(2\theta)$ and create a graph.

Solution

Suppose the point (r, θ) is on the graph of $r = 3\sin(2\theta)$.

i. To test for symmetry about the polar axis, first try replacing θ with $-\theta$. This gives $r = 3\sin(2(-\theta)) = -3\sin(2\theta)$. Since this changes the original equation, this test is not satisfied. However, returning to the original equation and replacing r with $-r$ and θ with $\pi - \theta$ yields

$$\begin{aligned} -r &= 3\sin(2(\pi-\theta)) \\ -r &= 3\sin(2\pi-2\theta) \\ -r &= 3\sin(-2\theta) \\ -r &= -3\sin 2\theta. \end{aligned}$$

Multiplying both sides of this equation by -1 gives $r = 3\sin 2\theta$, which is the original equation. This demonstrates that the graph is symmetric with respect to the polar axis.

ii. To test for symmetry with respect to the pole, first replace r with $-r$, which yields $-r = 3\sin(2\theta)$. Multiplying both sides by -1 gives $r = -3\sin(2\theta)$, which does not agree with the original equation. Therefore the equation does not pass the test for this symmetry. However, returning to the original equation and replacing θ with $\theta+\pi$ gives

$$\begin{aligned} r &= 3\sin(2(\theta+\pi)) \\ &= 3\sin(2\theta+2\pi) \\ &= 3(\sin 2\theta\cos 2\pi+\cos 2\theta\sin 2\pi) \\ &= 3\sin 2\theta. \end{aligned}$$

Since this agrees with the original equation, the graph is symmetric about the pole.

iii. To test for symmetry with respect to the vertical line $\theta = \dfrac{\pi}{2}$, first replace both r with $-r$ and θ with $-\theta$.

$$\begin{aligned} -r &= 3\sin(2(-\theta)) \\ -r &= 3\sin(-2\theta) \\ -r &= -3\sin 2\theta. \end{aligned}$$

Multiplying both sides of this equation by -1 gives $r = 3\sin 2\theta$, which is the original equation. Therefore the graph is symmetric about the vertical line $\theta = \dfrac{\pi}{2}$.

This graph has symmetry with respect to the polar axis, the origin, and the vertical line going through the pole. To graph the function, tabulate values of θ between 0 and $\pi/2$ and then reflect the resulting graph.

Table 10.3.2: Table of values for this polar function.

θ	$r = 3\sin(2\theta)$
0	0
$\dfrac{\pi}{6}$	$\dfrac{3\sqrt{3}}{2} \approx 2.6$
$\dfrac{\pi}{4}$	3
$\dfrac{\pi}{3}$	$\dfrac{3\sqrt{3}}{2} \approx 2.6$
$\dfrac{\pi}{2}$	0

This gives one petal of the rose, as shown in the following graph.

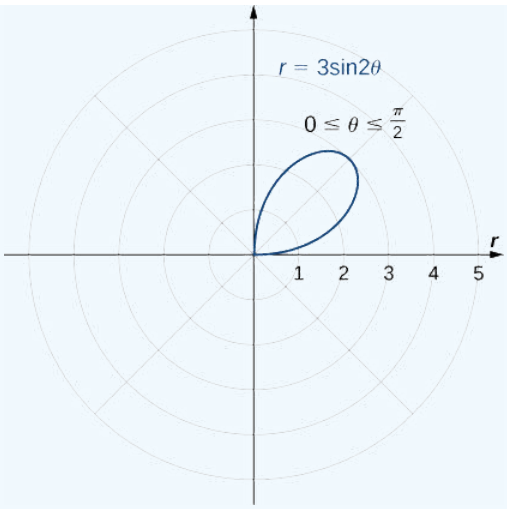


Figure 10.3.11: The graph of the equation between $\theta = 0$ and $\theta = \pi/2$.

Reflecting this image into the other three quadrants gives the entire graph as shown.

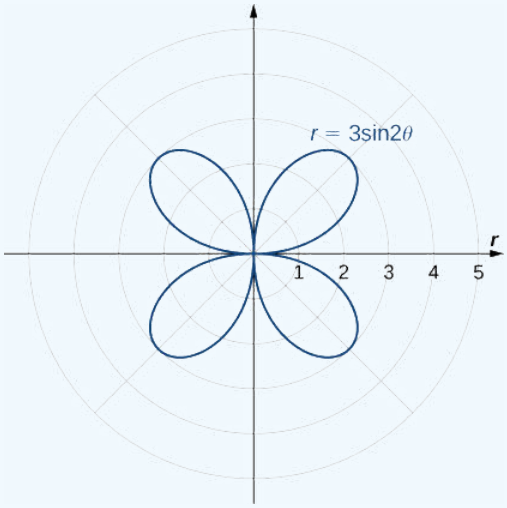


Figure 10.3.12: The entire graph of the equation is called a four-petaled rose.

? Exercise 10.3.5Symmetry

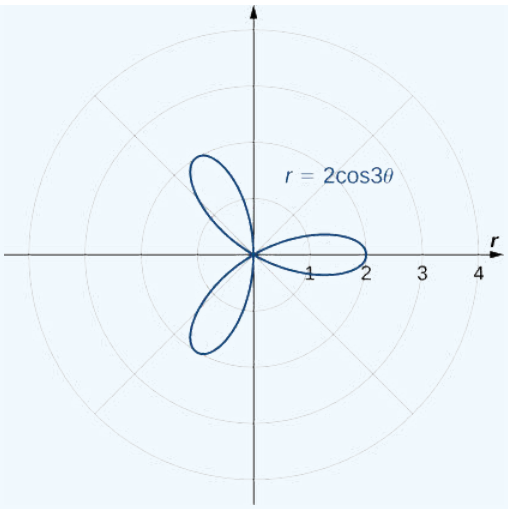
Determine the symmetry of the graph determined by the equation $r = 2\cos(3\theta)$ and create a graph.

Hint

Use Note.

Answer

Symmetric with respect to the polar axis.



Key Concepts

- The polar coordinate system provides an alternative way to locate points in the plane.
- Convert points between rectangular and polar coordinates using the formulas

$$x = r\cos\theta \text{ and } y = r\sin\theta$$

and

$$r = \sqrt{x^2 + y^2} \text{ and } \tan\theta = \frac{y}{x}.$$

- To sketch a polar curve from a given polar function, make a table of values and take advantage of periodic properties.
- Use the conversion formulas to convert equations between rectangular and polar coordinates.
- Identify symmetry in polar curves, which can occur through the pole, the horizontal axis, or the vertical axis.

Glossary

angular coordinate

θ the angle formed by a line segment connecting the origin to a point in the polar coordinate system with the positive radial x-axis, measured counterclockwise

cardioid

a plane curve traced by a point on the perimeter of a circle that is rolling around a fixed circle of the same radius; the equation of a cardioid is $r = a(1 + \sin\theta)$ or $r = a(1 + \cos\theta)$

limaçon

the graph of the equation $r = a + b\sin\theta$ or $r = a + b\cos\theta$. If $a = b$ then the graph is a cardioid

polar axis

the horizontal axis in the polar coordinate system corresponding to $r \geq 0$

polar coordinate system

a system for locating points in the plane. The coordinates are r, the radial coordinate, and θ, the angular coordinate

polar equation

an equation or function relating the radial coordinate to the angular coordinate in the polar coordinate system

pole

the central point of the polar coordinate system, equivalent to the origin of a Cartesian system

radial coordinate

r the coordinate in the polar coordinate system that measures the distance from a point in the plane to the pole

rose

graph of the polar equation $r = a\cos 2\theta$ or $r = a\sin 2\theta$ for a positive constant a

space-filling curve

a curve that completely occupies a two-dimensional subset of the real plane

---

10.3: Polar Coordinates is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **11.3: Polar Coordinates** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

10.4: Graphing in Polar Coordinates

10.4: Graphing in Polar Coordinates is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.5: Areas and Lengths in Polar Coordinates

Learning Objectives

- Apply the formula for area of a region in polar coordinates.
- Determine the arc length of a polar curve.

In the rectangular coordinate system, the definite integral provides a way to calculate the area under a curve. In particular, if we have a function $y = f(x)$ defined from $x = a$ to $x = b$ where $f(x) > 0$ on this interval, the area between the curve and the x-axis is given by

$$A = \int_a^b f(x)dx.$$

This fact, along with the formula for evaluating this integral, is summarized in the Fundamental Theorem of Calculus. Similarly, the arc length of this curve is given by

$$L = \int_a^b \sqrt{1 + (f'(x))^2}dx.$$

In this section, we study analogous formulas for area and arc length in the polar coordinate system.

Areas of Regions Bounded by Polar Curves

We have studied the formulas for area under a curve defined in rectangular coordinates and parametrically defined curves. Now we turn our attention to deriving a formula for the area of a region bounded by a polar curve. Recall that the proof of the Fundamental Theorem of Calculus used the concept of a Riemann sum to approximate the area under a curve by using rectangles. For polar curves we use the Riemann sum again, but the rectangles are replaced by sectors of a circle.

Consider a curve defined by the function $r = f(\theta)$, where $\alpha \leq \theta \leq \beta$. Our first step is to partition the interval $[\alpha, \beta]$ into n equal-width subintervals. The width of each subinterval is given by the formula $\Delta\theta = (\beta - \alpha)/n$, and the ith partition point θ_i is given by the formula $\theta_i = \alpha + i\Delta\theta$. Each partition point $\theta = \theta_i$ defines a line with slope $\tan\theta_i$ passing through the pole as shown in the following graph.

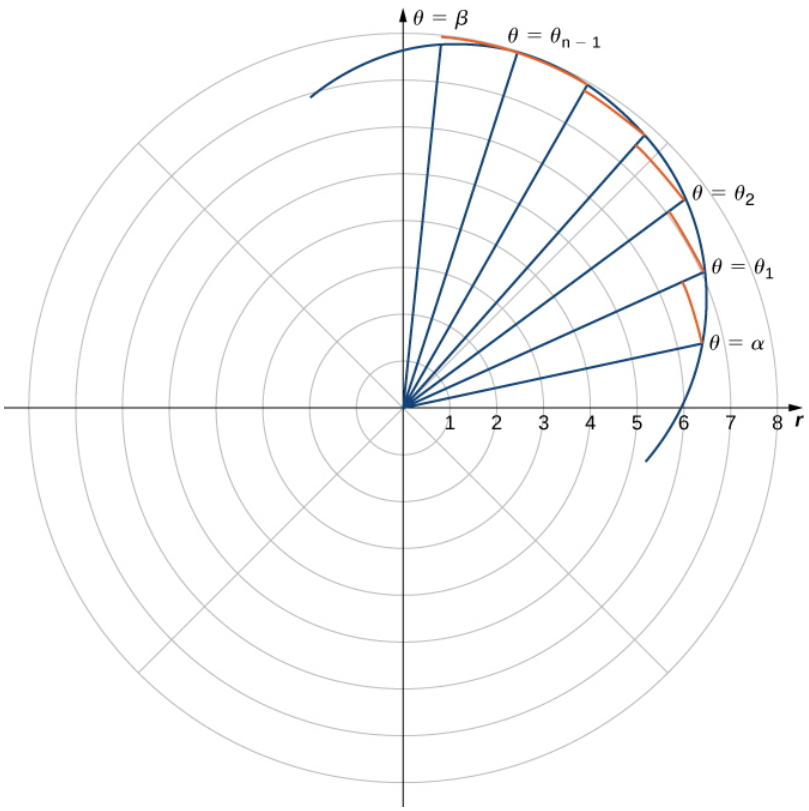


Figure 10.5.1: A partition of a typical curve in polar coordinates.

The line segments are connected by arcs of constant radius. This defines sectors whose areas can be calculated by using a geometric formula. The area of each sector is then used to approximate the area between successive line segments. We then sum the areas of the sectors to approximate the total area. This approach gives a Riemann sum approximation for the total area. The formula for the area of a sector of a circle is illustrated in the following figure.

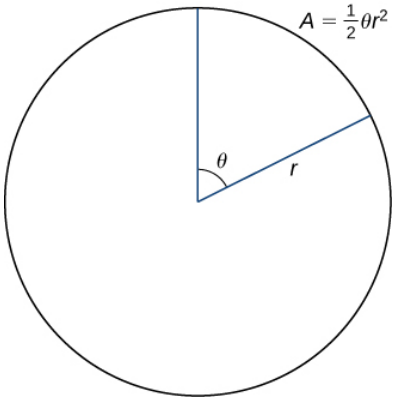


Figure 10.5.2: *The area of a sector of a circle is given by* $A = \frac{1}{2}\theta r^2$.

Recall that the area of a circle is $A = \pi r^2$. When measuring angles in radians, 360 degrees is equal to 2π radians. Therefore a fraction of a circle can be measured by the central angle θ. The fraction of the circle is given by $\frac{\theta}{2\pi}$, so the area of the sector is this fraction multiplied by the total area:

$$A = (\frac{\theta}{2\pi})\pi r^2 = \frac{1}{2}\theta r^2.$$

Since the radius of a typical sector in Figure $10.5.1$is given by $r_i = f(\theta_i)$, the area of the *i*th sector is given by

$$A_i = \frac{1}{2}(\Delta\theta)(f(\theta_i))^2.$$

Therefore a Riemann sum that approximates the area is given by

$$A_n = \sum_{i=1}^{n} A_i \approx \sum_{i=1}^{n} \frac{1}{2}(\Delta\theta)(f(\theta_i))^2.$$

We take the limit as $n \to \infty$ to get the exact area:

$$A = \lim_{n\to\infty} A_n = \frac{1}{2}\int_{\alpha}^{\beta} (f(\theta))^2 d\theta.$$

This gives the following theorem.

Area of a Region Bounded by a Polar Curve

Suppose f is continuous and nonnegative on the interval $\alpha \le \theta \le \beta$ with $0 < \beta - \alpha \le 2\pi$. The area of the region bounded by the graph of $r = f(\theta)$ between the radial lines $\theta = \alpha$ and $\theta = \beta$ is

$$A = \frac{1}{2}\int_{\alpha}^{\beta} [f(\theta)]^2 d\theta \tag{10.5.1}$$

$$= \frac{1}{2}\int_{\alpha}^{\beta} r^2 d\theta. \tag{10.5.2}$$

Example 10.5.1: Finding an Area of a Polar Region

Find the area of one petal of the rose defined by the equation $r = 3\sin(2\theta)$.

Solution

The graph of $r = 3\sin(2\theta)$ follows.

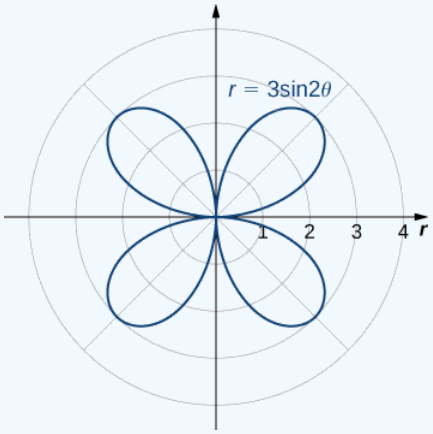


Figure 10.5.3: The graph of $r = 3\sin(2\theta)$.

When $\theta = 0$ we have $r = 3\sin(2(0)) = 0$. The next value for which $r = 0$ is $\theta = \pi/2$. This can be seen by solving the equation $3\sin(2\theta) = 0$ for θ. Therefore the values $\theta = 0$ to $\theta = \pi/2$ trace out the first petal of the rose. To find the area inside this petal, use Equation 10.5.2 with $f(\theta) = 3\sin(2\theta), \alpha = 0,$ and $\beta = \pi/2$:

$$A = \frac{1}{2}\int_{\alpha}^{\beta} [f(\theta)]^2 d\theta$$

$$= \frac{1}{2}\int_{0}^{\pi/2} [3\sin(2\theta)]^2 d\theta$$

$$= \frac{1}{2}\int_{0}^{\pi/2} 9\sin^2(2\theta) d\theta.$$

To evaluate this integral, use the formula $\sin^2\alpha = (1 - \cos(2\alpha))/2$ with $\alpha = 2\theta$:

$$
\begin{aligned}
A &= \frac{1}{2}\int_0^{\pi/2} 9\sin^2(2\theta)d\theta \\
&= \frac{9}{2}\int_0^{\pi/2} \frac{(1-\cos(4\theta))}{2}d\theta \\
&= \frac{9}{4}(\int_0^{\pi/2} 1-\cos(4\theta)d\theta) \\
&= \frac{9}{4}(\theta - \frac{\sin(4\theta)}{4}\Big|_0^{\pi/2} \\
&= \frac{9}{4}(\frac{\pi}{2} - \frac{\sin 2\pi}{4}) - \frac{9}{4}(0 - \frac{\sin 4(0)}{4}) \\
&= \frac{9\pi}{8}
\end{aligned}
$$

? Exercise 10.5.1

Find the area inside the cardioid defined by the equation $r = 1 - \cos\theta$.

Hint

Use Equation 10.5.2 Be sure to determine the correct limits of integration before evaluating.

Answer

$A = 3\pi/2$

Example 10.5.1 involved finding the area inside one curve. We can also use Equation 10.5.2 to find the area between two polar curves. However, we often need to find the points of intersection of the curves and determine which function defines the outer curve or the inner curve between these two points.

✓ Example 10.5.2: Finding the Area between Two Polar Curves

Find the area outside the cardioid $r = 2 + 2\sin\theta$ and inside the circle $r = 6\sin\theta$.

Solution

First draw a graph containing both curves as shown.

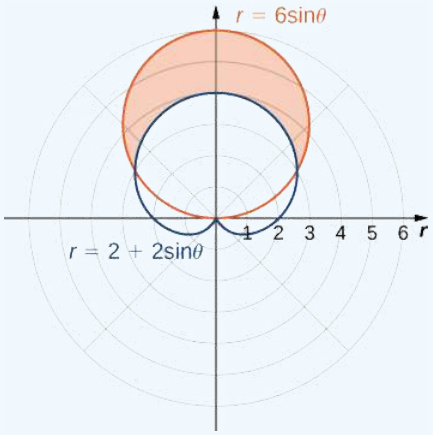


Figure 10.5.4: *The region between the curves* $r = 2 + 2\sin\theta$ *and* $r = 6\sin\theta$.

To determine the limits of integration, first find the points of intersection by setting the two functions equal to each other and solving for θ:

$$6\sin\theta = 2 + 2\sin\theta$$
$$4\sin\theta = 2$$
$$\sin\theta = \frac{1}{2}.$$

This gives the solutions $\theta = \frac{\pi}{6}$ and $\theta = \frac{5\pi}{6}$, which are the limits of integration. The circle $r = 3\sin\theta$ is the red graph, which is the outer function, and the cardioid $r = 2 + 2\sin\theta$ is the blue graph, which is the inner function. To calculate the area between the curves, start with the area inside the circle between $\theta = \frac{\pi}{6}$ and $\theta = \frac{5\pi}{6}$, then subtract the area inside the cardioid between $\theta = \frac{\pi}{6}$ and $\theta = \frac{5\pi}{6}$:

$$\begin{aligned} A &= \text{circle} - \text{cardioid} \\ &= \frac{1}{2}\int_{\pi/6}^{5\pi/6}[6\sin\theta]^2 d\theta - \frac{1}{2}\int_{\pi/6}^{5\pi/6}[2+2\sin\theta]^2 d\theta \\ &= \frac{1}{2}\int_{\pi/6}^{5\pi/6} 36\sin^2\theta\, d\theta - \frac{1}{2}\int_{\pi/6}^{5\pi/6} 4 + 8\sin\theta + 4\sin^2\theta\, d\theta \\ &= 18\int_{\pi/6}^{5\pi/6} \frac{1-\cos(2\theta)}{2} d\theta - 2\int_{\pi/6}^{5\pi/6} 1 + 2\sin\theta + \frac{1-\cos(2\theta)}{2} d\theta \\ &= 9[\theta - \frac{\sin(2\theta)}{2}]_{\pi/6}^{5\pi/6} - 2[\frac{3\theta}{2} - 2\cos\theta - \frac{\sin(2\theta)}{4}]_{\pi/6}^{5\pi/6} \\ &= 9(\frac{5\pi}{6} - \frac{\sin(10\pi/6)}{2}) - 9(\frac{\pi}{6} - \frac{\sin(2\pi/6)}{2}) - (3(\frac{5\pi}{6}) - 4\cos\frac{5\pi}{6} - \frac{\sin(10\pi/6)}{2}) \\ &+ (3(\frac{\pi}{6}) - 4\cos\frac{\pi}{6} - \frac{\sin(2\pi/6)}{2}) \\ &= 4\pi. \end{aligned}$$

? Exercise 10.5.2

Find the area inside the circle $r = 4\cos\theta$ and outside the circle $r = 2$.

Hint

Use Equation 10.5.2 and take advantage of symmetry.

Answer

$$A = \frac{4\pi}{3} + 2\sqrt{3}$$

In Example 10.5.2 we found the area inside the circle and outside the cardioid by first finding their intersection points. Notice that solving the equation directly for θ yielded two solutions: $\theta = \frac{\pi}{6}$ and $\theta = \frac{5\pi}{6}$. However, in the graph there are three intersection points. The third intersection point is the origin. The reason why this point did not show up as a solution is because the origin is on both graphs but for different values of θ. For example, for the cardioid we get

$$2 + 2\sin\theta = 0$$
$$\sin\theta = -1,$$

so the values for θ that solve this equation are $\theta = \frac{3\pi}{2} + 2n\pi$, where n is any integer. For the circle we get

$$6\sin\theta = 0.$$

The solutions to this equation are of the form $\theta = n\pi$ for any integer value of n. These two solution sets have no points in common. Regardless of this fact, the curves intersect at the origin. This case must always be taken into consideration.

Arc Length in Polar Curves

Here we derive a formula for the arc length of a curve defined in polar coordinates. In rectangular coordinates, the arc length of a parameterized curve $(x(t), y(t))$ for $a \le t \le b$ is given by

$$L = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt.$$

In polar coordinates we define the curve by the equation $r = f(\theta)$, where $\alpha \le \theta \le \beta$. In order to adapt the arc length formula for a polar curve, we use the equations

$$x = r\cos\theta = f(\theta)\cos\theta$$

and

$$y = r\sin\theta = f(\theta)\sin\theta,$$

and we replace the parameter t by θ. Then

$$\frac{dx}{d\theta} = f'(\theta)\cos\theta - f(\theta)\sin\theta$$

$$\frac{dy}{d\theta} = f'(\theta)\sin\theta + f(\theta)\cos\theta.$$

We replace dt by $d\theta$, and the lower and upper limits of integration are α and β, respectively. Then the arc length formula becomes

$$\begin{aligned} L &= \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt \\ &= \int_\alpha^\beta \sqrt{\left(\frac{dx}{d\theta}\right)^2 + \left(\frac{dy}{d\theta}\right)^2}\, d\theta \\ &= \int_\alpha^\beta \sqrt{(f'(\theta)\cos\theta - f(\theta)\sin\theta)^2 + (f'(\theta)\sin\theta + f(\theta)\cos\theta)^2}\, d\theta \\ &= \int_\alpha^\beta \sqrt{(f'(\theta))^2(\cos^2\theta + \sin^2\theta) + (f(\theta))^2(\cos^2\theta + \sin^2\theta)}\, d\theta \\ &= \int_\alpha^\beta \sqrt{(f'(\theta))^2 + (f(\theta))^2}\, d\theta \\ &= \int_\alpha^\beta \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2}\, d\theta \end{aligned}$$

This gives us the following theorem.

Arc Length of a Curve Defined by a Polar Function

Let f be a function whose derivative is continuous on an interval $\alpha \le \theta \le \beta$. The arc length of the graph of $r = f(\theta)$ from $\theta = \alpha$ to $\theta = \beta$ is

$$L = \int_\alpha^\beta \sqrt{[f(\theta)]^2 + [f'(\theta)]^2}\, d\theta \tag{10.5.3}$$

$$= \int_\alpha^\beta \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2}\, d\theta. \tag{10.5.4}$$

Example 10.5.3: Finding the Arc Length of a cardioid

Find the arc length of the cardioid $r = 2 + 2\cos\theta$.

Solution

When $\theta = 0, r = 2 + 2\cos 0 = 4$. Furthermore, as θ goes from 0 to 2π, the cardioid is traced out exactly once. Therefore these are the limits of integration. Using $f(\theta) = 2 + 2\cos\theta, \alpha = 0,$ and $\beta = 2\pi$, Equation 10.5.3 becomes

$$\begin{aligned} L &= \int_{\alpha}^{\beta} \sqrt{[f(\theta)]^2 + [f'(\theta)]^2}\, d\theta \\ &= \int_0^{2\pi} \sqrt{[2 + 2\cos\theta]^2 + [-2\sin\theta]^2}\, d\theta \\ &= \int_0^{2\pi} \sqrt{4 + 8\cos\theta + 4\cos^2\theta + 4\sin^2\theta}\, d\theta \\ &= \int_0^{2\pi} \sqrt{4 + 8\cos\theta + 4(\cos^2\theta + \sin^2\theta)}\, d\theta \\ &= \int_0^{2\pi} \sqrt{8 + 8\cos\theta}\, d\theta \\ &= 2\int_0^{2\pi} \sqrt{2 + 2\cos\theta}\, d\theta. \end{aligned}$$

Next, using the identity $\cos(2\alpha) = 2\cos^2\alpha - 1,$ add 1 to both sides and multiply by 2. This gives $2 + 2\cos(2\alpha) = 4\cos^2\alpha$. Substituting $\alpha = \theta/2$ gives $2 + 2\cos\theta = 4\cos^2(\theta/2)$, so the integral becomes

$$\begin{aligned} L &= 2\int_0^{2\pi} \sqrt{2 + 2\cos\theta}\, d\theta \\ &= 2\int_0^{2\pi} \sqrt{4\cos^2(\frac{\theta}{2})}\, d\theta \\ &= 4\int_0^{2\pi} |\cos(\frac{\theta}{2})|\, d\theta. \end{aligned}$$

The absolute value is necessary because the cosine is negative for some values in its domain. To resolve this issue, change the limits from 0 to π and double the answer. This strategy works because cosine is positive between 0 and $\frac{\pi}{2}$. Thus,

$$\begin{aligned} L &= 4\int_0^{2\pi} |\cos(\frac{\theta}{2})|\, d\theta \\ &= 8\int_0^{\pi} \cos(\frac{\theta}{2})\, d\theta \\ &= 8(2\sin(\frac{\theta}{2})\Big|_0^{\pi} \\ &= 16 \end{aligned}$$

Exercise 10.5.3

Find the total arc length of the graph of $r = 3\sin\theta$.

Hint

Use Equation 10.5.3 To determine the correct limits, make a table of values.

Answer

$s = 3\pi$

Key Concepts

- The area of a region in polar coordinates defined by the equation $r = f(\theta)$ with $\alpha \leq \theta \leq \beta$ is given by the integral $A = \frac{1}{2}\int_{\alpha}^{\beta}[f(\theta)]^2 d\theta$.
- To find the area between two curves in the polar coordinate system, first find the points of intersection, then subtract the corresponding areas.
- The arc length of a polar curve defined by the equation $r = f(\theta)$ with $\alpha \leq \theta \leq \beta$ is given by the integral $L = \int_{\alpha}^{\beta}\sqrt{[f(\theta)]^2 + [f\,'(\theta)]^2}\,d\theta = \int_{\alpha}^{\beta}\sqrt{r^2 + (\frac{dr}{d\theta})^2}\,d\theta$.

Key Equations

- **Area of a region bounded by a polar curve**

$$A = \frac{1}{2}\int_{\alpha}^{\beta}[f(\theta)]^2 d\theta = \frac{1}{2}\int_{\alpha}^{\beta} r^2 d\theta$$

- **Arc length of a polar curve**

$$L = \int_{\alpha}^{\beta}\sqrt{[f(\theta)]^2 + [f\,'(\theta)]^2}\,d\theta = \int_{\alpha}^{\beta}\sqrt{r^2 + (\frac{dr}{d\theta})^2}\,d\theta$$

10.5: Areas and Lengths in Polar Coordinates is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **11.4: Area and Arc Length in Polar Coordinates** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

10.6: Conics in Polar Coordinates

Learning Objectives

- Identify a conic in polar form.
- Graph the polar equations of conics.
- Define conics in terms of a focus and a directrix.

Most of us are familiar with orbital motion, such as the motion of a planet around the sun or an electron around an atomic nucleus. Within the planetary system, orbits of planets, asteroids, and comets around a larger celestial body are often elliptical. Comets, however, may take on a parabolic or hyperbolic orbit instead. And, in reality, the characteristics of the planets' orbits may vary over time. Each orbit is tied to the location of the celestial body being orbited and the distance and direction of the planet or other object from that body. As a result, we tend to use polar coordinates to represent these orbits.

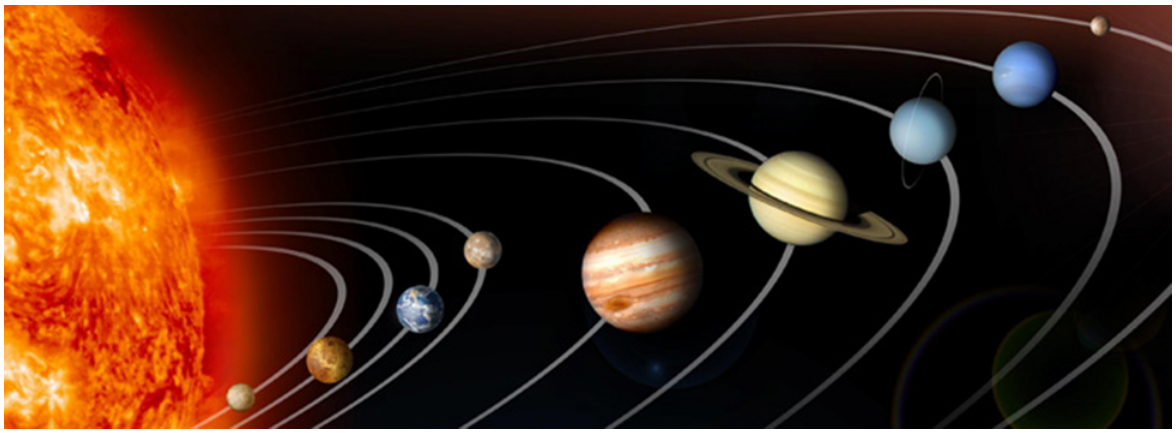


Figure 10.6.1: Planets orbiting the sun follow elliptical paths. (credit: NASA Blueshift, Flickr)

In an elliptical orbit, the **periapsis** is the point at which the two objects are closest, and the **apoapsis** is the point at which they are farthest apart. Generally, the velocity of the orbiting body tends to increase as it approaches the periapsis and decrease as it approaches the apoapsis. Some objects reach an escape velocity, which results in an infinite orbit. These bodies exhibit either a parabolic or a hyperbolic orbit about a body; the orbiting body breaks free of the celestial body's gravitational pull and fires off into space. Each of these orbits can be modeled by a conic section in the polar coordinate system.

Identifying a Conic in Polar Form

Any conic may be determined by three characteristics: a single **focus**, a fixed line called the **directrix**, and the ratio of the distances of each to a point on the graph. Consider the **parabola** $x = 2 + y^2$ shown in Figure $10.6.2$

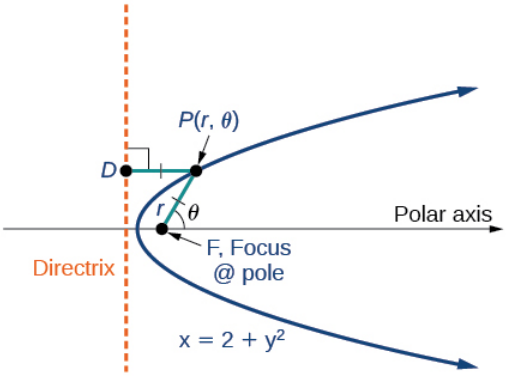


Figure 10.6.2

We previously learned how a parabola is defined by the focus (a fixed point) and the directrix (a fixed line). In this section, we will learn how to define any conic in the polar coordinate system in terms of a fixed point, the focus $P(r, \theta)$ at the pole, and a line, the directrix, which is perpendicular to the polar axis.

If F is a fixed point, the focus, and D is a fixed line, the directrix, then we can let e be a fixed positive number, called the **eccentricity**, which we can define as the ratio of the distances from a point on the graph to the focus and the point on the graph to the directrix. Then the set of all points P such that $e = \frac{PF}{PD}$ is a conic. In other words, we can define a conic as the set of all points P with the property that the ratio of the distance from P to F to the distance from P to D is equal to the constant e.

For a conic with eccentricity e,

- if $0 \le e < 1$, the conic is an ellipse
- if $e = 1$, the conic is a parabola
- if $e > 1$, the conic is an hyperbola

With this definition, we may now define a conic in terms of the directrix, $x = \pm p$, the eccentricity e, and the angle θ. Thus, each conic may be written as a **polar equation**, an equation written in terms of r and θ.

THE POLAR EQUATION FOR A CONIC

For a conic with a focus at the origin, if the directrix is $x = \pm p$, where p is a positive real number, and the eccentricity is a positive real number e, the conic has a polar equation

$$r = \frac{ep}{1 \pm e\cos\theta} \tag{10.6.1}$$

For a conic with a focus at the origin, if the directrix is $y = \pm p$, where p is a positive real number, and the eccentricity is a positive real number e, the conic has a polar equation

$$r = \frac{ep}{1 \pm e\sin\theta} \tag{10.6.2}$$

How to: Given the polar equation for a conic, identify the type of conic, the directrix, and the eccentricity.

1. Multiply the numerator and denominator by the reciprocal of the constant in the denominator to rewrite the equation in standard form.
2. Identify the eccentricity e as the coefficient of the trigonometric function in the denominator.
3. Compare e with 1 to determine the shape of the conic.
4. Determine the directrix as $x = p$ if cosine is in the denominator and $y = p$ if sine is in the denominator. Set ep equal to the numerator in standard form to solve for x or y.

Example 10.6.1: Identifying a Conic Given the Polar Form

For each of the following equations, identify the conic with focus at the origin, the directrix, and the eccentricity.

a. $r = \dfrac{6}{3 + 2\sin\theta}$

b. $r = \dfrac{12}{4 + 5\cos\theta}$

c. $r = \dfrac{7}{2 - 2\sin\theta}$

Solution

For each of the three conics, we will rewrite the equation in standard form. Standard form has a 1 as the constant in the denominator. Therefore, in all three parts, the first step will be to multiply the numerator and denominator by the reciprocal of the constant of the original equation, $\frac{1}{c}$, where c is that constant.

a. Multiply the numerator and denominator by $\frac{1}{3}$.

$$r=\frac{6}{3+2\sin\theta}\cdot\frac{\left(\frac{1}{3}\right)}{\left(\frac{1}{3}\right)}=\frac{6\left(\frac{1}{3}\right)}{3\left(\frac{1}{3}\right)+2\left(\frac{1}{3}\right)\sin\theta}=\frac{2}{1+\frac{2}{3}\sin\theta}$$

Because $\sin\theta$ is in the denominator, the directrix is $y=p$. Comparing to standard form, note that $e=\frac{2}{3}$ .Therefore, from the numerator,

$$\begin{aligned}2&=ep\\2&=\frac{2}{3}p\\\left(\frac{3}{2}\right)2&=\left(\frac{3}{2}\right)\frac{2}{3}p\\3&=p\end{aligned}$$

Since $e<1$, the conic is an **ellipse**. The eccentricity is $e=\frac{2}{3}$ and the directrix is $y=3$.

b. Multiply the numerator and denominator by $\frac{1}{4}$.

$$\begin{aligned}r&=\frac{12}{4+5\cos\theta}\cdot\frac{\left(\frac{1}{4}\right)}{\left(\frac{1}{4}\right)}\\r&=\frac{12\left(\frac{1}{4}\right)}{4\left(\frac{1}{4}\right)+5\left(\frac{1}{4}\right)\cos\theta}\\r&=\frac{3}{1+\frac{5}{4}\cos\theta}\end{aligned}$$

Because $\cos\theta$ is in the denominator, the directrix is $x=p$. Comparing to standard form, $e=\frac{5}{4}$. Therefore, from the numerator,

$$\begin{aligned}3&=ep\\3&=\frac{5}{4}p\\\left(\frac{4}{5}\right)3&=\left(\frac{4}{5}\right)\frac{5}{4}p\\\frac{12}{5}&=p\end{aligned}$$

Since $e>1$, the conic is a **hyperbola**. The eccentricity is $e=\frac{5}{4}$ and the directrix is $x=\frac{12}{5}=2.4$.

c. Multiply the numerator and denominator by $\frac{1}{2}$.

$$r=\frac{7}{2-2\sin\theta}\cdot\frac{\left(\frac{1}{2}\right)}{\left(\frac{1}{2}\right)}$$

$$r=\frac{7\left(\frac{1}{2}\right)}{2\left(\frac{1}{2}\right)-2\left(\frac{1}{2}\right)\sin\theta}$$

$$r=\frac{\frac{7}{2}}{1-\sin\theta}$$

Because sine is in the denominator, the directrix is $y=-p$. Comparing to standard form, $e=1$. Therefore, from the numerator,

$$\frac{7}{2}=ep$$
$$\frac{7}{2}=(1)p$$
$$\frac{7}{2}=p$$

Because $e=1$, the conic is a parabola. The eccentricity is $e=1$ and the directrix is $y=-\frac{7}{2}=-3.5$.

Exercise 10.6.1

Identify the conic with focus at the origin, the directrix, and the eccentricity for $r=\frac{2}{3-\cos\theta}$.

Answer

ellipse; $e=\frac{1}{3}$; $x=-2$

Graphing the Polar Equations of Conics

When graphing in Cartesian coordinates, each conic section has a unique equation. This is not the case when graphing in polar coordinates. We must use the eccentricity of a conic section to determine which type of curve to graph, and then determine its specific characteristics. The first step is to rewrite the conic in standard form as we have done in the previous example. In other words, we need to rewrite the equation so that the denominator begins with 1. This enables us to determine e and, therefore, the shape of the curve. The next step is to substitute values for θ and solve for r to plot a few key points. Setting θ equal to 0, $\frac{\pi}{2}$, π, and $\frac{3\pi}{2}$ provides the vertices so we can create a rough sketch of the graph.

Example 10.6.2A: Graphing a Parabola in Polar Form

Graph $r=\frac{5}{3+3\cos\theta}$.

Solution

First, we rewrite the conic in standard form by multiplying the numerator and denominator by the reciprocal of 3, which is $\frac{1}{3}$.

$$r=\frac{5}{3+3\cos\theta}=\frac{5\left(\frac{1}{3}\right)}{3\left(\frac{1}{3}\right)+3\left(\frac{1}{3}\right)\cos\theta}$$

$$r=\frac{\frac{5}{3}}{1+\cos\theta}$$

Because $e=1$,we will graph a **parabola** with a focus at the origin. The function has a $\cos\theta$, and there is an addition sign in the denominator, so the directrix is $x=p$.

$$\frac{5}{3}=ep$$
$$\frac{5}{3}=(1)p$$
$$\frac{5}{3}=p$$

The directrix is $x=\frac{5}{3}$.

Plotting a few key points as in Table 10.6.1 will enable us to see the vertices. See Figure 10.6.3

Table 10.6.1

	A	B	C	D
θ	0	$\frac{\pi}{2}$	π	$\frac{3\pi}{2}$
$r=\frac{5}{3+3\cos\theta}$	$\frac{5}{6}\approx 0.83$	$\frac{5}{3}\approx 1.67$	undefined	$\frac{5}{3}\approx 1.67$

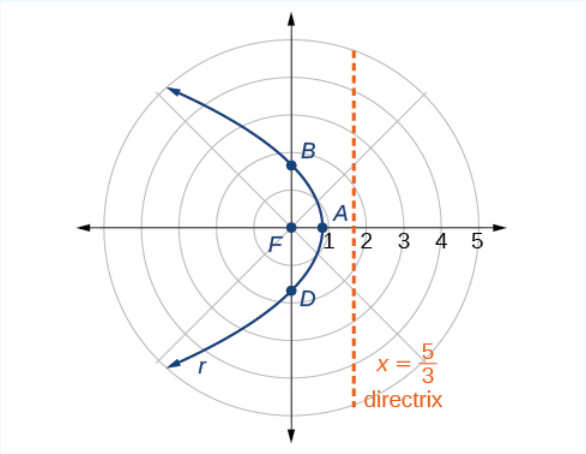


Figure 10.6.3

We can check our result with a graphing utility. See Figure 10.6.4

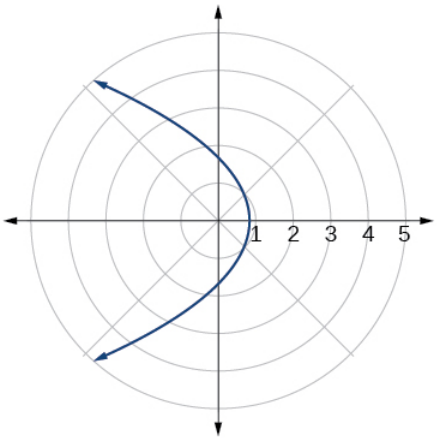


Figure 10.6.4

Example $10.6.2B$: Graphing a Hyperbola in Polar Form

Graph $r = \dfrac{8}{2-3\sin\theta}$.

Solution

First, we rewrite the conic in standard form by multiplying the numerator and denominator by the reciprocal of 2, which is $\frac{1}{2}$.

$$r = \frac{8}{2-3\sin\theta} = \frac{8\left(\frac{1}{2}\right)}{2\left(\frac{1}{2}\right)-3\left(\frac{1}{2}\right)\sin\theta}$$

$$r = \frac{4}{1-\frac{3}{2}\sin\theta}$$

Because $e = \frac{3}{2}$, $e > 1$, so we will graph a hyperbola with a focus at the origin. The function has a $\sin\theta$ term and there is a subtraction sign in the denominator, so the directrix is $y = -p$.

$$4 = ep$$
$$4 = \left(\frac{3}{2}\right)p$$
$$4\left(\frac{2}{3}\right) = p$$
$$\frac{8}{3} = p$$

The directrix is $y = -\frac{8}{3}$.

Plotting a few key points as in Table $10.6.2$ will enable us to see the vertices. See Figure $10.6.5$.

Table 10.6.2

	A	B	C	D
θ	0	$\frac{\pi}{2}$	π	$\frac{3\pi}{2}$
$r = \frac{8}{2-3\sin\theta}$	4	-8	4	$\frac{8}{5} = 1.6$

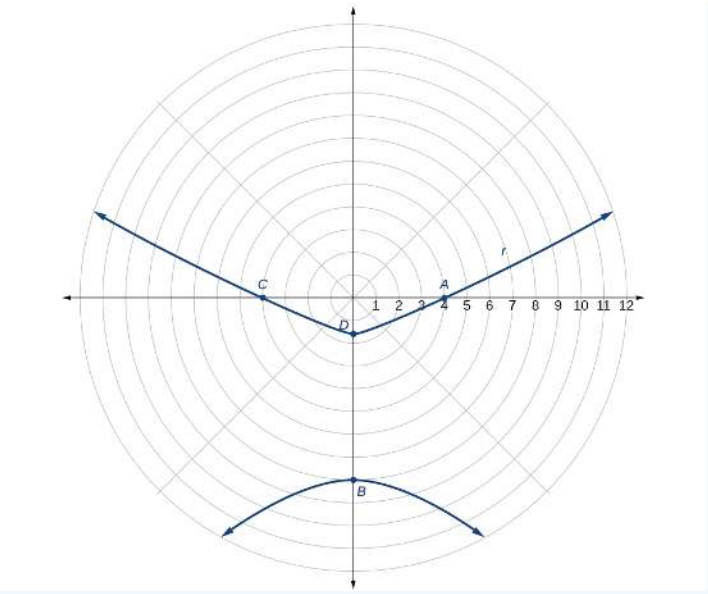


Figure 10.6.5

Example 10.6.2C: Graphing an Ellipse in Polar Form

Graph $r = \dfrac{10}{5 - 4\cos\theta}$.

Solution

First, we rewrite the conic in standard form by multiplying the numerator and denominator by the reciprocal of 5, which is $\frac{1}{5}$.

$$r = \frac{10}{5 - 4\cos\theta} = \frac{10\left(\frac{1}{5}\right)}{5\left(\frac{1}{5}\right) - 4\left(\frac{1}{5}\right)\cos\theta}$$

$$r = \frac{2}{1 - \frac{4}{5}\cos\theta}$$

Because $e = \frac{4}{5}$, $e < 1$, so we will graph an **ellipse** with a **focus** at the origin. The function has a $\cos\theta$, and there is a subtraction sign in the denominator, so the **directrix** is $x = -p$.

$$2 = ep$$
$$2 = \left(\frac{4}{5}\right)p$$
$$2\left(\frac{5}{4}\right) = p$$
$$\frac{5}{2} = p$$

The directrix is $x = -\frac{5}{2}$.

Plotting a few key points as in Table 10.6.3 will enable us to see the vertices. See Figure 10.6.6

Table 10.6.3

	A	B	C	D
θ	0	$\frac{\pi}{2}$	π	$\frac{3\pi}{2}$

	A	B	C	D
$r = \dfrac{10}{5 - 4\cos\theta}$	10	2	$\dfrac{10}{9} \approx 1.1$	2

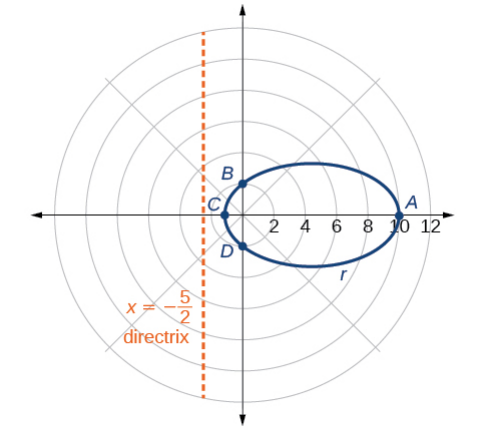


Figure 10.6.6

Analysis

We can check our result using a graphing utility. See Figure 10.6.7.

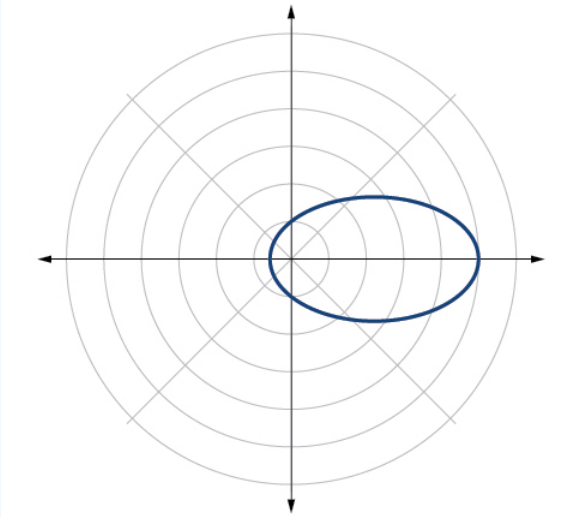


Figure 10.6.6: $r = \dfrac{10}{5 - 4\cos\theta}$ graphed on a viewing window of $[-3, 12, 1]$ by $[-4, 4, 1]$, $\theta min = 0$ and $\theta max = 2\pi$.

? Exercise 10.6.2

Graph $r = \dfrac{2}{4 - \cos\theta}$.

Answer

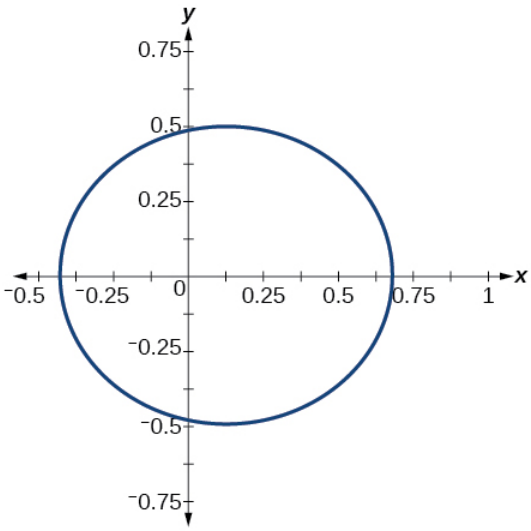


Figure 10.6.7

Defining Conics in Terms of a Focus and a Directrix

So far we have been using polar equations of conics to describe and graph the curve. Now we will work in reverse; we will use information about the origin, eccentricity, and directrix to determine the polar equation.

How to: Given the focus, eccentricity, and directrix of a conic, determine the polar equation

1. Determine whether the directrix is horizontal or vertical. If the directrix is given in terms of y, we use the general polar form in terms of sine. If the directrix is given in terms of x, we use the general polar form in terms of cosine.
2. Determine the sign in the denominator. If $p < 0$, use subtraction. If $p > 0$, use addition.
3. Write the coefficient of the trigonometric function as the given eccentricity.
4. Write the absolute value of p in the numerator, and simplify the equation.

Example $10.6.3A$: Finding the Polar Form of a Vertical Conic Given a Focus at the Origin and the Eccentricity and Directrix

Find the polar form of the conic given a focus at the origin, $e = 3$ and directrix $y = -2$.

Solution

The directrix is $y = -p$, so we know the trigonometric function in the denominator is sine.

Because $y = -2, -2 < 0$, so we know there is a subtraction sign in the denominator. We use the standard form of

$$r = \frac{ep}{1 - e\sin\theta}$$

and $e = 3$ and $|-2| = 2 = p$.

Therefore,

$$r = \frac{(3)(2)}{1 - 3\sin\theta}$$
$$r = \frac{6}{1 - 3\sin\theta}$$

Example $10.6.3B$: Finding the Polar Form of a Horizontal Conic Given a Focus at the Origin and the Eccentricity and Directrix

Find the polar form of a conic given a focus at the origin, $e = \frac{3}{5}$, and directrix $x = 4$.

Solution

Because the directrix is $x = p$, we know the function in the denominator is cosine. Because $x = 4$, $4 > 0$, so we know there is an addition sign in the denominator. We use the standard form of

$$r = \frac{ep}{1 + e\cos\theta}$$

and $e = \frac{3}{5}$ and $|4| = 4 = p$.

Therefore,

$$r = \frac{\left(\frac{3}{5}\right)(4)}{1 + \frac{3}{5}\cos\theta}$$

$$r = \frac{\frac{12}{5}}{1 + \frac{3}{5}\cos\theta}$$

$$r = \frac{\frac{12}{5}}{1\left(\frac{5}{5}\right) + \frac{3}{5}\cos\theta}$$

$$r = \frac{\frac{12}{5}}{\frac{5}{5} + \frac{3}{5}\cos\theta}$$

$$r = \frac{12}{5} \cdot \frac{5}{5 + 3\cos\theta}$$

$$r = \frac{12}{5 + 3\cos\theta}$$

? Exercise 10.6.3

Find the polar form of the conic given a focus at the origin, $e = 1$, and directrix $x = -1$.

Answer

$$r = \frac{1}{1 - \cos\theta}$$

✓ Example 10.6.4: Converting a Conic in Polar Form to Rectangular Form

Convert the conic $r = \frac{1}{5 - 5\sin\theta}$ to rectangular form.

Solution

We will rearrange the formula to use the identities $r = \sqrt{x^2 + y^2}$, $x = r\cos\theta$,and $y = r\sin\theta$.

$$
\begin{aligned}
r &= \frac{1}{5-5\sin\theta} \\
r\cdot(5-5\sin\theta) &= \frac{1}{5-5\sin\theta}\cdot(5-5\sin\theta) && \text{Eliminate the fraction.} \\
5r-5r\sin\theta &= 1 && \text{Distribute.} \\
5r &= 1+5r\sin\theta && \text{Isolate } 5r. \\
25r^2 &= (1+5r\sin\theta)^2 && \text{Square both sides.} \\
25(x^2+y^2) &= (1+5y)^2 && \text{Substitute } r=\sqrt{x^2+y^2} \text{ and } y=r\sin\theta. \\
25x^2+25y^2 &= 1+10y+25y^2 && \text{Distribute and use FOIL.} \\
25x^2-10y &= 1 && \text{Rearrange terms and set equal to 1.}
\end{aligned}
$$

Exercise 10.6.4

Convert the conic $r=\dfrac{2}{1+2\cos\theta}$ to rectangular form.

Answer

$4-8x+3x^2-y^2=0$

Media

Access these online resources for additional instruction and practice with conics in polar coordinates.

- Polar Equations of Conic Sections
- Graphing Polar Equations of Conics - 1
- Graphing Polar Equations of Conics - 2

Visit this website for additional practice questions from Learningpod.

Key Concepts

- Any conic may be determined by a single focus, the corresponding eccentricity, and the directrix. We can also define a conic in terms of a fixed point, the focus $P(r,\theta)$ at the pole, and a line, the directrix, which is perpendicular to the polar axis.
- A conic is the set of all points $e=\dfrac{PF}{PD}$, where eccentricity e is a positive real number. Each conic may be written in terms of its polar equation. See Example 10.6.1.
- The polar equations of conics can be graphed. See Example 10.6.2, Example 10.6.3, and Example 10.6.4.
- Conics can be defined in terms of a focus, a directrix, and eccentricity. See Example 10.6.5 and Example 10.6.6.
- We can use the identities $r=\sqrt{x^2+y^2}$, $x=r\cos\theta$, and $y=r\sin\theta$ to convert the equation for a conic from polar to rectangular form. See Example 10.6.7.

10.6: Conics in Polar Coordinates is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **12.5: Conic Sections in Polar Coordinates** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/precalculus.

CHAPTER OVERVIEW

11: Vectors and the Geometry of Space

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

11.1: Three-Dimensional Coordinate Systems
11.2: Vectors
11.3: The Dot Product
11.4: The Cross Product
11.5: Lines and Planes in Space
11.6: Cylinders and Quadric Surfaces

11: Vectors and the Geometry of Space is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

11.1: Three-Dimensional Coordinate Systems

In single-variable calculus, the functions that one encounters are functions of a variable (usually x or t) that varies over some subset of the real number line (which we denote by $\mathbb{R}$). For such a function, say, $y = f(x)$, the **graph** of the function f consists of the points $(x, y) = (x, f(x))$. These points lie in the **Euclidean plane**, which, in the **Cartesian** or **rectangular** coordinate system, consists of all ordered pairs of real numbers (a, b). We use the word ``Euclidean" to denote a system in which all the usual rules of Euclidean geometry hold. We denote the Euclidean plane by $\mathbb{R}^2$; the "2" represents the number of *dimensions* of the plane. The Euclidean plane has two perpendicular **coordinate axes**: the x-axis and the y-axis.

In vector (or multivariable) calculus, we will deal with functions of two or three variables (usually x, y or x, y, z, respectively). The graph of a function of two variables, say, $z = f(x, y)$, lies in **Euclidean space**, which in the Cartesian coordinate system consists of all ordered triples of real numbers (a, b, c). Since Euclidean space is 3-dimensional, we denote it by $\mathbb{R}^3$. The graph of f consists of the points $(x, y, z) = (x, y, f(x, y))$. The 3-dimensional coordinate system of Euclidean space can be represented on a flat surface, such as this page or a blackboard, only by giving the illusion of three dimensions, in the manner shown in Figure 11.1.1 . Euclidean space has three mutually perpendicular coordinate axes (x, y and z), and three mutually perpendicular coordinate planes\index{plane!coordinate}: the xy-plane, yz-plane and xz-plane (Figure 11.1.2).

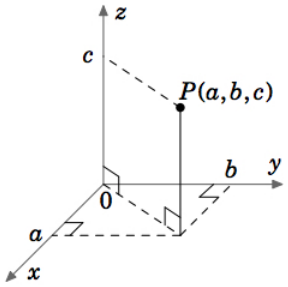


Figure 11.1.1

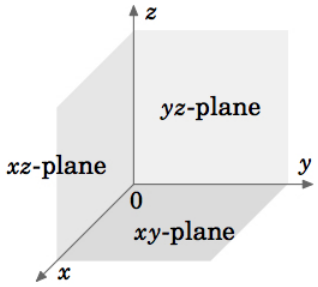


Figure 11.1.2

The coordinate system shown in Figure 11.1.1 is known as a **right-handed coordinate system**, because it is possible, using the right hand, to point the index finger in the positive direction of the x-axis, the middle finger in the positive direction of the y-axis, and the thumb in the positive direction of the z-axis, as in Figure 11.1.3

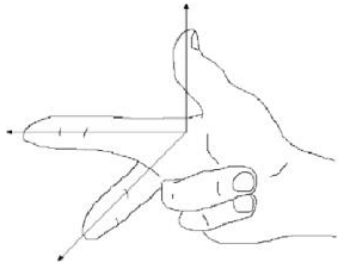


Fig 11.1.3 : *Right-handed coordinate system.*

An equivalent way of defining a right-handed system is if you can point your thumb upwards in the positive z-axis direction while using the remaining four fingers to rotate the x-axis towards the y-axis. Doing the same thing with the left hand is what defines a **left-handed coordinate system**. Notice that switching the x- and y-axes in a right-handed system results in a left-handed system, and that rotating either type of system does not change its ``handedness". Throughout the book we will use a right-handed system.

For functions of three variables, the graphs exist in 4-dimensional space (i.e. $\mathbb{R}^4$), which we can not see in our 3-dimensional space, let alone simulate in 2-dimensional space. So we can only think of 4-dimensional space abstractly. For an entertaining discussion of this subject, see the book by ABBOT.

So far, we have discussed the *position* of an object in 2-dimensional or 3-dimensional space. But what about something such as the velocity of the object, or its acceleration? Or the gravitational force acting on the object? These phenomena all seem to involve motion and *direction* in some way. This is where the idea of a *vector* comes in.

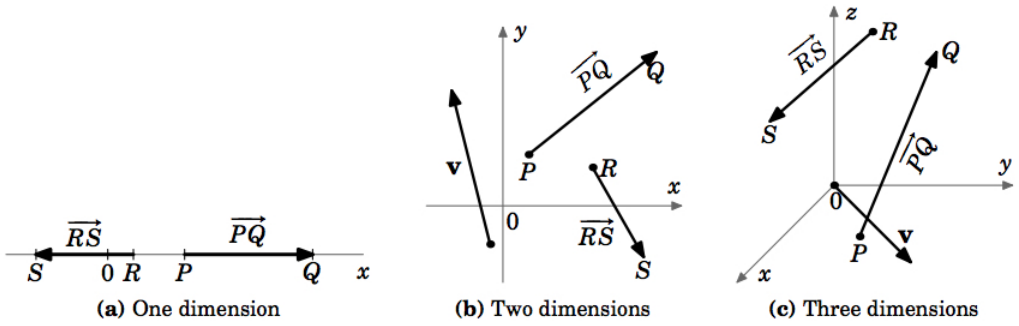
You have already dealt with velocity and acceleration in single-variable calculus. For example, for motion along a straight line, if $y = f(t)$ gives the displacement of an object after time t, then $dy/dt = f'(t)$ is the velocity of the object at time t. The derivative $f'(t)$ is just a number, which is positive if the object is moving in an agreed-upon "positive" direction, and negative if it moves in the opposite of that direction. So you can think of that number, which was called the velocity of the object, as having two components: a *magnitude*, indicated by a nonnegative number, preceded by a *direction*, indicated by a plus or minus symbol (representing motion in the positive direction or the negative direction, respectively), i.e. $f'(t) = \pm a$ for some number $a \geq 0$. Then a is the magnitude of the velocity (normally called the *speed* of the object), and the $\pm$ represents the direction of the velocity (though the $+$ is usually omitted for the positive direction).

For motion along a straight line, i.e. in a 1-dimensional space, the velocities are also contained in that 1-dimensional space, since they are just numbers. For general motion along a curve in 2- or 3-dimensional space, however, velocity will need to be represented by a multidimensional object which should have both a magnitude and a direction. A geometric object which has those features is an arrow, which in elementary geometry is called a ``directed line segment". This is the motivation for how we will define a vector.

Definition 11.1.1

A (nonzero) **vector** is a directed line segment drawn from a point P (called its **initial point**) to a point Q (called its **terminal point**), with P and Q being distinct points. The vector is denoted by $\overrightarrow{PQ}$. Its **magnitude** is the length of the line segment, denoted by $\|\overrightarrow{PQ}\|$, and its **direction** is the same as that of the directed line segment. The **zero vector** is just a point, and it is denoted by $\mathbf{0}$.

To indicate the direction of a vector, we draw an arrow from its initial point to its terminal point. We will often denote a vector by a single bold-faced letter (e.g. $\mathbf{v}$) and use the terms ``magnitude" and ``length" interchangeably. Note that our definition could apply to systems with any number of dimensions (Figure 1.1.4 (a)-(c)).



(a) One dimension **(b)** Two dimensions **(c)** Three dimensions

Figure 11.1.4 *Vectors in different dimensions*

A few things need to be noted about the zero vector. Our motivation for what a vector is included the notions of magnitude and direction. What is the magnitude of the zero vector? We define it to be zero, i.e. $\|\mathbf{0}\| = 0$. This agrees with the definition of the zero vector as just a point, which has zero length. What about the direction of the zero vector? A single point really has no well-defined direction. Notice that we were careful to only define the direction of a *nonzero* vector, which is well-defined since the initial and terminal points are distinct. Not everyone agrees on the direction of the zero vector. Some contend that the zero vector has *arbitrary* direction (i.e. can take any direction), some say that it has *indeterminate* direction (i.e. the direction cannot be determined), while others say that it has *no* direction. Our definition of the zero vector, however, does not require it to have a direction, and we will leave it at that.

Now that we know what a vector is, we need a way of determining when two vectors are equal. This leads us to the following definition.

Definition 11.1.2

Two nonzero vectors are **equal** if they have the same magnitude and the same direction. Any vector with zero magnitude is equal to the zero vector.

By this definition, vectors with the same magnitude and direction but with different initial points would be equal. For example, in Figure 1.1.5 the vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ all have the same magnitude $\sqrt{5}$ (by the Pythagorean Theorem). And we see that $\mathbf{u}$ and $\mathbf{w}$ are parallel, since they lie on lines having the same slope $\frac{1}{2}$, and they point in the same direction. So $\mathbf{u} = \mathbf{w}$, even though they have different initial points. We also see that $\mathbf{v}$ is parallel to $\mathbf{u}$ but points in the opposite direction. So $\mathbf{u} \neq \mathbf{v}$.

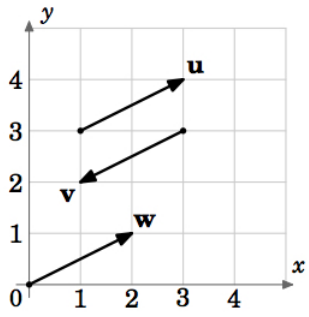


Figure 11.1.5

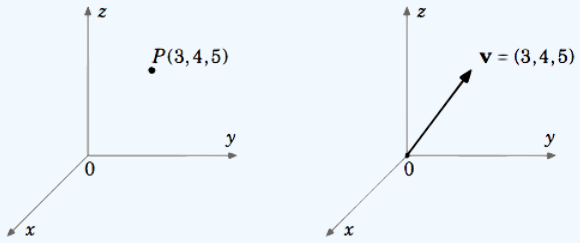
So we can see that there are an infinite number of vectors for a given magnitude and direction, those vectors all being equal and differing only by their initial and terminal points. Is there a single vector which we can choose to represent all those equal vectors? The answer is yes, and is suggested by the vector $\mathbf{w}$ in Figure 11.1.5

> Unless otherwise indicated, when speaking of "the vector" with a given magnitude and direction, we will mean the one whose initial point is at the origin of the coordinate system.

Thinking of vectors as starting from the origin provides a way of dealing with vectors in a standard way, since every coordinate system has an origin. But there will be times when it is convenient to consider a different initial point for a vector (for example, when adding vectors, which we will do in the next section). Another advantage of using the origin as the initial point is that it provides an easy correspondence between a vector and its terminal point.

Example 11.1.1

Let $\mathbf{v}$ be the vector in $\mathbb{R}^3$ whose initial point is at the origin and whose terminal point is $(3, 4, 5)$. Though the *point* $(3, 4, 5)$ and the vector $\mathbf{v}$ are different objects, it is convenient to write $\mathbf{v} = (3, 4, 5)$. When doing this, it is understood that the initial point of $\mathbf{v}$ is at the origin $(0, 0, 0)$ and the terminal point is $(3, 4, 5)$.



(a) The point (3,4,5) **(b)** The vector (3,4,5)

Figure 11.1.6 Correspondence between points and vectors

Unless otherwise stated, when we refer to vectors as $\mathbf{v} = (a, b)$ in $\mathbb{R}^2$ or $\mathbf{v} = (a, b, c)$ in $\mathbb{R}^3$, we mean vectors in Cartesian coordinates starting at the origin. Also, we will write the zero vector $\mathbf{0}$ in $\mathbb{R}^2$ and $\mathbb{R}^3$ as $(0, 0)$ and $(0, 0, 0)$, respectively.

The point-vector correspondence provides an easy way to check if two vectors are equal, without having to determine their magnitude and direction. Similar to seeing if two points are the same, you are now seeing if the terminal points of vectors starting at the origin are the same. For each vector, find the (unique!) vector it equals whose initial point is the origin. Then compare the coordinates of the terminal points of these ``new" vectors: if those coordinates are the same, then the original vectors are equal. To get the ``new" vectors starting at the origin, you *translate* each vector to start at the origin by subtracting the coordinates of the

original initial point from the original terminal point. The resulting point will be the terminal point of the ``new" vector whose initial point is the origin. Do this for each original vector then compare.

Example 11.1.2

Consider the vectors $\overrightarrow{PQ}$ and $\overrightarrow{RS}$ in $\mathbb{R}^3$, where $P=(2,1,5), Q=(3,5,7), R=(1,-3,-2)$ and $S=(2,1,0)$. Does $\overrightarrow{PQ}=\overrightarrow{RS}$?

The vector $\overrightarrow{PQ}$ is equal to the vector $\mathbf{v}$ with initial point $(0,0,0)$ and terminal point $Q-P=(3,5,7)-(2,1,5)=(3-2,5-1,7-5)=(1,4,2)$.

Similarly, $\overrightarrow{RS}$ is equal to the vector $\mathbf{w}$ with initial point $(0,0,0)$ and terminal point $S-R=(2,1,0)-(1,-3,-2)=(2-1,1-(-3),0-(-2))=(1,4,2)$.

So $\overrightarrow{PQ}=\mathbf{v}=(1,4,2)$ and $\overrightarrow{RS}=\mathbf{w}=(1,4,2)$.

$\therefore \overrightarrow{PQ}=\overrightarrow{RS}$

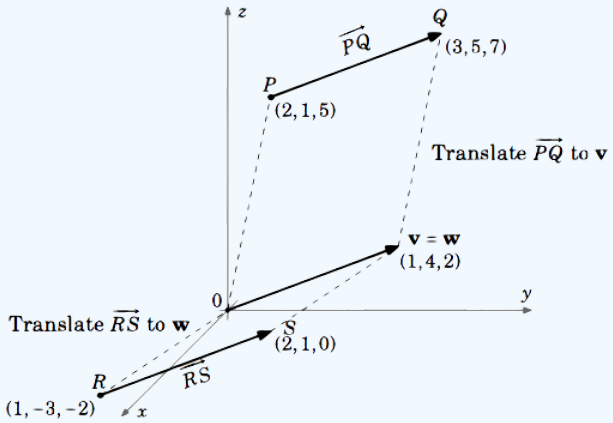


Figure 11.1.7

Recall the distance formula for points in the Euclidean plane:

For points $P=(x_1,y_1)$, $Q=(x_2,y_2)$ in $\mathbb{R}^2$, the distance d between P and Q is:

$$d=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2} \tag{11.1.1}$$

By this formula, we have the following result:

Note

For a vector $\overrightarrow{PQ}$ in $\mathbb{R}^2$ with initial point $P=(x_1,y_1)$ and terminal point $Q=(x_2,y_2)$, the magnitude of $\overrightarrow{PQ}$ is:

$$\|\overrightarrow{PQ}\|=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2} \tag{11.1.2}$$

Finding the magnitude of a vector $\mathbf{v}=(a,b)$ in $\mathbb{R}^2$ is a special case of the above formula with $P=(0,0)$ and $Q=(a,b)$:

For a vector $\mathbf{v}=(a,b)$ in $\mathbb{R}^2$, the magnitude of $\mathbf{v}$ is:

$$\|\mathbf{v}\|=\sqrt{a^2+b^2} \tag{11.1.3}$$

To calculate the magnitude of vectors in $\mathbb{R}^3$, we need a distance formula for points in Euclidean space (we will postpone the proof until the next section):

Theorem 11.1.1

The distance d between points $P=(x_1,y_1,z_1)$ and $Q=(x_2,y_2,z_2)$ in $\mathbb{R}^3$ is:

$$d=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2} \tag{11.1.4}$$

The proof will use the following result:

Theorem 11.1.2

For a vector $\mathbf{v}=(a,b,c)$ in $\mathbb{R}^3$, the magnitude of $\mathbf{v}$ is:

$$\|\mathbf{v}\|=\sqrt{a^2+b^2+c^2} \tag{11.1.5}$$

Proof: There are four cases to consider:

Case 1: $a=b=c=0$. Then $\mathbf{v}=\mathbf{0}$, so $\|\mathbf{v}\|=0=\sqrt{0^2+0^2+0^2}=\sqrt{a^2+b^2+c^2}$.

Case 2: exactly two of a,b,c are 0. Without loss of generality, we assume that $a=b=0$ and $c\neq 0$ (the other two possibilities are handled in a similar manner). Then $\mathbf{v}=(0,0,c)$, which is a vector of length $|c|$ along the z-axis. So $\|\mathbf{v}\|=|c|=\sqrt{c^2}=\sqrt{0^2+0^2+c^2}=\sqrt{a^2+b^2+c^2}$.

Case 3: exactly one of a,b,c is 0. Without loss of generality, we assume that $a=0$, $b\neq 0$ and $c\neq 0$ (the other two possibilities are handled in a similar manner). Then $\mathbf{v}=(0,b,c)$, which is a vector in the yz-plane, so by the Pythagorean Theorem we have $\|\mathbf{v}\|=\sqrt{b^2+c^2}=\sqrt{0^2+b^2+c^2}=\sqrt{a^2+b^2+c^2}$.

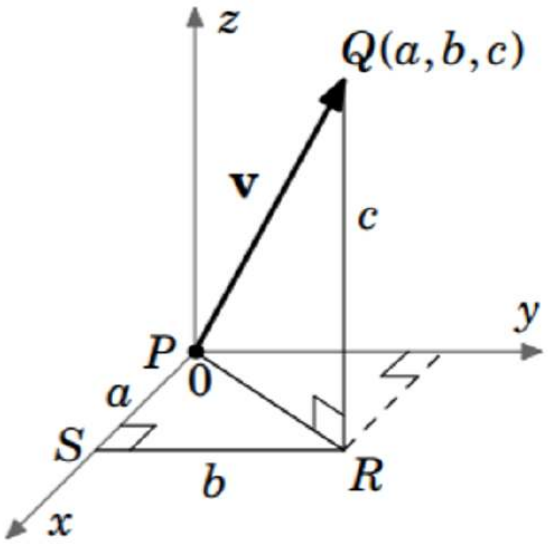


Figure 11.1.8

Case 4: none of a,b,c are 0. Without loss of generality, we can assume that a,b,c are all positive (the other seven possibilities are handled in a similar manner). Consider the points $P=(0,0,0)$, $Q=(a,b,c)$, $R=(a,b,0)$, and $S=(a,0,0)$, as shown in Figure 1.1.8. Applying the Pythagorean Theorem to the right triangle $\triangle PSR$ gives $|PR|^2=a^2+b^2$. A second application of the Pythagorean Theorem, this time to the right triangle $\triangle PQR$, gives $\|\mathbf{v}\|=|PQ|=\sqrt{|PR|^2+|QR|^2}=\sqrt{a^2+b^2+c^2}$. This proves the theorem.

(QED)

Example 1.3

Calculate the following:

1. The magnitude of the vector $\overrightarrow{PQ}$ in $\mathbb{R}^2$ with $P=(-1,2)$ and $Q=(5,5)$.
   Solution: By formula (1.2), $\|\overrightarrow{PQ}\| = \sqrt{(5-(-1))^2+(5-2)^2} = \sqrt{36+9} = \sqrt{45} = 3\sqrt{5}$.
2. The magnitude of the vector $\mathbf{v}=(8,3)$ in $\mathbb{R}^2$.
   Solution: By formula (1.3), $\|\mathbf{v}\| = \sqrt{8^2+3^2} = \sqrt{73}$.
3. The distance between the points $P=(2,-1,4)$ and $Q=(4,2,-3)$ in $\mathbb{R}^2$.
   Solution: By formula (1.4), the distance $d = \sqrt{(4-2)^2+(2-(-1))^2+(-3-4)^2} = \sqrt{4+9+49} = \sqrt{62}$.
4. The magnitude of the vector $\mathbf{v}=(5,8,-2)$ in $\mathbb{R}^3$.
   Solution: By formula (1.5), $\|\mathbf{v}\| = \sqrt{5^2+8^2+(-2)^2} = \sqrt{25+64+4} = \sqrt{93}$.

11.1: Three-Dimensional Coordinate Systems is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **1.1: Introduction** by Michael Corral is licensed GNU FDL. Original source: http://www.mecmath.net/.

11.2: Vectors

Learning Objectives

- Describe three-dimensional space mathematically.
- Locate points in space using coordinates.
- Write the distance formula in three dimensions.
- Write the equations for simple planes and spheres.
- Perform vector operations in $\mathbb{R}^3$.

Vectors are useful tools for solving two-dimensional problems. Life, however, happens in three dimensions. To expand the use of vectors to more realistic applications, it is necessary to create a framework for describing three-dimensional space. For example, although a two-dimensional map is a useful tool for navigating from one place to another, in some cases the topography of the land is important. Does your planned route go through the mountains? Do you have to cross a river? To appreciate fully the impact of these geographic features, you must use three dimensions. This section presents a natural extension of the two-dimensional Cartesian coordinate plane into three dimensions.

Three-Dimensional Coordinate Systems

As we have learned, the two-dimensional rectangular coordinate system contains two perpendicular axes: the horizontal x-axis and the vertical y-axis. We can add a third dimension, the z-axis, which is perpendicular to both the x-axis and the y-axis. We call this system the three-dimensional rectangular coordinate system. It represents the three dimensions we encounter in real life.

Definition: Three-dimensional Rectangular Coordinate System

The three-dimensional rectangular coordinate system consists of three perpendicular axes: the x-axis, the y-axis, and the z-axis. Because each axis is a number line representing all real numbers in $\mathbb{R}$, the three-dimensional system is often denoted by $\mathbb{R}^3$.

In Figure $11.2.1a$, the positive z-axis is shown above the plane containing the x- and y-axes. The positive x-axis appears to the left and the positive y-axis is to the right. A natural question to ask is: How was this arrangement determined? The system displayed follows the **right-hand rule**. If we take our right hand and align the fingers with the positive x-axis, then curl the fingers so they point in the direction of the positive y-axis, our thumb points in the direction of the positive z-axis (Figure $11.2.1b$). In this text, we always work with coordinate systems set up in accordance with the right-hand rule. Some systems do follow a left-hand rule, but the right-hand rule is considered the standard representation.

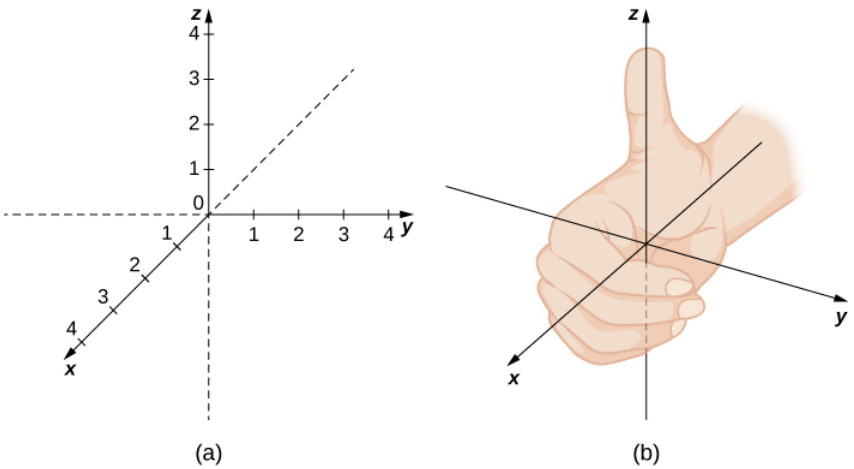


Figure 11.2.1: (a) We can extend the two-dimensional rectangular coordinate system by adding a third axis, the z-axis, that is perpendicular to both the x-axis and the y-axis. (b) The right-hand rule is used to determine the placement of the coordinate axes in the standard Cartesian plane.

In two dimensions, we describe a point in the plane with the coordinates (x, y). Each coordinate describes how the point aligns with the corresponding axis. In three dimensions, a new coordinate, z, is appended to indicate alignment with the z-axis: (x, y, z).

A point in space is identified by all three coordinates (Figure 11.2.2). To plot the point (x, y, z), go x units along the x-axis, then y units in the direction of the y-axis, then z units in the direction of the z-axis.

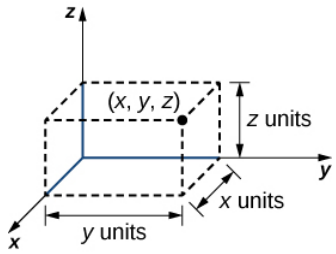


Figure 11.2.2: To plot the point (x, y, z) go x units along the x-axis, then y units in the direction of the y-axis, then z units in the direction of the z-axis.

Example 11.2.1: Locating Points in Space

Sketch the point $(1, -2, 3)$ in three-dimensional space.

Solution

To sketch a point, start by sketching three sides of a rectangular prism along the coordinate axes: one unit in the positive x direction, 2 units in the negative y direction, and 3 units in the positive z direction. Complete the prism to plot the point (Figure 11.2.3).

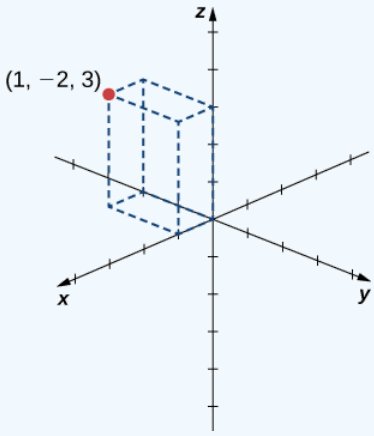


Figure 11.2.3: Sketching the point $(1, -2, 3)$.

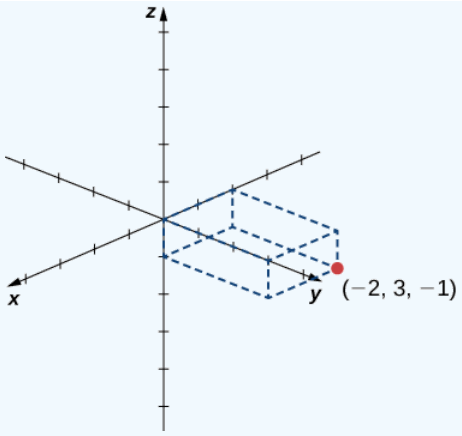
Exercise 11.2.1

Sketch the point $(-2, 3, -1)$ in three-dimensional space.

Hint

Start by sketching the coordinate axes. e.g., Figure 11.2.3 Then sketch a rectangular prism to help find the point in space.

Answer



In two-dimensional space, the coordinate plane is defined by a pair of perpendicular axes. These axes allow us to name any location within the plane. In three dimensions, we define **coordinate planes** by the coordinate axes, just as in two dimensions. There are three axes now, so there are three intersecting pairs of axes. Each pair of axes forms a coordinate plane: the xy-plane, the xz-plane, and the yz-plane (Figure 11.2.4). We define the xy-plane formally as the following set: $\{(x, y, 0) : x, y \in \mathbb{R}\}$. Similarly, the xz-plane and the yz-plane are defined as $\{(x, 0, z) : x, z \in \mathbb{R}\}$ and $\{(0, y, z) : y, z \in \mathbb{R}\}$, respectively.

To visualize this, imagine you're building a house and are standing in a room with only two of the four walls finished. (Assume the two finished walls are adjacent to each other.) If you stand with your back to the corner where the two finished walls meet, facing out into the room, the floor is the xy-plane, the wall to your right is the xz-plane, and the wall to your left is the yz-plane.

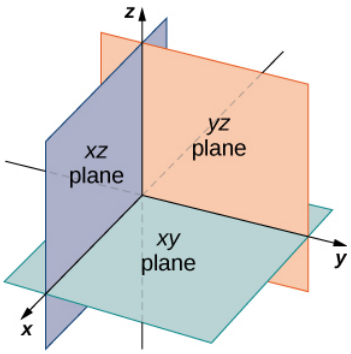


Figure 11.2.4: The plane containing the x- *and* y-axes is called the xy-plane. The plane containing the x- and z-axes is called the xz-plane, and the y- and z-axes define the yz-plane.

In two dimensions, the coordinate axes partition the plane into four quadrants. Similarly, the coordinate planes divide space between them into eight regions about the origin, called **octants.** The octants fill $\mathbb{R}^3$ in the same way that quadrants fill $\mathbb{R}^2$, as shown in Figure 11.2.5.

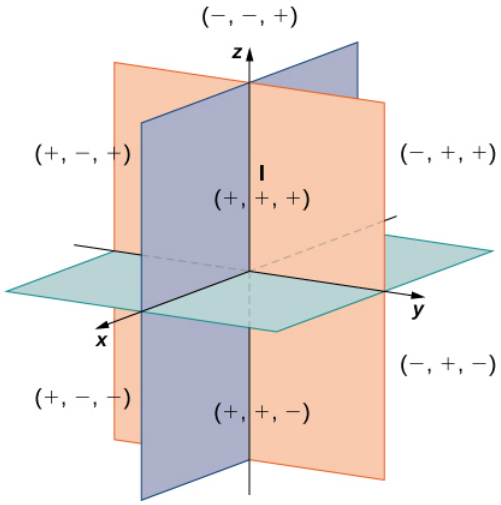


Figure 11.2.5: Points that lie in octants have three nonzero coordinates.

Most work in three-dimensional space is a comfortable extension of the corresponding concepts in two dimensions. In this section, we use our knowledge of circles to describe spheres, then we expand our understanding of vectors to three dimensions. To accomplish these goals, we begin by adapting the distance formula to three-dimensional space.

If two points lie in the same coordinate plane, then it is straightforward to calculate the distance between them. We know that the distance d between two points (x_1, y_1) and (x_2, y_2) in the xy-coordinate plane is given by the formula

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}\,.$$

The formula for the distance between two points in space is a natural extension of this formula.

The Distance between Two Points in Space

The distance d between points (x_1, y_1, z_1) and (x_2, y_2, z_2) is given by the formula

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}\,. \tag{11.2.1}$$

The proof of this theorem is left as an exercise. (Hint: First find the distance d_1 between the points (x_1, y_1, z_1) and (x_2, y_2, z_1) as shown in Figure 11.2.6)

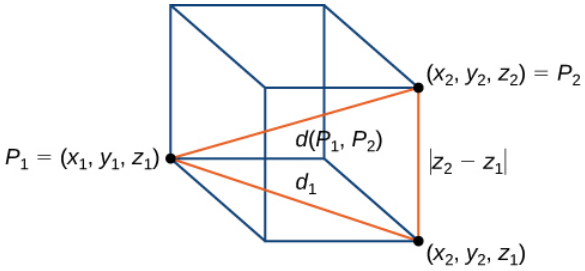


Figure 11.2.6: The distance between P_1 and P_2 is the length of the diagonal of the rectangular prism having P_1 and P_2 as opposite corners.

Example 11.2.2: Distance in Space

Find the distance between points $P_1 = (3, -1, 5)$ and $P_2 = (2, 1, -1)$.

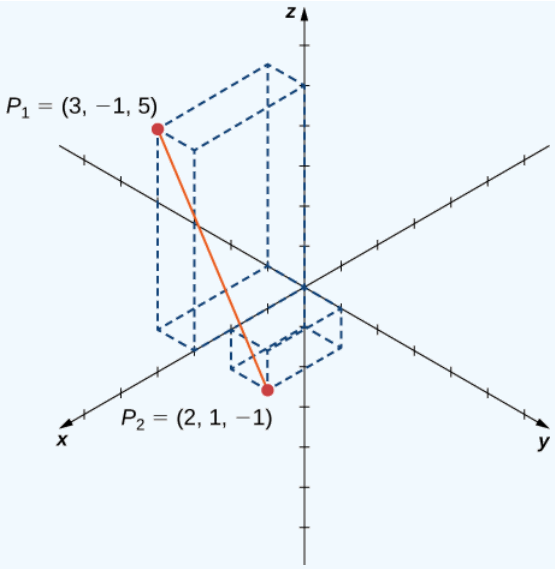


Figure 11.2.7: Find the distance between the two points.

Solution

Substitute values directly into the distance formula (Equation 11.2.1):

$$\begin{aligned} d(P_1, P_2) &= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \\ &= \sqrt{(2-3)^2 + (1-(-1))^2 + (-1-5)^2} \\ &= \sqrt{(-1)^2 + 2^2 + (-6)^2} \\ &= \sqrt{41}. \end{aligned}$$

? Exercise 11.2.2

Find the distance between points $P_1 = (1, -5, 4)$ and $P_2 = (4, -1, -1)$.

Hint

$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$

Answer

$$5\sqrt{2}$$

Before moving on to the next section, let's get a feel for how $\mathbb{R}^3$ differs from $\mathbb{R}^2$. For example, in $\mathbb{R}^2$, lines that are not parallel must always intersect. This is not the case in $\mathbb{R}^3$. For example, consider the lines shown in Figure 11.2.8 These two lines are not parallel, nor do they intersect.

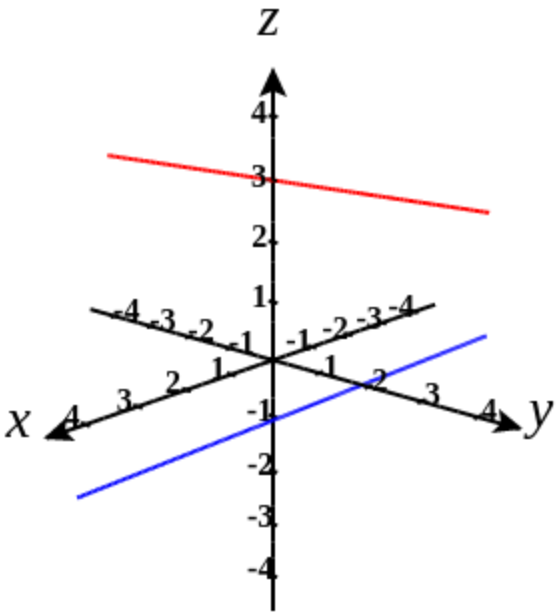


Figure 11.2.8: These two lines are not parallel, but still do not intersect.

You can also have circles that are interconnected but have no points in common, as in Figure 11.2.9.

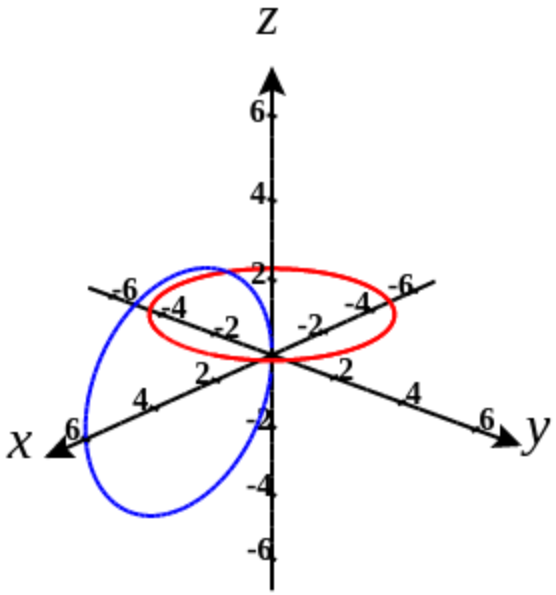


Figure 11.2.9: These circles are interconnected, but have no points in common.

We have a lot more flexibility working in three dimensions than we do if we stuck with only two dimensions.

Writing Equations in $\mathbb{R}^3$

Now that we can represent points in space and find the distance between them, we can learn how to write equations of geometric objects such as lines, planes, and curved surfaces in $\mathbb{R}^3$. First, we start with a simple equation. Compare the graphs of the equation $x = 0$ in $\mathbb{R}$, $\mathbb{R}^2$,and $\mathbb{R}^3$ (Figure 11.2.10). From these graphs, we can see the same equation can describe a point, a line, or a plane.

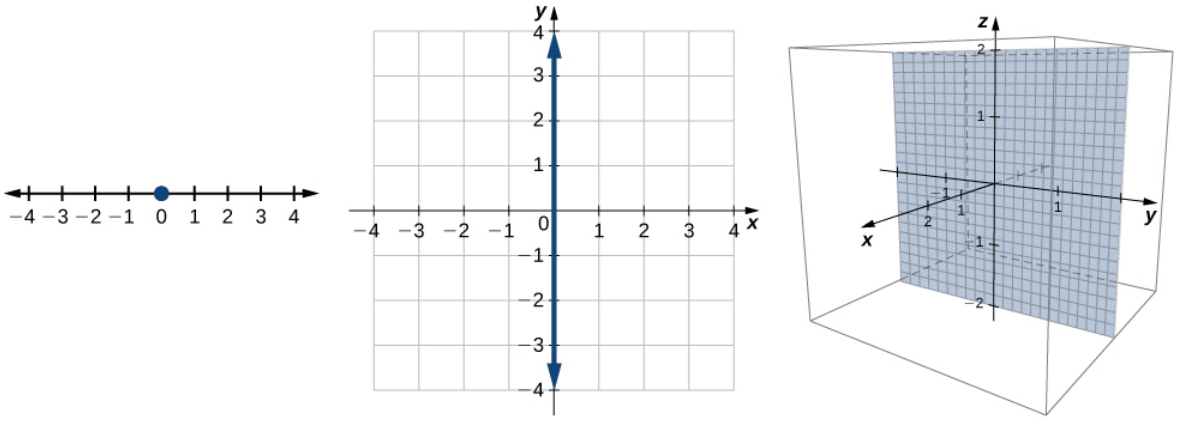


Figure 11.2.10: (a) In $\mathbb{R}$, the equation $x = 0$ describes a single point. (b) In $\mathbb{R}^2$, the equation $x = 0$ describes a line, the y-axis. (c) In $\mathbb{R}^3$, the equation $x = 0$ describes a plane, the yz-plane.

In space, the equation $x = 0$ describes all points $(0, y, z)$. This equation defines the yz-plane. Similarly, the xy-plane contains all points of the form $(x, y, 0)$. The equation $z = 0$ defines the xy-plane and the equation $y = 0$ describes the xz-plane (Figure 11.2.11).

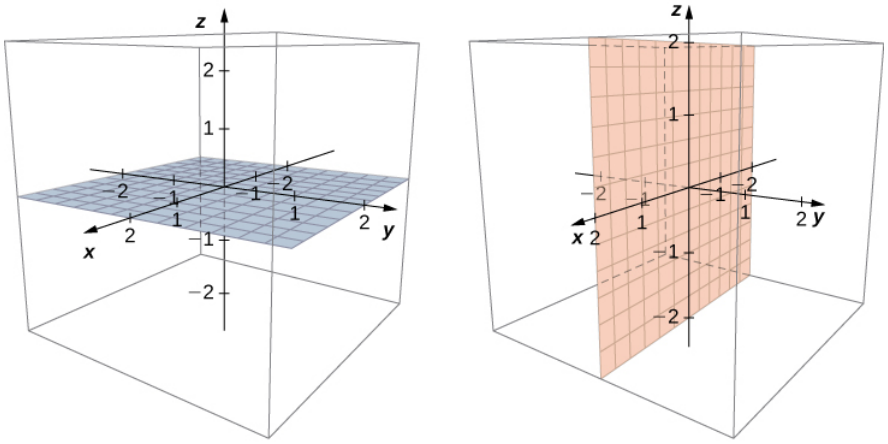


Figure 11.2.11: (a) In space, the equation $z = 0$ describes the xy-plane. (b) All points in the xz-plane satisfy the equation $y = 0$.

Understanding the equations of the coordinate planes allows us to write an equation for any plane that is parallel to one of the coordinate planes. When a plane is parallel to the xy-plane, for example, the z-coordinate of each point in the plane has the same constant value. Only the x- and y-coordinates of points in that plane vary from point to point.

Equations of Planes Parallel to Coordinate Planes

1. The plane in space that is parallel to the xy-plane and contains point (a, b, c) can be represented by the equation $z = c$.
2. The plane in space that is parallel to the xz-plane and contains point (a, b, c) can be represented by the equation $y = b$.
3. The plane in space that is parallel to the yz-plane and contains point (a, b, c) can be represented by the equation $x = a$.

Example 11.2.3: Writing Equations of Planes Parallel to Coordinate Planes

a. Write an equation of the plane passing through point $(3, 11, 7)$ that is parallel to the yz-plane.
b. Find an equation of the plane passing through points $(6, -2, 9)$, $(0, -2, 4)$,and $(1, -2, -3)$.

Solution

a. When a plane is parallel to the yz-plane, only the y- and z-coordinates may vary. The x-coordinate has the same constant value for all points in this plane, so this plane can be represented by the equation $x = 3$.
b. Each of the points $(6, -2, 9)$, $(0, -2, 4)$,and $(1, -2, -3)$ has the same y-coordinate. This plane can be represented by the equation $y = -2$.

Exercise 11.2.3

Write an equation of the plane passing through point $(1, -6, -4)$ that is parallel to the xy-plane.

Hint

If a plane is parallel to the xy-plane, the z-coordinates of the points in that plane do not vary.

Answer

$$z = -4$$

As we have seen, in $\mathbb{R}^2$ the equation $x = 5$ describes the vertical line passing through point $(5, 0)$. This line is parallel to the y-axis. In a natural extension, the equation $x = 5$ in $\mathbb{R}^3$ describes the plane passing through point $(5, 0, 0)$, which is parallel to the yz-plane. Another natural extension of a familiar equation is found in the equation of a sphere.

Definition: Sphere

A sphere is the set of all points in space equidistant from a fixed point, the center of the sphere (Figure 11.2.12), just as the set of all points in a plane that are equidistant from the center represents a circle. In a sphere, as in a circle, the distance from the center to a point on the sphere is called the *radius.*

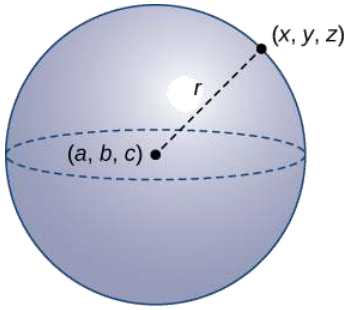


Figure 11.2.12: Each point (x, y, z) on the surface of a sphere is r units away from the center (a, b, c).

The equation of a circle is derived using the distance formula in two dimensions. In the same way, the equation of a sphere is based on the three-dimensional formula for distance.

Standard Equation of a Sphere

The sphere with center (a, b, c) and radius r can be represented by the equation

$$(x-a)^2+(y-b)^2+(z-c)^2=r^2.$$

This equation is known as the *standard equation of a sphere*.

Example 11.2.4: Finding an Equation of a Sphere

Find the standard equation of the sphere with center $(10, 7, 4)$ and point $(-1, 3, -2)$, as shown in Figure 11.2.13

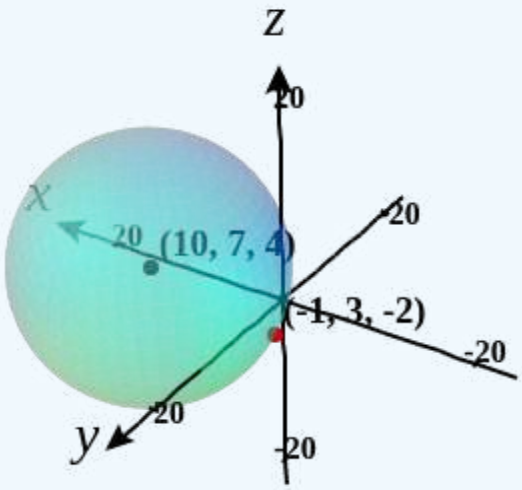


Figure 11.2.13 The sphere centered at $(10, 7, 4)$ containing point $(-1, 3, -2)$.

Solution

Use the distance formula to find the radius r of the sphere:

$$\begin{aligned} r &= \sqrt{(-1-10)^2+(3-7)^2+(-2-4)^2} \\ &= \sqrt{(-11)^2+(-4)^2+(-6)^2} \\ &= \sqrt{173} \end{aligned}$$

The standard equation of the sphere is

$$(x-10)^2+(y-7)^2+(z-4)^2 = 173.$$

? Exercise 11.2.4

Find the standard equation of the sphere with center $(-2, 4, -5)$ containing point $(4, 4, -1)$.

Hint

First use the distance formula to find the radius of the sphere.

Answer

$$(x+2)^2+(y-4)^2+(z+5)^2 = 52$$

Example 11.2.5: Finding the Equation of a Sphere

Let $P=(-5,2,3)$ and $Q=(3,4,-1)$, and suppose line segment $\overline{PQ}$ forms the diameter of a sphere (Figure 11.2.14). Find the equation of the sphere.

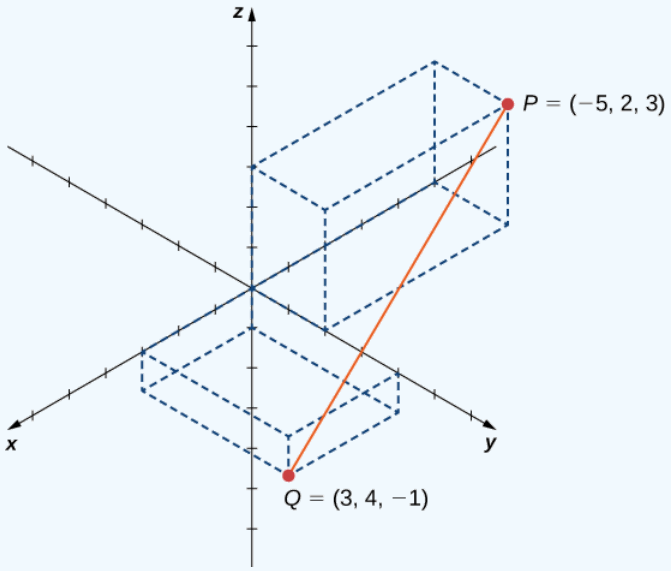


Figure 11.2.14: Line segment $\overline{PQ}$.

Solution:

Since $\overline{PQ}$ is a diameter of the sphere, we know the center of the sphere is the midpoint of $\overline{PQ}$.Then,

$$C=\left(\frac{-5+3}{2},\frac{2+4}{2},\frac{3+(-1)}{2}\right)=(-1,3,1).$$

Furthermore, we know the radius of the sphere is half the length of the diameter. This gives

$$\begin{aligned} r &= \frac{1}{2}\sqrt{(-5-3)^2+(2-4)^2+(3-(-1))^2} \\ &= \frac{1}{2}\sqrt{64+4+16} \\ &= \sqrt{21} \end{aligned}$$

Then, the equation of the sphere is $(x+1)^2+(y-3)^2+(z-1)^2=21$.

Exercise 11.2.5

Find the equation of the sphere with diameter $\overline{PQ}$, where $P=(2,-1,-3)$ and $Q=(-2,5,-1)$.

Hint

Find the midpoint of the diameter first.

Answer

$$x^2+(y-2)^2+(z+2)^2=14$$

Example 11.2.6: Graphing Other Equations in Three Dimensions

Describe the set of points that satisfies $(x-4)(z-2)=0$, and graph the set.

Solution

We must have either $x-4=0$ or $z-2=0$, so the set of points forms the two planes $x=4$ and $z=2$ (Figure 11.2.15).

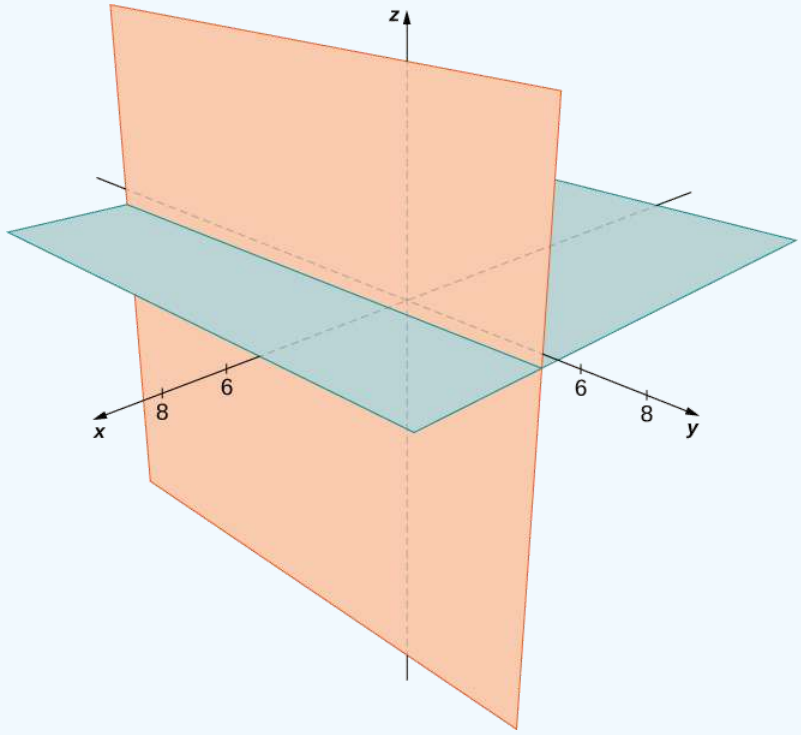


Figure 11.2.15: The set of points satisfying $(x-4)(z-2)=0$ forms the two planes $x=4$ and $z=2$.

Exercise 11.2.6

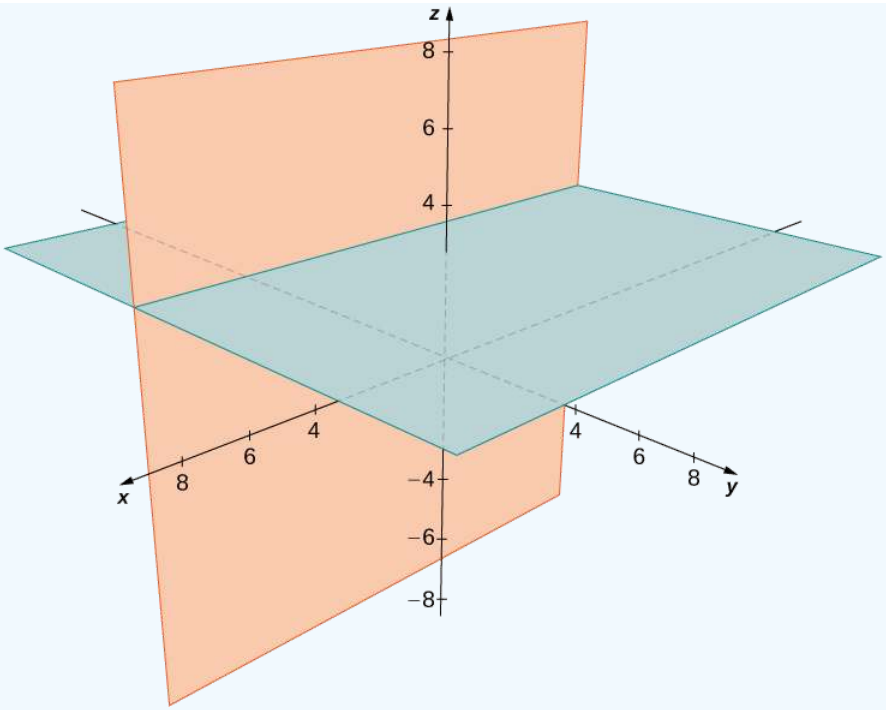
Describe the set of points that satisfies $(y+2)(z-3)=0,$ and graph the set.

Hint

One of the factors must be zero.

Answer

The set of points forms the two planes $y=-2$ and $z=3$.



Example 11.2.7: Graphing Other Equations in Three Dimensions

Describe the set of points in three-dimensional space that satisfies $(x-2)^2+(y-1)^2=4$, and graph the set.

Solution

The x- and y-coordinates form a circle in the xy-plane of radius 2, centered at $(2,1)$. Since there is no restriction on the z-coordinate, the three-dimensional result is a circular cylinder of radius 2 centered on the line with $x=2$ and $y=1$. The cylinder extends indefinitely in the z-direction (Figure 11.2.16).

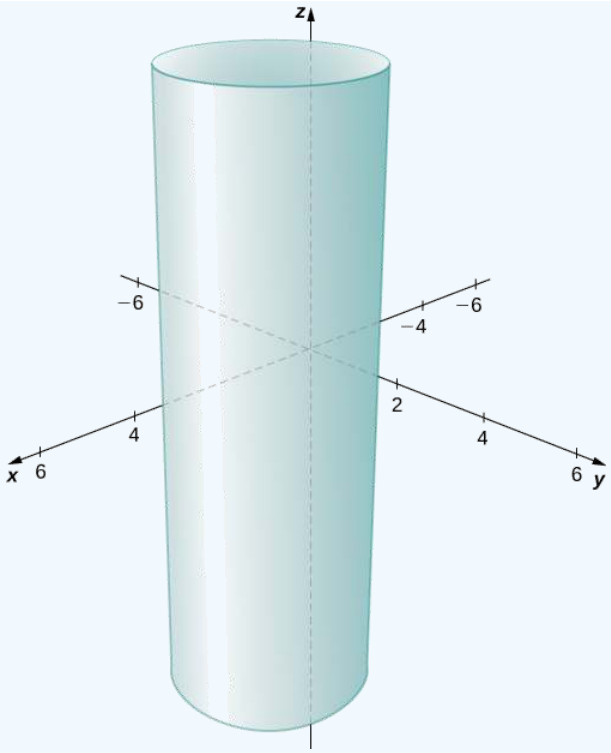


Figure 11.2.16: The set of points satisfying $(x-2)^2+(y-1)^2=4$. This is a cylinder of radius 2 centered on the line with $x=2$ and $y=1$.

Exercise 11.2.7

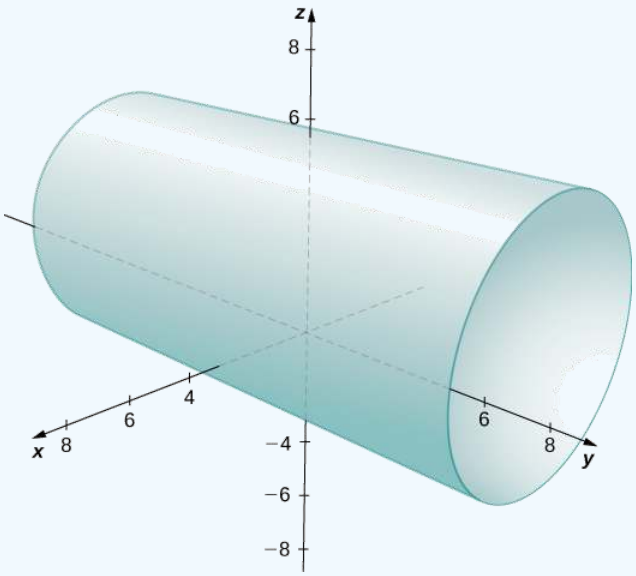
Describe the set of points in three dimensional space that satisfies $x^2+(z-2)^2=16$, and graph the surface.

Hint

Think about what happens if you plot this equation in two dimensions in the xz-plane.

Answer

A cylinder of radius 4 centered on the line with $x=0$ and $z=2$.



Working with Vectors in $\mathbb{R}^3$

Just like two-dimensional vectors, three-dimensional vectors are quantities with both magnitude and direction, and they are represented by directed line segments (arrows). With a three-dimensional vector, we use a three-dimensional arrow.

Three-dimensional vectors can also be represented in component form. The notation $\vec{\mathbf{v}} = \langle x, y, z \rangle$ is a natural extension of the two-dimensional case, representing a vector with the initial point at the origin, $(0, 0, 0)$, and terminal point (x, y, z). The zero vector is $\vec{\mathbf{0}} = \langle 0, 0, 0 \rangle$. So, for example, the three dimensional vector $\vec{\mathbf{v}} = \langle 2, 4, 1 \rangle$ is represented by a directed line segment from point $(0, 0, 0)$ to point $(2, 4, 1)$ (Figure 11.2.17).

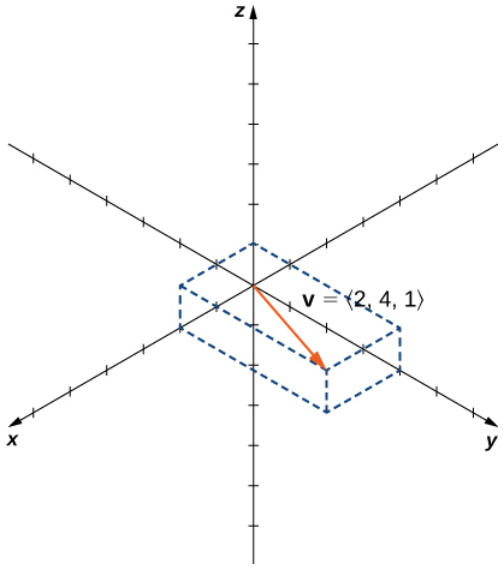


Figure 11.2.17: Vector $\vec{\mathbf{v}} = \langle 2, 4, 1 \rangle$ is represented by a directed line segment from point $(0, 0, 0)$ to point $(2, 4, 1)$.

Vector addition and scalar multiplication are defined analogously to the two-dimensional case. If $\vec{\mathbf{v}} = \langle x_1, y_1, z_1 \rangle$ and $\vec{\mathbf{w}} = \langle x_2, y_2, z_2 \rangle$ are vectors, and k is a scalar, then

$$\vec{\mathbf{v}} + \vec{\mathbf{w}} = \langle x_1 + x_2, y_1 + y_2, z_1 + z_2 \rangle$$

and

$$k\vec{\mathbf{v}} = \langle kx_1, ky_1, kz_1 \rangle.$$

If $k = -1$, then $k\vec{\mathbf{v}} = (-1)\vec{\mathbf{v}}$ is written as $-\vec{\mathbf{v}}$, and vector subtraction is defined by $\vec{\mathbf{v}} - \vec{\mathbf{w}} = \vec{\mathbf{v}} + (-\vec{\mathbf{w}}) = \vec{\mathbf{v}} + (-1)\vec{\mathbf{w}}$.

The standard unit vectors extend easily into three dimensions as well, $\hat{\mathbf{i}} = \langle 1, 0, 0 \rangle$, $\hat{\mathbf{j}} = \langle 0, 1, 0 \rangle$, and $\hat{\mathbf{k}} = \langle 0, 0, 1 \rangle$, and we use them in the same way we used the standard unit vectors in two dimensions. Thus, we can represent a vector in $\mathbb{R}^3$ in the following ways:

$$\vec{\mathbf{v}} = \langle x, y, z \rangle = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$$

.

Example 11.2.8: Vector Representations

Let $\overrightarrow{PQ}$ be the vector with initial point $P = (3, 12, 6)$ and terminal point $Q = (-4, -3, 2)$ as shown in Figure 11.2.18 Express $\overrightarrow{PQ}$ in both component form and using standard unit vectors.

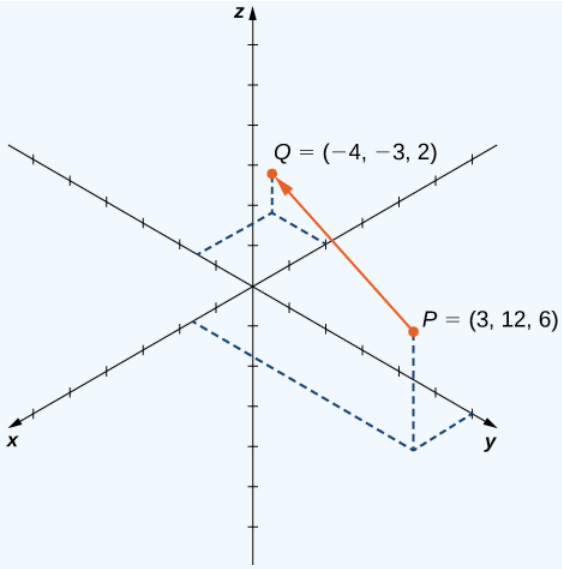


Figure 11.2.18: The vector with initial point $P=(3,12,6)$ and terminal point $Q=(-4,-3,2)$.

Solution

In component form,

$$\begin{aligned}\overrightarrow{PQ} &= \langle x_2-x_1, y_2-y_1, z_2-z_1\rangle \\ &= \langle -4-3, -3-12, 2-6\rangle \\ &= \langle -7, -15, -4\rangle.\end{aligned}$$

In standard unit form,

$$\overrightarrow{PQ} = -7\hat{\mathbf{i}} - 15\hat{\mathbf{j}} - 4\hat{\mathbf{k}}.$$

? Exercise 11.2.8

Let $S=(3,8,2)$ and $T=(2,-1,3)$. Express $\overrightarrow{ST}$ in component form and in standard unit form.

Hint

Write $\overrightarrow{ST}$ in component form first. T is the terminal point of $\overrightarrow{ST}$.

Answer

$\overrightarrow{ST} = \langle -1,-9,1\rangle = -\hat{\mathbf{i}} - 9\hat{\mathbf{j}} + \hat{\mathbf{k}}$

As described earlier, vectors in three dimensions behave in the same way as vectors in a plane. The geometric interpretation of vector addition, for example, is the same in both two- and three-dimensional space (Figure 11.2.19).

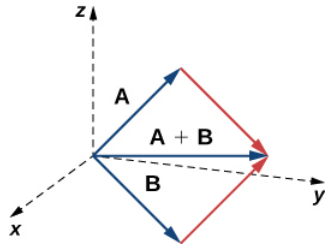


Figure 11.2.19: To add vectors in three dimensions, we follow the same procedures we learned for two dimensions.

We have already seen how some of the algebraic properties of vectors, such as vector addition and scalar multiplication, can be extended to three dimensions. Other properties can be extended in similar fashion. They are summarized here for our reference.

Properties of Vectors in Space

Let $\vec{\mathbf{v}} = \langle x_1, y_1, z_1 \rangle$ and $\vec{\mathbf{w}} = \langle x_2, y_2, z_2 \rangle$ be vectors, and let k be a scalar.

- **Scalar multiplication**:

$$k\vec{\mathbf{v}} = \langle kx_1, ky_1, kz_1 \rangle$$

- **Vector addition**:

$$\vec{\mathbf{v}} + \vec{\mathbf{w}} = \langle x_1, y_1, z_1 \rangle + \langle x_2, y_2, z_2 \rangle = \langle x_1 + x_2, y_1 + y_2, z_1 + z_2 \rangle$$

- **Vector subtraction**:

$$\vec{\mathbf{v}} - \vec{\mathbf{w}} = \langle x_1, y_1, z_1 \rangle - \langle x_2, y_2, z_2 \rangle = \langle x_1 - x_2, y_1 - y_2, z_1 - z_2 \rangle$$

- **Vector magnitude**:

$$\|\vec{\mathbf{v}}\| = \sqrt{x_1^2 + y_1^2 + z_1^2}$$

- **Unit vector in the direction of $\vec{\mathbf{v}}$**:

$$\frac{1}{\|\vec{\mathbf{v}}\|}\vec{\mathbf{v}} = \frac{1}{\|\vec{\mathbf{v}}\|}\langle x_1, y_1, z_1 \rangle = \langle \frac{x_1}{\|\vec{\mathbf{v}}\|}, \frac{y_1}{\|\vec{\mathbf{v}}\|}, \frac{z_1}{\|\vec{\mathbf{v}}\|} \rangle, \quad \text{if } \vec{\mathbf{v}} \neq \vec{\mathbf{0}}$$

We have seen that vector addition in two dimensions satisfies the commutative, associative, and additive inverse properties. These properties of vector operations are valid for three-dimensional vectors as well. Scalar multiplication of vectors satisfies the distributive property, and the zero vector acts as an additive identity. The proofs to verify these properties in three dimensions are straightforward extensions of the proofs in two dimensions.

Example 11.2.9: Vector Operations in Three Dimensions

Let $\vec{\mathbf{v}} = \langle -2, 9, 5 \rangle$ and $\vec{\mathbf{w}} = \langle 1, -1, 0 \rangle$ (Figure 11.2.20). Find the following vectors.

a. $3\vec{\mathbf{v}} - 2\vec{\mathbf{w}}$
b. $5\|\vec{\mathbf{w}}\|$
c. $\|5\vec{\mathbf{w}}\|$
d. A unit vector in the direction of $\vec{\mathbf{v}}$

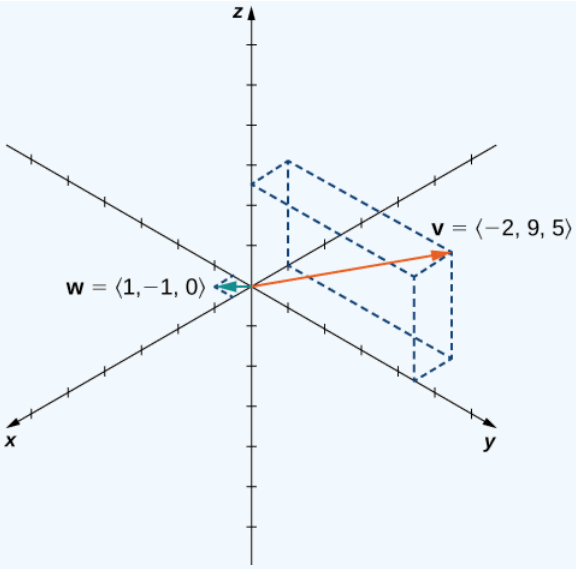


Figure 11.2.20: The vectors $\vec{\mathbf{v}} = \langle -2, 9, 5 \rangle$ and $\vec{\mathbf{w}} = \langle 1, -1, 0 \rangle$.

Solution

a. First, use scalar multiplication of each vector, then subtract:

$$\begin{aligned} 3\vec{\mathbf{v}} - 2\vec{\mathbf{w}} &= 3\langle -2, 9, 5 \rangle - 2\langle 1, -1, 0 \rangle \\ &= \langle -6, 27, 15 \rangle - \langle 2, -2, 0 \rangle \\ &= \langle -6 - 2, 27 - (-2), 15 - 0 \rangle \\ &= \langle -8, 29, 15 \rangle. \end{aligned}$$

b. Write the equation for the magnitude of the vector, then use scalar multiplication:

$$5\|\vec{\mathbf{w}}\| = 5\sqrt{1^2 + (-1)^2 + 0^2} = 5\sqrt{2}.$$

c. First, use scalar multiplication, then find the magnitude of the new vector. Note that the result is the same as for part b.:

$$\|5\vec{\mathbf{w}}\| = \| \langle 5, -5, 0 \rangle \| = \sqrt{5^2 + (-5)^2 + 0^2} = \sqrt{50} = 5\sqrt{2}$$

d. Recall that to find a unit vector in two dimensions, we divide a vector by its magnitude. The procedure is the same in three dimensions:

$$\begin{aligned} \frac{\vec{\mathbf{v}}}{\|\vec{\mathbf{v}}\|} &= \frac{1}{\|\vec{\mathbf{v}}\|}\langle -2, 9, 5 \rangle \\ &= \frac{1}{\sqrt{(-2)^2 + 9^2 + 5^2}}\langle -2, 9, 5 \rangle \\ &= \frac{1}{\sqrt{110}}\langle -2, 9, 5 \rangle \\ &= \left\langle \frac{-2}{\sqrt{110}}, \frac{9}{\sqrt{110}}, \frac{5}{\sqrt{110}} \right\rangle. \end{aligned}$$

? Exercise 11.2.9:

Let $\vec{\mathbf{v}} = \langle -1, -1, 1 \rangle$ and $\vec{\mathbf{w}} = \langle 2, 0, 1 \rangle$. Find a unit vector in the direction of $5\vec{\mathbf{v}} + 3\vec{\mathbf{w}}$.

Hint

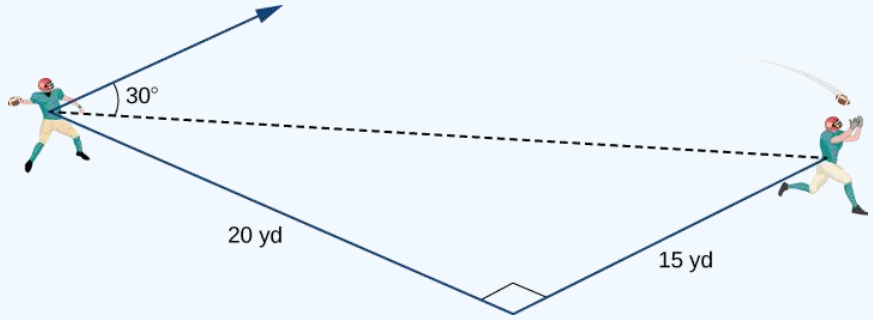
Start by writing $5\vec{\mathbf{v}}+3\vec{\mathbf{w}}$ in component form.

Answer

$$\left\langle \frac{1}{3\sqrt{10}}, -\frac{5}{3\sqrt{10}}, \frac{8}{3\sqrt{10}} \right\rangle$$

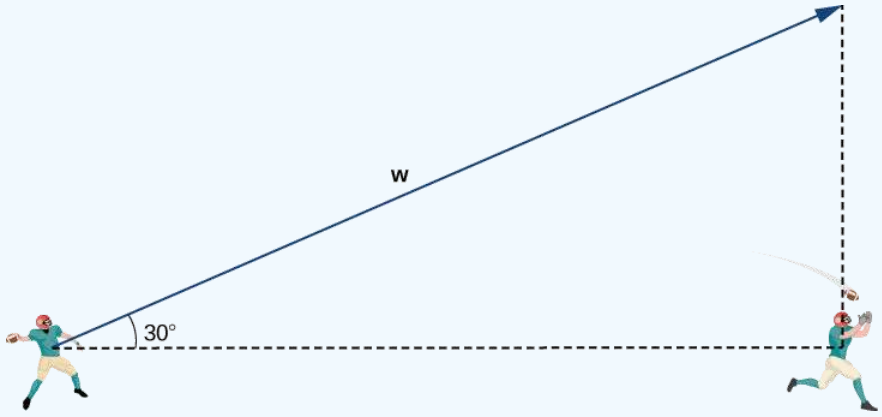
Example 11.2.10: Throwing a Forward Pass

A quarterback is standing on the football field preparing to throw a pass. His receiver is standing 20 yd down the field and 15 yd to the quarterback's left. The quarterback throws the ball at a velocity of 60 mph toward the receiver at an upward angle of 30° (see the following figure). Write the initial velocity vector of the ball, $\vec{\mathbf{v}}$, in component form.



Solution

The first thing we want to do is find a vector in the same direction as the velocity vector of the ball. We then scale the vector appropriately so that it has the right magnitude. Consider the vector $\vec{\mathbf{w}}$ extending from the quarterback's arm to a point directly above the receiver's head at an angle of 30° (see the following figure). This vector would have the same direction as $\vec{\mathbf{v}}$, but it may not have the right magnitude.



The receiver is 20 yd down the field and 15 yd to the quarterback's left. Therefore, the straight-line distance from the quarterback to the receiver is

$$\textbf{Dist from QB to receiver} = \sqrt{15^2+20^2} = \sqrt{225+400} = \sqrt{625} = 25 \text{ yd.}$$

We have $\dfrac{25}{\|\vec{\mathbf{w}}\|} = \cos 30^\circ$. Then the magnitude of $\vec{\mathbf{w}}$ is given by

$$\|\vec{\mathbf{w}}\| = \frac{25}{\cos 30^\circ} = \frac{25 \cdot 2}{\sqrt{3}} = \frac{50}{\sqrt{3}} \text{ yd}$$

and the vertical distance from the receiver to the terminal point of $\vec{\mathbf{w}}$ is

$$\textbf{Vert dist from receiver to terminal point of } \vec{\mathbf{w}} = \|\vec{\mathbf{w}}\| \sin 30^\circ = \frac{50}{\sqrt{3}} \cdot \frac{1}{2} = \frac{25}{\sqrt{3}} \text{ yd.}$$

Then $\vec{\mathbf{w}} = \langle 20, 15, \frac{25}{\sqrt{3}} \rangle$, and has the same direction as $\vec{\mathbf{v}}$.

Recall, though, that we calculated the magnitude of $\vec{\mathbf{w}}$ to be $\|\vec{\mathbf{w}}\| = \frac{50}{\sqrt{3}}$ yd, and $\vec{\mathbf{v}}$ has magnitude 60 mph. So, we need to multiply vector $\vec{\mathbf{w}}$ by an appropriate constant, k. We want to find a value of k so that $\|k\vec{\mathbf{w}}\| = 60$ mph*. We have

$$\|k\vec{\mathbf{w}}\| = k\|\vec{\mathbf{w}}\| = k\frac{50}{\sqrt{3}} \text{ yd},$$

so we want

$$k\left(\frac{50}{\sqrt{3}} \text{ yd}\right) = 60 \text{ mph}$$

$$k = \frac{60\sqrt{3}}{50} \text{ mph / yd}$$

$$k = \frac{6\sqrt{3}}{5} \text{ mph / yd}.$$

Then

$$\vec{\mathbf{v}} = k\vec{\mathbf{w}} = k\langle 20, 15, \frac{25}{\sqrt{3}} \rangle = \frac{6\sqrt{3}}{5} \langle 20, 15, \frac{25}{\sqrt{3}} \rangle = \langle 24\sqrt{3}, 18\sqrt{3}, 30 \rangle.$$

Let's double-check that $\|\vec{\mathbf{v}}\| = 60$ mph. We have

$$\|\vec{\mathbf{v}}\| = \sqrt{(24\sqrt{3})^2 + (18\sqrt{3})^2 + (30)^2} = \sqrt{1728 + 972 + 900} = \sqrt{3600} = 60 \text{ mph}.$$

So, we have found the correct components for $\vec{\mathbf{v}}$.

Note *

Readers who have been watching the units of measurement may be wondering what exactly is going on at this point: haven't we just mixed yards and miles per hour? We haven't, but the reason is subtle. One way to understand it is to realize that there are really two parallel coordinate systems in this problem: one gives positions down the field, across the field, and up into the air in units of yards; the other gives speeds down the field, across the field, and up into the air in units of miles per hour. The vector $\vec{\mathbf{w}}$ is calculated in the position coordinate system; vector $\vec{\mathbf{v}}$ will be in the speed system. Because corresponding axes in each system are parallel, directions in the two systems are also parallel, so the claim that $\vec{\mathbf{w}}$ and $\vec{\mathbf{v}}$ point in the same direction is correct. The constant k that we're looking for is a conversion factor between the magnitudes of these two vectors, converting from the position system to the speed one in the process. And as seen above, our calculation of k produces the right units for such a conversion, namely miles per hour per yard.

Exercise 11.2.10

Assume the quarterback and the receiver are in the same place as in the previous example. This time, however, the quarterback throws the ball at velocity of 40 mph and an angle of 45°. Write the initial velocity vector of the ball, $\vec{\mathbf{v}}$, in component form.

Hint

Follow the process used in the previous example.

Answer

$v = \langle 16\sqrt{2}, 12\sqrt{2}, 20\sqrt{2} \rangle$

Key Concepts

- The three-dimensional coordinate system is built around a set of three axes that intersect at right angles at a single point, the origin. Ordered triples (x, y, z) are used to describe the location of a point in space.
- The distance d between points (x_1, y_1, z_1) and (x_2, y_2, z_2) is given by the formula

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}\ .$$

- In three dimensions, the equations $x = a,\ y = b,$ and $z = c$ describe planes that are parallel to the coordinate planes.
- The standard equation of a sphere with center (a, b, c) and radius r is

$$(x - a)^2 + (y - b)^2 + (z - c)^2 = r^2.$$

- In three dimensions, as in two, vectors are commonly expressed in component form, $\vec{\mathbf{v}} = \langle x, y, z \rangle$, or in terms of the standard unit vectors, $\vec{\mathbf{v}} = x\,\hat{\mathbf{i}} + y\,\hat{\mathbf{j}} + z\,\hat{\mathbf{k}}$.
- Properties of vectors in space are a natural extension of the properties for vectors in a plane. Let $\vec{\mathbf{v}} = \langle x_1, y_1, z_1 \rangle$ and $\vec{\mathbf{w}} = \langle x_2, y_2, z_2 \rangle$ be vectors, and let k be a scalar.

  Scalar multiplication:

$$k\vec{\mathbf{v}} = \langle kx_1, ky_1, kz_1 \rangle$$

  Vector addition:

$$\vec{\mathbf{v}} + \vec{\mathbf{w}} = \langle x_1, y_1, z_1 \rangle + \langle x_2, y_2, z_2 \rangle = \langle x_1 + x_2, y_1 + y_2, z_1 + z_2 \rangle$$

  Vector subtraction:

$$\vec{\mathbf{v}} - \vec{\mathbf{w}} = \langle x_1, y_1, z_1 \rangle - \langle x_2, y_2, z_2 \rangle = \langle x_1 - x_2, y_1 - y_2, z_1 - z_2 \rangle$$

  Vector magnitude:

$$\|\vec{\mathbf{v}}\| = \sqrt{x_1^2 + y_1^2 + z_1^2}$$

  Unit vector in the direction of $\vec{\mathbf{v}}$:

$$\frac{\vec{\mathbf{v}}}{\|\vec{\mathbf{v}}\|} = \frac{1}{\|\vec{\mathbf{v}}\|}\langle x_1, y_1, z_1 \rangle = \langle \frac{x_1}{\|\vec{\mathbf{v}}\|}, \frac{y_1}{\|\vec{\mathbf{v}}\|}, \frac{z_1}{\|\vec{\mathbf{v}}\|} \rangle,\ \vec{\mathbf{v}} \neq \vec{\mathbf{0}}$$

Key Equations

Distance between two points in space:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Sphere with center (a, b, c) and radius r:

$$(x - a)^2 + (y - b)^2 + (z - c)^2 = r^2$$

Glossary

coordinate plane
a plane containing two of the three coordinate axes in the three-dimensional coordinate system, named by the axes it contains: the xy-plane, xz-plane, or the yz-plane

right-hand rule
a common way to define the orientation of the three-dimensional coordinate system; when the right hand is curved around the z-axis in such a way that the fingers curl from the positive x-axis to the positive y-axis, the thumb points in the direction of the positive z-axis

octants
the eight regions of space created by the coordinate planes

sphere
the set of all points equidistant from a given point known as the *center*

standard equation of a sphere

$(x-a)^2+(y-b)^2+(z-c)^2=r^2$ describes a sphere with center (a,b,c) and radius r

three-dimensional rectangular coordinate system

a coordinate system defined by three lines that intersect at right angles; every point in space is described by an ordered triple (x,y,z) that plots its location relative to the defining axes

Contributors

Gilbert Strang (MIT) and Edwin “Jed” Herman (Harvey Mudd) with many contributing authors. This content by OpenStax is licensed with a CC-BY-SA-NC 4.0 license. Download for free at http://cnx.org.

Example 11.2.10has been modified by Doug Baldwin and Paul Seeburger to clarify the units of measurement that it uses and how it uses them.

Paul Seeburger also created dynamic versions of Figures 11.2.8, 11.2.9and 11.2.13using CalcPlot3D.

11.2: Vectors is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **12.2: Vectors in Three Dimensions** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

11.3: The Dot Product

Learning Objectives

- Calculate the dot product of two given vectors.
- Determine whether two given vectors are perpendicular.
- Find the direction cosines of a given vector.
- Explain what is meant by the vector projection of one vector onto another vector, and describe how to compute it.
- Calculate the work done by a given force.

If we apply a force to an object so that the object moves, we say that work is done by the force. Previously, we looked at a constant force and we assumed the force was applied in the direction of motion of the object. Under those conditions, work can be expressed as the product of the force acting on an object and the distance the object moves. In this chapter, however, we have seen that both force and the motion of an object can be represented by vectors.

In this section, we develop an operation called the dot product, which allows us to calculate work in the case when the force vector and the motion vector have different directions. The dot product essentially tells us how much of the force vector is applied in the direction of the motion vector. The dot product can also help us measure the angle formed by a pair of vectors and the position of a vector relative to the coordinate axes. It even provides a simple test to determine whether two vectors meet at a right angle.

The Dot Product and Its Properties

We have already learned how to add and subtract vectors. In this chapter, we investigate two types of vector multiplication. The first type of vector multiplication is called the dot product, based on the notation we use for it, and it is defined as follows:

Definition: dot product

The *dot product* of vectors $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$ is given by the sum of the products of the components

$$\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = u_1 v_1 + u_2 v_2 + u_3 v_3.$$

Note that if u and v are two-dimensional vectors, we calculate the dot product in a similar fashion. Thus, if $\vec{\mathbf{u}} = \langle u_1, u_2 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2 \rangle$, then

$$\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = u_1 v_1 + u_2 v_2.$$

When two vectors are combined under addition or subtraction, the result is a vector. When two vectors are combined using the dot product, the result is a scalar. For this reason, the dot product is often called the *scalar product.* It may also be called the *inner product.*

Example 11.3.1: Calculating Dot Products

a. Find the dot product of $\vec{\mathbf{u}} = \langle 3, 5, 2 \rangle$ and $\vec{\mathbf{v}} = \langle -1, 3, 0 \rangle$.

b. Find the scalar product of $\vec{\mathbf{p}} = 10\hat{\mathbf{i}} - 4\hat{\mathbf{j}} + 7\hat{\mathbf{k}}$ and $\vec{\mathbf{q}} = -2\hat{\mathbf{i}} + \hat{\mathbf{j}} + 6\hat{\mathbf{k}}$.

Solution:

a. Substitute the vector components into the formula for the dot product:

$$\begin{aligned} \vec{\mathbf{u}} \cdot \vec{\mathbf{v}} &= u_1 v_1 + u_2 v_2 + u_3 v_3 \\ &= 3(-1) + 5(3) + 2(0) \\ &= -3 + 15 + 0 \\ &= 12. \end{aligned}$$

b. The calculation is the same if the vectors are written using standard unit vectors. We still have three components for each vector to substitute into the formula for the dot product:

$$\begin{aligned}\vec{\mathbf{p}}\cdot\vec{\mathbf{q}} &= p_1q_1+p_2q_2+p_3q_3\\ &= 10(-2)+(-4)(1)+(7)(6)\\ &= -20-4+42\\ &= 18.\end{aligned}$$

Exercise 11.3.1

Find $\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}$, where $\vec{\mathbf{u}}=\langle 2,9,-1\rangle$ and $\vec{\mathbf{v}}=\langle -3,1,-4\rangle$.

Hint

Multiply corresponding components and then add their products.

Answer

7

Like vector addition and subtraction, the dot product has several algebraic properties. We prove three of these properties and leave the rest as exercises.

Properties of the Dot Product

Let $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$ be vectors, and let c be a scalar.

i. *Commutative property*

$$\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}=\vec{\mathbf{v}}\cdot\vec{\mathbf{u}}$$

ii. *Distributive property*

$$\vec{\mathbf{u}}\cdot(\vec{\mathbf{v}}+\vec{\mathbf{w}})=\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}+\vec{\mathbf{u}}\cdot\vec{\mathbf{w}}$$

iii. *Associative property*

$$c(\vec{\mathbf{u}}\cdot\vec{\mathbf{v}})=(c\vec{\mathbf{u}})\cdot\vec{\mathbf{v}}=\vec{\mathbf{u}}\cdot(c\vec{\mathbf{v}})$$

iv. *Property of magnitude*

$$\vec{\mathbf{v}}\cdot\vec{\mathbf{v}}=\|\vec{\mathbf{v}}\|^2$$

Proof

Let $\vec{\mathbf{u}}=\langle u_1,u_2,u_3\rangle$ and $\vec{\mathbf{v}}=\langle v_1,v_2,v_3\rangle$. Then

$$\begin{aligned}\vec{\mathbf{u}}\cdot\vec{\mathbf{v}} &= \langle u_1,u_2,u_3\rangle\cdot\langle v_1,v_2,v_3\rangle\\ &= u_1v_1+u_2v_2+u_3v_3\\ &= v_1u_1+v_2u_2+v_3u_3\\ &= \langle v_1,v_2,v_3\rangle\cdot\langle u_1,u_2,u_3\rangle\\ &= \vec{\mathbf{v}}\cdot\vec{\mathbf{u}}.\end{aligned}$$

The associative property looks like the associative property for real-number multiplication, but pay close attention to the difference between scalar and vector objects:

$$\begin{aligned} c(\vec{\mathbf{u}} \cdot \vec{\mathbf{v}}) &= c(u_1v_1 + u_2v_2 + u_3v_3) \\ &= c(u_1v_1) + c(u_2v_2) + c(u_3v_3) \\ &= (cu_1)v_1 + (cu_2)v_2 + (cu_3)v_3 \\ &= \langle cu_1, cu_2, cu_3 \rangle \cdot \langle v_1, v_2, v_3 \rangle \\ &= c\langle u_1, u_2, u_3 \rangle \cdot \langle v_1, v_2, v_3 \rangle \\ &= (c\vec{\mathbf{u}}) \cdot \vec{\mathbf{v}}. \end{aligned}$$

The proof that $c(\vec{\mathbf{u}} \cdot \vec{\mathbf{v}}) = \vec{\mathbf{u}} \cdot (c\vec{\mathbf{v}})$ is similar.

The fourth property shows the relationship between the magnitude of a vector and its dot product with itself:

$$\begin{aligned} \vec{\mathbf{v}} \cdot \vec{\mathbf{v}} &= \langle v_1, v_2, v_3 \rangle \cdot \langle v_1, v_2, v_3 \rangle \\ &= (v_1)^2 + (v_2)^2 + (v_3)^2 \\ &= \left[\sqrt{(v_1)^2 + (v_2)^2 + (v_3)^2}\right]^2 \\ &= \|\vec{\mathbf{v}}\|^2. \end{aligned}$$

□

Note that the definition of the dot product yields $\vec{\mathbf{0}} \cdot \vec{\mathbf{v}} = 0$. By property iv. if $\vec{\mathbf{v}} \cdot \vec{\mathbf{v}} = 0$, then $\vec{\mathbf{v}} = \vec{\mathbf{0}}$.

Example 11.3.2: Using Properties of the Dot Product

Let $\vec{\mathbf{a}} = \langle 1, 2, -3 \rangle$, $\vec{\mathbf{b}} = \langle 0, 2, 4 \rangle$, and $\vec{\mathbf{c}} = \langle 5, -1, 3 \rangle$.

Find each of the following products.

a. $(\vec{\mathbf{a}} \cdot \vec{\mathbf{b}})\vec{\mathbf{c}}$
b. $\vec{\mathbf{a}} \cdot (2\vec{\mathbf{c}})$
c. $\|\vec{\mathbf{b}}\|^2$

Solution

a. Note that this expression asks for the scalar multiple of $\vec{\mathbf{c}}$ by $\vec{\mathbf{a}} \cdot \vec{\mathbf{b}}$:

$$\begin{aligned} (\vec{\mathbf{a}} \cdot \vec{\mathbf{b}})\vec{\mathbf{c}} &= (\langle 1, 2, -3 \rangle \cdot \langle 0, 2, 4 \rangle)\langle 5, -1, 3 \rangle \\ &= (1(0) + 2(2) + (-3)(4))\langle 5, -1, 3 \rangle \\ &= -8\langle 5, -1, 3 \rangle \\ &= \langle -40, 8, -24 \rangle. \end{aligned}$$

b. This expression is a dot product of vector $\vec{\mathbf{a}}$ and scalar multiple $2\vec{\mathbf{c}}$:

$$\begin{aligned} \vec{\mathbf{a}} \cdot (2\vec{\mathbf{c}}) &= 2(\vec{\mathbf{a}} \cdot \vec{\mathbf{c}}) \\ &= 2(\langle 1, 2, -3 \rangle \cdot \langle 5, -1, 3 \rangle) \\ &= 2(1(5) + 2(-1) + (-3)(3)) \\ &= 2(-6) = -12. \end{aligned}$$

c. Simplifying this expression is a straightforward application of the dot product:

$$\begin{aligned}\|\vec{\mathbf{b}}\|^2 &= \vec{\mathbf{b}} \cdot \vec{\mathbf{b}} \\ &= \langle 0,2,4\rangle \cdot \langle 0,2,4\rangle \\ &= 0^2+2^2+4^2 \\ &= 0+4+16 \\ &= 20.\end{aligned}$$

Exercise 11.3.2

Find the following products for $\vec{\mathbf{p}} = \langle 7,0,2\rangle$, $\vec{\mathbf{q}} = \langle -2,2,-2\rangle$, and $\vec{\mathbf{r}} = \langle 0,2,-3\rangle$.

a. $(\vec{\mathbf{r}} \cdot \vec{\mathbf{p}})\vec{\mathbf{q}}$
b. $\|\vec{\mathbf{p}}\|^2$

Hint

$\vec{\mathbf{r}} \cdot \vec{\mathbf{p}}$ is a scalar.

Answer

a. $(\vec{\mathbf{r}} \cdot \vec{\mathbf{p}})\vec{\mathbf{q}} = \langle 12,-12,12\rangle$; *b.* $\|\vec{\mathbf{p}}\|^2 = 53$

Using the Dot Product to Find the Angle between Two Vectors

When two nonzero vectors are placed in standard position, whether in two dimensions or three dimensions, they form an angle between them (Figure 11.3.1). The dot product provides a way to find the measure of this angle. This property is a result of the fact that we can express the dot product in terms of the cosine of the angle formed by two vectors.

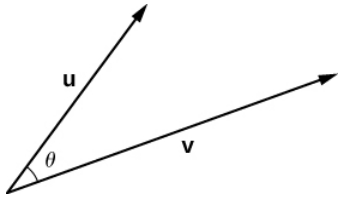


Figure 11.3.1: Let θ be the angle between two nonzero vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ such that $0 \le \theta \le \pi$.

Evaluating a Dot Product

The dot product of two vectors is the product of the magnitude of each vector and the cosine of the angle between them:

$$\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = \|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\| \cos\theta. \tag{11.3.1}$$

Proof

Place vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ in standard position and consider the vector $\vec{\mathbf{v}} - \vec{\mathbf{u}}$ (Figure 11.3.2). These three vectors form a triangle with side lengths $\|\vec{\mathbf{u}}\|$, $\|\vec{\mathbf{v}}\|$, and $\|\vec{\mathbf{v}} - \vec{\mathbf{u}}\|$.

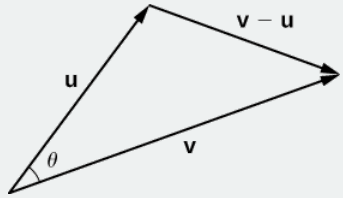


Figure 11.3.2: The lengths of the sides of the triangle are given by the magnitudes of the vectors that form the triangle.

Recall from trigonometry that the law of cosines describes the relationship among the side lengths of the triangle and the angle θ. Applying the law of cosines here gives

$$\|\vec{\mathbf{v}} - \vec{\mathbf{u}}\|^2 = \|\vec{\mathbf{u}}\|^2 + \|\vec{\mathbf{v}}\|^2 - 2\|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\| \cos\theta. \tag{11.3.2}$$

The dot product provides a way to rewrite the left side of Equation 11.3.2:

$$\begin{aligned}\|\vec{\mathbf{v}}-\vec{\mathbf{u}}\|^2 &= (\vec{\mathbf{v}}-\vec{\mathbf{u}})\cdot(\vec{\mathbf{v}}-\vec{\mathbf{u}})\\ &= (\vec{\mathbf{v}}-\vec{\mathbf{u}})\cdot\vec{\mathbf{v}}-(\vec{\mathbf{v}}-\vec{\mathbf{u}})\cdot\vec{\mathbf{u}}\\ &= \vec{\mathbf{v}}\cdot\vec{\mathbf{v}}-\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}-\vec{\mathbf{v}}\cdot\vec{\mathbf{u}}+\vec{\mathbf{u}}\cdot\vec{\mathbf{u}}\\ &= \vec{\mathbf{v}}\cdot\vec{\mathbf{v}}-\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}-\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}+\vec{\mathbf{u}}\cdot\vec{\mathbf{u}}\\ &= \|\vec{\mathbf{v}}\|^2-2\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}+\|\vec{\mathbf{u}}\|^2.\end{aligned}$$

Substituting into the law of cosines yields

$$\begin{aligned}\|\vec{\mathbf{v}}-\vec{\mathbf{u}}\|^2 &= \|\vec{\mathbf{u}}\|^2+\|\vec{\mathbf{v}}\|^2-2\|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\|\cos\theta\\ \|\vec{\mathbf{v}}\|^2-2\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}+\|\vec{\mathbf{u}}\|^2 &= \|\vec{\mathbf{u}}\|^2+\|\vec{\mathbf{v}}\|^2-2\|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\|\cos\theta\\ -2\vec{\mathbf{u}}\cdot\vec{\mathbf{v}} &= -2\|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\|\cos\theta\\ \vec{\mathbf{u}}\cdot\vec{\mathbf{v}} &= \|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\|\cos\theta.\end{aligned}$$

□

We can use the form of the dot product in Equation 11.3.1 to find the measure of the angle between two nonzero vectors by rearranging Equation 11.3.1 to solve for the cosine of the angle:

$$\cos\theta=\frac{\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\|}. \tag{11.3.3}$$

Using this equation, we can find the cosine of the angle between two nonzero vectors. Since we are considering the smallest angle between the vectors, we assume $0^\circ\le\theta\le180^\circ$ (or $0\le\theta\le\pi$ if we are working in radians). The inverse cosine is unique over this range, so we are then able to determine the measure of the angle θ.

Example 11.3.3: Finding the Angle between Two Vectors

Find the measure of the angle between each pair of vectors.

a. $\hat{\mathbf{i}}+\hat{\mathbf{j}}+\hat{\mathbf{k}}$ and $2\hat{\mathbf{i}}-\hat{\mathbf{j}}-3\hat{\mathbf{k}}$
b. $\langle 2,5,6\rangle$ and $\langle -2,-4,4\rangle$

Solution

a. To find the cosine of the angle formed by the two vectors, substitute the components of the vectors into Equation 11.3.3:

$$\begin{aligned}\cos\theta &= \frac{(\hat{\mathbf{i}}+\hat{\mathbf{j}}+\hat{\mathbf{k}})\cdot(2\hat{\mathbf{i}}-\hat{\mathbf{j}}-3\hat{\mathbf{k}})}{\left\|\hat{\mathbf{i}}+\hat{\mathbf{j}}+\hat{\mathbf{k}}\right\|\cdot\left\|2\hat{\mathbf{i}}-\hat{\mathbf{j}}-3\hat{\mathbf{k}}\right\|}\\ &= \frac{1(2)+(1)(-1)+(1)(-3)}{\sqrt{1^2+1^2+1^2}\sqrt{2^2+(-1)^2+(-3)^2}}\\ &= \frac{-2}{\sqrt{3}\sqrt{14}}=\frac{-2}{\sqrt{42}}.\end{aligned}$$

Therefore, $\theta=\arccos\dfrac{-2}{\sqrt{42}}$ rad.

b. Start by finding the value of the cosine of the angle between the vectors:

$$\begin{aligned}\cos\theta &= \frac{\langle 2,5,6\rangle \cdot \langle -2,-4,4\rangle}{\|\langle 2,5,6\rangle\| \cdot \|\langle -2,-4,4\rangle\|} \\ &= \frac{2(-2)+(5)(-4)+(6)(4)}{\sqrt{2^2+5^2+6^2}\sqrt{(-2)^2+(-4)^2+4^2}} \\ &= \frac{0}{\sqrt{65}\sqrt{36}} = 0.\end{aligned}$$

Now, $\cos\theta = 0$ and $0 \le \theta \le \pi$, so $\theta = \pi/2$.

Exercise 11.3.3

Find the measure of the angle, in radians, formed by vectors $\vec{\mathbf{a}} = \langle 1,2,0\rangle$ and $\vec{\mathbf{b}} = \langle 2,4,1\rangle$. Round to the nearest hundredth.

Hint

Use the Equation 11.3.3

Answer

$\theta \approx 0.22$ rad

The angle between two vectors can be acute $(0 < \cos\theta < 1)$, obtuse $(-1 < \cos\theta < 0)$, or straight $(\cos\theta = -1)$. If $\cos\theta = 1$, then both vectors have the same direction. If $\cos\theta = 0$, then the vectors, when placed in standard position, form a right angle (Figure 11.3.3). We can formalize this result into a theorem regarding orthogonal (perpendicular) vectors.

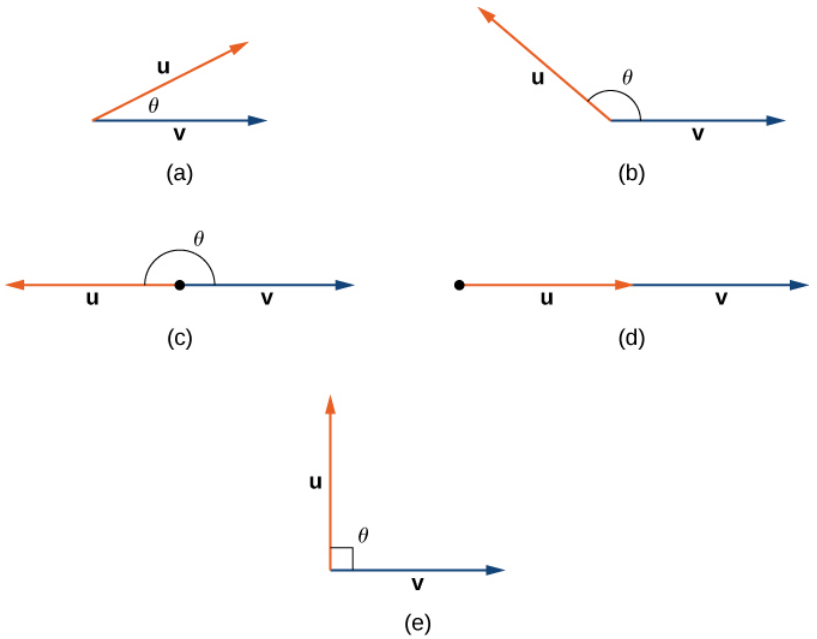


Figure 11.3.3: (a) An acute angle has $0 < \cos\theta < 1$. (b) An obtuse angle has $-1 < \cos\theta < 0$. (c) A straight line has $\cos\theta = -1$. (d) If the vectors have the same direction, $\cos\theta = 1$. (e) If the vectors are orthogonal (perpendicular), $\cos\theta = 0$.

Orthogonal Vectors

The nonzero vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ are **orthogonal vectors** if and only if $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = 0$.

Proof

Let $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ be nonzero vectors, and let θ denote the angle between them. First, assume $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = 0$. Then

$$\|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\| \cos\theta = 0.$$

However, $\|\vec{\mathbf{u}}\| \neq 0$ and $\|\vec{\mathbf{v}}\| \neq 0$, so we must have $\cos\theta = 0$. Hence, $\theta = 90^\circ$, and the vectors are orthogonal.

Now assume $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ are orthogonal. Then $\theta = 90^\circ$ and we have

$$\begin{aligned} \vec{\mathbf{u}} \cdot \vec{\mathbf{v}} &= \|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\| \cos\theta \\ &= \|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\| \cos 90^\circ \\ &= \|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\| (0) \\ &= 0. \end{aligned}$$

□

The terms **orthogonal, perpendicular,** and **normal** each indicate that mathematical objects are intersecting at right angles. The use of each term is determined mainly by its context. We say that vectors are orthogonal and lines are perpendicular. The term *normal* is used most often when measuring the angle made with a plane or other surface.

Example 11.3.4: Identifying Orthogonal Vectors

Determine whether $\vec{\mathbf{p}} = \langle 1, 0, 5 \rangle$ and $\vec{\mathbf{q}} = \langle 10, 3, -2 \rangle$ are orthogonal vectors.

Solution

Using the definition, we need only check the dot product of the vectors:

$$\vec{\mathbf{p}} \cdot \vec{\mathbf{q}} = 1(10) + (0)(3) + (5)(-2) = 10 + 0 - 10 = 0.$$

Because $\vec{\mathbf{p}} \cdot \vec{\mathbf{q}} = 0,$ the vectors are orthogonal (Figure 11.3.4).

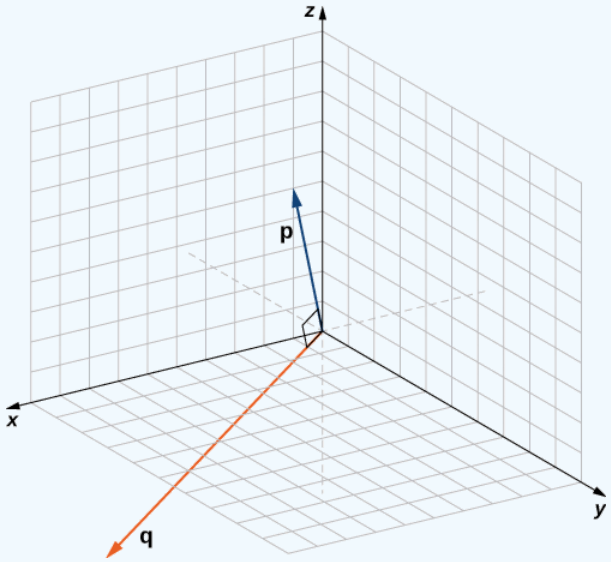


Figure 11.3.4: Vectors $\vec{\mathbf{p}}$ and $\vec{\mathbf{q}}$ form a right angle when their initial points are aligned.

Exercise 11.3.4

For which value of x is $\vec{\mathbf{p}} = \langle 2, 8, -1 \rangle$ orthogonal to $\vec{\mathbf{q}} = \langle x, -1, 2 \rangle$?

Hint

Vectors $\vec{\mathbf{p}}$ and $\vec{\mathbf{q}}$ are orthogonal if and only if $\vec{\mathbf{p}} \cdot \vec{\mathbf{q}} = 0$.

Answer

$x = 5$

Example 11.3.5: Measuring the Angle Formed by Two Vectors

Let $\vec{\mathbf{v}} = \langle 2, 3, 3 \rangle$. Find the measures of the angles formed by the following vectors.

a. $\vec{\mathbf{v}}$ and $\hat{\mathbf{i}}$
b. $\vec{\mathbf{v}}$ and $\hat{\mathbf{j}}$
c. $\vec{\mathbf{v}}$ and $\hat{\mathbf{k}}$

Solution

a.Let α be the angle formed by $\vec{\mathbf{v}}$ and $\hat{\mathbf{i}}$:

$$\cos\alpha = \frac{\vec{\mathbf{v}} \cdot \hat{\mathbf{i}}}{\|\vec{\mathbf{v}}\| \cdot \|\hat{\mathbf{i}}\|} = \frac{\langle 2, 3, 3 \rangle \cdot \langle 1, 0, 0 \rangle}{\sqrt{2^2 + 3^2 + 3^2}\sqrt{1}} = \frac{2}{\sqrt{22}}$$

$$\alpha = \arccos \frac{2}{\sqrt{22}} \approx 1.130 \text{ rad.}$$

b. Let β represent the angle formed by $\vec{\mathbf{v}}$ and $\hat{\mathbf{j}}$:

$$\cos\beta = \frac{\vec{\mathbf{v}} \cdot \hat{\mathbf{j}}}{\|\vec{\mathbf{v}}\| \cdot \|\hat{\mathbf{j}}\|} = \frac{\langle 2, 3, 3 \rangle \cdot \langle 0, 1, 0 \rangle}{\sqrt{2^2 + 3^2 + 3^2}\sqrt{1}} = \frac{3}{\sqrt{22}}$$

$$\beta = \arccos \frac{3}{\sqrt{22}} \approx 0.877 \text{ rad.}$$

c. Let γ represent the angle formed by $\vec{\mathbf{v}}$ and $\hat{\mathbf{k}}$:

$$\cos\gamma = \frac{\vec{\mathbf{v}} \cdot \hat{\mathbf{k}}}{\|\vec{\mathbf{v}}\| \cdot \|\hat{\mathbf{k}}\|} = \frac{\langle 2, 3, 3 \rangle \cdot \langle 0, 0, 1 \rangle}{\sqrt{2^2 + 3^2 + 3^2}\sqrt{1}} = \frac{3}{\sqrt{22}}$$

$$\gamma = \arccos \frac{3}{\sqrt{22}} \approx 0.877 \text{ rad.}$$

Exercise 11.3.5

Let $\vec{\mathbf{v}} = \langle 3, -5, 1 \rangle$. Find the measure of the angles formed by each pair of vectors.

a. $\vec{\mathbf{v}}$ and $\hat{\mathbf{i}}$
b. $\vec{\mathbf{v}}$ and $\hat{\mathbf{j}}$
c. $\vec{\mathbf{v}}$ and $\hat{\mathbf{k}}$

Hint

$\hat{\mathbf{i}} = \langle 1, 0, 0 \rangle$, $\hat{\mathbf{j}} = \langle 0, 1, 0 \rangle$, and $\hat{\mathbf{k}} = \langle 0, 0, 1 \rangle$

Answer

a. $\alpha \approx 1.04$ rad; b. $\beta \approx 2.58$ rad; c. $\gamma \approx 1.40$ rad

The angle a vector makes with each of the coordinate axes, called a direction angle, is very important in practical computations, especially in a field such as engineering. For example, in astronautical engineering, the angle at which a rocket is launched must be determined very precisely. A very small error in the angle can lead to the rocket going hundreds of miles off course. Direction angles are often calculated by using the dot product and the cosines of the angles, called the direction cosines. Therefore, we define both these angles and their cosines.

Definition: direction angles

The angles formed by a nonzero vector and the coordinate axes are called the **direction angles** for the vector (Figure 11.3.5). The cosines for these angles are called the **direction cosines.**

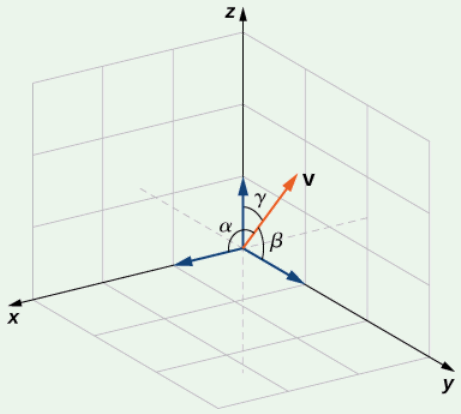


Figure 11.3.5: Angle α is formed by vector $\vec{\mathbf{v}}$ and unit vector $\hat{\mathbf{i}}$. Angle β is formed by vector $\vec{\mathbf{v}}$ and unit vector $\hat{\mathbf{j}}$. Angle γ is formed by vector $\vec{\mathbf{v}}$ and unit vector $\hat{\mathbf{k}}$.

In Example 11.3.5, the direction cosines of $\vec{\mathbf{v}} = \langle 2, 3, 3 \rangle$ are $\cos\alpha = \dfrac{2}{\sqrt{22}}, \cos\beta = \dfrac{3}{\sqrt{22}}$, and $\cos\gamma = \dfrac{3}{\sqrt{22}}$. The direction angles of $\vec{\mathbf{v}}$ are $\alpha = 1.130$ rad, $\beta = 0.877$ rad, and $\gamma = 0.877$ rad.

So far, we have focused mainly on vectors related to force, movement, and position in three-dimensional physical space. However, vectors are often used in more abstract ways. For example, suppose a fruit vendor sells apples, bananas, and oranges. On a given day, he sells 30 apples, 12 bananas, and 18 oranges. He might use a quantity vector, $\vec{\mathbf{q}} = \langle 30, 12, 18 \rangle$, to represent the quantity of fruit he sold that day. Similarly, he might want to use a price vector, $\vec{\mathbf{p}} = \langle 0.50, 0.25, 1 \rangle$, to indicate that he sells his apples for 50¢ each, bananas for 25¢ each, and oranges for $1 apiece. In this example, although we could still graph these vectors, we do not interpret them as literal representations of position in the physical world. We are simply using vectors to keep track of particular pieces of information about apples, bananas, and oranges.

This idea might seem a little strange, but if we simply regard vectors as a way to order and store data, we find they can be quite a powerful tool. Going back to the fruit vendor, let's think about the dot product, $\vec{\mathbf{q}} \cdot \vec{\mathbf{p}}$. We compute it by multiplying the number of apples sold (30) by the price per apple (50¢), the number of bananas sold by the price per banana, and the number of oranges sold by the price per orange. We then add all these values together. So, in this example, the dot product tells us how much money the fruit vendor had in sales on that particular day.

When we use vectors in this more general way, there is no reason to limit the number of components to three. What if the fruit vendor decides to start selling grapefruit? In that case, he would want to use four-dimensional quantity and price vectors to represent the number of apples, bananas, oranges, and grapefruit sold, and their unit prices. As you might expect, to calculate the dot product of four-dimensional vectors, we simply add the products of the components as before, but the sum has four terms instead of three.

Example 11.3.6: Using Vectors in an Economic Context

AAA Party Supply Store sells invitations, party favors, decorations, and food service items such as paper plates and napkins. When AAA buys its inventory, it pays 25¢ per package for invitations and party favors. Decorations cost AAA 50¢ each, and food service items cost 20¢ per package. AAA sells invitations for $2.50 per package and party favors for $1.50 per package. Decorations sell for $4.50 each and food service items for $1.25 per package.

During the month of May, AAA Party Supply Store sells 1258 invitations, 342 party favors, 2426 decorations, and 1354 food service items. Use vectors and dot products to calculate how much money AAA made in sales during the month of May. How much did the store make in profit?

Solution

The cost, price, and quantity vectors are

$$\vec{\mathbf{c}} = \langle 0.25, 0.25, 0.50, 0.20 \rangle$$
$$\vec{\mathbf{p}} = \langle 2.50, 1.50, 4.50, 1.25 \rangle$$
$$\vec{\mathbf{q}} = \langle 1258, 342, 2426, 1354 \rangle.$$

AAA sales for the month of May can be calculated using the dot product $\vec{\mathbf{p}} \cdot \vec{\mathbf{q}}$. We have

$$\begin{aligned} \vec{\mathbf{p}} \cdot \vec{\mathbf{q}} &= \langle 2.50, 1.50, 4.50, 1.25 \rangle \cdot \langle 1258, 342, 2426, 1354 \rangle \\ &= 3145 + 513 + 10917 + 1692.5 \\ &= 16267.5. \end{aligned}$$

So, AAA took in \$16,267.50 during the month of May. To calculate the profit, we must first calculate how much AAA paid for the items sold. We use the dot product $\vec{\mathbf{c}} \cdot \vec{\mathbf{q}}$ to get

$$\begin{aligned} \vec{\mathbf{c}} \cdot \vec{\mathbf{q}} &= \langle 0.25, 0.25, 0.50, 0.20 \rangle \cdot \langle 1258, 342, 2426, 1354 \rangle \\ &= 314.5 + 85.5 + 1213 + 270.8 \\ &= 1883.8. \end{aligned}$$

So, AAA paid \$1,883.80 for the items they sold. Their profit, then, is given by

$$\vec{\mathbf{p}} \cdot \vec{\mathbf{q}} - \vec{\mathbf{c}} \cdot \vec{\mathbf{q}} = 16267.5 - 1883.8 = 14383.7.$$

Therefore, AAA Party Supply Store made \$14,383.70 in May.

? Exercise 11.3.6

On June 1, AAA Party Supply Store decided to increase the price they charge for party favors to \$2 per package. They also changed suppliers for their invitations, and are now able to purchase invitations for only 10¢ per package. All their other costs and prices remain the same. If AAA sells 1408 invitations, 147 party favors, 2112 decorations, and 1894 food service items in the month of June, use vectors and dot products to calculate their total sales and profit for June.

Hint

Use four-dimensional vectors for cost, price, and quantity sold.

Answer

Sales = \$15,685.50; profit = \$14,073.15

Projections

As we have seen, addition combines two vectors to create a resultant vector. But what if we are given a vector and we need to find its component parts? We use vector projections to perform the opposite process; they can break down a vector into its components. The magnitude of a vector projection is a scalar projection. For example, if a child is pulling the handle of a wagon at a 55° angle, we can use projections to determine how much of the force on the handle is actually moving the wagon forward (11.3.6). We return to this example and learn how to solve it after we see how to calculate projections.

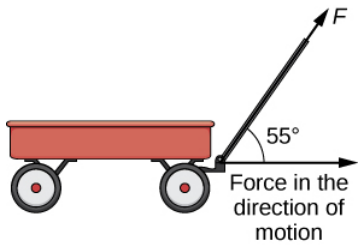


Figure 11.3.6: When a child pulls a wagon, only the horizontal component of the force propels the wagon forward.

Definition: Vector and Projection

The **vector projection** of $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$ is the vector labeled $\text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}}$ in Figure 11.3.7. It has the same initial point as $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ and the same direction as $\vec{\mathbf{u}}$, and represents the component of $\vec{\mathbf{v}}$ that acts in the direction of $\vec{\mathbf{u}}$. If θ represents the angle between $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$, then, by properties of triangles, we know the length of $\text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}}$ is $\|\text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}}\| = \|\vec{\mathbf{v}}\|\cos\theta$. When expressing $\cos\theta$ in terms of the dot product, this becomes

$$\|\text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}}\| = \|\vec{\mathbf{v}}\|\cos\theta = \|\vec{\mathbf{v}}\|\left(\frac{\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\|}\right) = \frac{\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|}.$$

We now multiply by a unit vector in the direction of $\vec{\mathbf{u}}$ to get $\text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}}$:

$$\text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}} = \frac{\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|}\left(\frac{1}{\|\vec{\mathbf{u}}\|}\vec{\mathbf{u}}\right) = \frac{\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|^2}\vec{\mathbf{u}}.$$

The length of this vector is also known as the **scalar projection** of $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$ and is denoted by

$$\|\text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}}\| = \text{comp}_{\vec{\mathbf{u}}}\vec{\mathbf{v}} = \frac{\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|}.$$

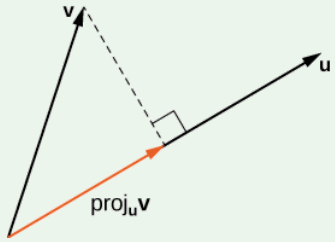


Figure 11.3.7: The projection of $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$ shows the component of vector $\vec{\mathbf{v}}$ in the direction of $\vec{\mathbf{u}}$.

Example 11.3.7: Finding Projections

Find the projection of $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$.

a. $\vec{\mathbf{v}} = \langle 3, 5, 1\rangle$ and $\vec{\mathbf{u}} = \langle -1, 4, 3\rangle$
b. $\vec{\mathbf{v}} = 3\hat{\mathbf{i}} - 2\hat{\mathbf{j}}$ and $\vec{\mathbf{u}} = \hat{\mathbf{i}} + 6\hat{\mathbf{j}}$

Solution

a. Substitute the components of $\vec{\mathbf{v}}$ and $\vec{\mathbf{u}}$ into the formula for the projection:

$$\begin{aligned}\text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}} &= \frac{\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|^2}\vec{\mathbf{u}}\\ &= \frac{\langle -1,4,3\rangle\cdot\langle 3,5,1\rangle}{\|\langle -1,4,3\rangle\|^2}\langle -1,4,3\rangle\\ &= \frac{-3+20+3}{(-1)^2+4^2+3^2}\langle -1,4,3\rangle\\ &= \frac{20}{26}\langle -1,4,3\rangle\\ &= \langle -\frac{10}{13}, \frac{40}{13}, \frac{30}{13}\rangle.\end{aligned}$$

b. To find the two-dimensional projection, simply adapt the formula to the two-dimensional case:

$$\begin{aligned}\text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}} &= \frac{\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|^2}\vec{\mathbf{u}} \\ &= \frac{(\hat{\mathbf{i}}+6\hat{\mathbf{j}})\cdot(3\hat{\mathbf{i}}-2\hat{\mathbf{j}})}{\left\|\hat{\mathbf{i}}+6\hat{\mathbf{j}}\right\|^2}(\hat{\mathbf{i}}+6\hat{\mathbf{j}}) \\ &= \frac{1(3)+6(-2)}{1^2+6^2}(\hat{\mathbf{i}}+6\hat{\mathbf{j}}) \\ &= -\frac{9}{37}(\hat{\mathbf{i}}+6\hat{\mathbf{j}}) \\ &= -\frac{9}{37}\hat{\mathbf{i}}-\frac{54}{37}\hat{\mathbf{j}}.\end{aligned}$$

Sometimes it is useful to decompose vectors—that is, to break a vector apart into a sum. This process is called the **resolution of a vector into components.** Projections allow us to identify two orthogonal vectors having a desired sum. For example, let $\vec{\mathbf{v}}=\langle 6,-4\rangle$ and let $\vec{\mathbf{u}}=\langle 3,1\rangle$. We want to decompose the vector $\vec{\mathbf{v}}$ into orthogonal components such that one of the component vectors has the same direction as $\vec{\mathbf{u}}$.

We first find the component that has the same direction as $\vec{\mathbf{u}}$ by projecting $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$. Let $\vec{\mathbf{p}}=\text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}}$. Then, we have

$$\begin{aligned}\vec{\mathbf{p}} &= \frac{\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|^2}\vec{\mathbf{u}} \\ &= \frac{18-4}{9+1}\vec{\mathbf{u}} \\ = \frac{7}{5}\vec{\mathbf{u}} = \frac{7}{5}\langle 3,1\rangle &= \langle\frac{21}{5},\frac{7}{5}\rangle.\end{aligned}$$

Now consider the vector $\vec{\mathbf{q}}=\vec{\mathbf{v}}-\vec{\mathbf{p}}$. We have

$$\begin{aligned}\vec{\mathbf{q}} &= \vec{\mathbf{v}}-\vec{\mathbf{p}} \\ &= \langle 6,-4\rangle-\langle\frac{21}{5},\frac{7}{5}\rangle \\ &= \langle\frac{9}{5},-\frac{27}{5}\rangle.\end{aligned}$$

Clearly, by the way we defined $\vec{\mathbf{q}}$, we have $\vec{\mathbf{v}}=\vec{\mathbf{q}}+\vec{\mathbf{p}}$, and

$$\begin{aligned}\vec{\mathbf{q}}\cdot\vec{\mathbf{p}} &= \langle\frac{9}{5},-\frac{27}{5}\rangle\cdot\langle\frac{21}{5},\frac{7}{5}\rangle \\ &= \frac{9(21)}{25}+-\frac{27(7)}{25} \\ &= \frac{189}{25}-\frac{189}{25}=0.\end{aligned}$$

Therefore, $\vec{\mathbf{q}}$ and $\vec{\mathbf{p}}$ are orthogonal.

Example 11.3.8: Resolving Vectors into Components

Express $\vec{\mathbf{v}}=\langle 8,-3,-3\rangle$ as a sum of orthogonal vectors such that one of the vectors has the same direction as $\vec{\mathbf{u}}=\langle 2,3,2\rangle$.

Solution

Let $\vec{\mathbf{p}}$ represent the projection of $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$:

$$\begin{aligned}\vec{\mathbf{p}} &= \text{proj}_{\vec{\mathbf{u}}}\vec{\mathbf{v}} \\ &= \frac{\vec{\mathbf{u}}\cdot\vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|^2}\vec{\mathbf{u}} \\ &= \frac{\langle 2,3,2\rangle\cdot\langle 8,-3,-3\rangle}{\|\langle 2,3,2\rangle\|^2}\langle 2,3,2\rangle \\ &= \frac{16-9-6}{2^2+3^2+2^2}\langle 2,3,2\rangle \\ &= \frac{1}{17}\langle 2,3,2\rangle \\ &= \langle \frac{2}{17},\frac{3}{17},\frac{2}{17}\rangle.\end{aligned}$$

Then,

$$\begin{aligned}\vec{\mathbf{q}} &= \vec{\mathbf{v}}-\vec{\mathbf{p}} = \langle 8,-3,-3\rangle - \langle \frac{2}{17},\frac{3}{17},\frac{2}{17}\rangle \\ &= \langle \frac{134}{17},-\frac{54}{17},-\frac{53}{17}\rangle.\end{aligned}$$

To check our work, we can use the dot product to verify that $\vec{\mathbf{p}}$ and $\vec{\mathbf{q}}$ are orthogonal vectors:

$$\begin{aligned}\vec{\mathbf{p}}\cdot\vec{\mathbf{q}} &= \langle \frac{2}{17},\frac{3}{17},\frac{2}{17}\rangle\cdot\langle \frac{134}{17},-\frac{54}{17},-\frac{53}{17}\rangle \\ &= \frac{268}{289}-\frac{162}{289}-\frac{106}{289}=0.\end{aligned}$$

Then,

$$\vec{\mathbf{v}} = \vec{\mathbf{p}}+\vec{\mathbf{q}} = \langle \frac{2}{17},\frac{3}{17},\frac{2}{17}\rangle + \langle \frac{134}{17},-\frac{54}{17},-\frac{53}{17}\rangle.$$

Exercise 11.3.7

Express $\vec{\mathbf{v}} = 5\hat{\mathbf{i}}-\hat{\mathbf{j}}$ as a sum of orthogonal vectors such that one of the vectors has the same direction as $\vec{\mathbf{u}} = 4\hat{\mathbf{i}}+2\hat{\mathbf{j}}$.

Hint

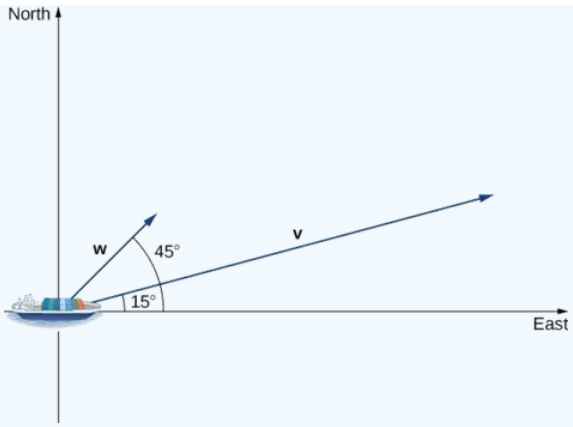
Start by finding the projection of $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$.

Answer

$\vec{\mathbf{v}} = \vec{\mathbf{p}}+\vec{\mathbf{q}}$, where $\vec{\mathbf{p}} = \frac{18}{5}\hat{\mathbf{i}}+\frac{9}{5}\hat{\mathbf{j}}$ and $\vec{\mathbf{q}} = \frac{7}{5}\hat{\mathbf{i}}-\frac{14}{5}\hat{\mathbf{j}}$

Example 11.3.9: Scalar Projection of Velocity

A container ship leaves port traveling 15° north of east. Its engine generates a speed of 20 knots along that path (see the following figure). In addition, the ocean current moves the ship northeast at a speed of 2 knots. Considering both the engine and the current, how fast is the ship moving in the direction 15° north of east? Round the answer to two decimal places.



Solution

Let $\vec{\mathbf{v}}$ be the velocity vector generated by the engine, and let $\vec{\mathbf{w}}$ be the velocity vector of the current. We already know $\|\vec{\mathbf{v}}\| = 20$ along the desired route. We just need to add in the scalar projection of $\vec{\mathbf{w}}$ onto $\vec{\mathbf{v}}$. We get

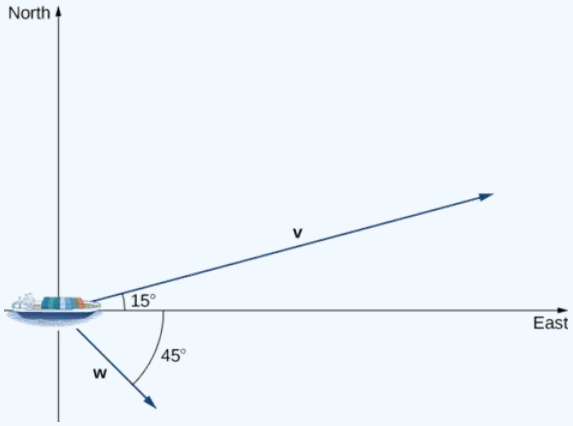
$$\text{comp}_{\vec{\mathbf{v}}}\vec{\mathbf{w}} = \frac{\vec{\mathbf{v}} \cdot \vec{\mathbf{w}}}{\|\vec{\mathbf{v}}\|}$$

$$= \frac{\|\vec{\mathbf{v}}\|\|\vec{\mathbf{w}}\| \cos(30^\circ)}{\|\vec{\mathbf{v}}\|} = \|\vec{\mathbf{w}}\| \cos(30^\circ) = 2\frac{\sqrt{3}}{2} = \sqrt{3} \approx 1.73 \text{ knots}.$$

The ship is moving at 21.73 knots in the direction 15° north of east.

? Exercise 11.3.8

Repeat the previous example, but assume the ocean current is moving southeast instead of northeast, as shown in the following figure.



Hint

Compute the scalar projection of $\vec{\mathbf{w}}$ onto $\vec{\mathbf{v}}$.

Answer

21 knots

Work

Now that we understand dot products, we can see how to apply them to real-life situations. The most common application of the dot product of two vectors is in the calculation of work.

From physics, we know that work is done when an object is moved by a force. When the force is constant and applied in the same direction the object moves, then we define the work done as the product of the force and the distance the object travels: $W = Fd$. We saw several examples of this type in earlier chapters. Now imagine the direction of the force is different from the direction of motion, as with the example of a child pulling a wagon. To find the work done, we need to multiply the component of the force that acts in the direction of the motion by the magnitude of the displacement. The dot product allows us to do just that. If we represent an applied force by a vector $\vec{\mathbf{F}}$ and the displacement of an object by a vector $\vec{\mathbf{s}}$, then the **work done by the force** is the dot product of $\vec{\mathbf{F}}$ and $\vec{\mathbf{s}}$.

Definition: Constant Force

When a constant force is applied to an object so the object moves in a straight line from point P to point Q, the work W done by the force $\vec{\mathbf{F}}$, acting at an angle θ from the line of motion, is given by

$$W = \vec{\mathbf{F}} \cdot \overrightarrow{PQ} = \| \vec{\mathbf{F}} \| \| \overrightarrow{PQ} \| \cos\theta.$$

Let's revisit the problem of the child's wagon introduced earlier. Suppose a child is pulling a wagon with a force having a magnitude of 8 lb on the handle at an angle of 55°. If the child pulls the wagon 50 ft, find the work done by the force (Figure 11.3.8).

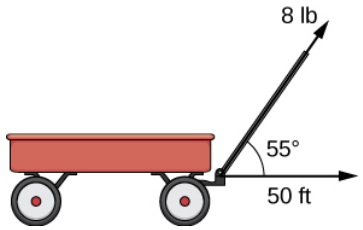


Figure 11.3.8: The horizontal component of the force is the projection of $\vec{\mathbf{F}}$ onto the positive x-axis.

We have

$$W = \| \vec{\mathbf{F}} \| \| \overrightarrow{PQ} \| \cos\theta = 8(50)(\cos(55^\circ)) \approx 229 \text{ ft·lb}.$$

In U.S. standard units, we measure the magnitude of force $\left\| \vec{\mathbf{F}} \right\|$ in pounds. The magnitude of the displacement vector $\left\| \overrightarrow{PQ} \right\|$ tells us how far the object moved, and it is measured in feet. The customary unit of measure for work, then, is the foot-pound. One foot-pound is the amount of work required to move an object weighing 1 lb a distance of 1 ft straight up. In the metric system, the unit of measure for force is the newton (N), and the unit of measure of magnitude for work is a newton-meter (N·m), or a joule (J).

Example 11.3.10: Calculating Work

A conveyor belt generates a force $\vec{\mathbf{F}} = 5\hat{\mathbf{i}} - 3\hat{\mathbf{j}} + \hat{\mathbf{k}}$ that moves a suitcase from point $(1, 1, 1)$ to point $(9, 4, 7)$ along a straight line. Find the work done by the conveyor belt. The distance is measured in meters and the force is measured in newtons.

Solution

The displacement vector $\overrightarrow{PQ}$ has initial point $(1, 1, 1)$ and terminal point $(9, 4, 7)$:

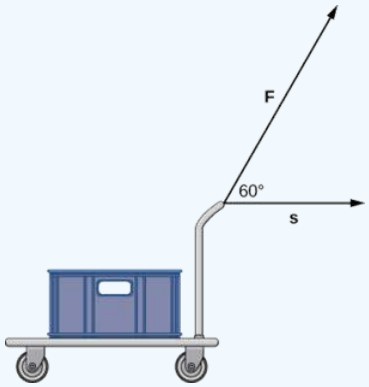
$$\overrightarrow{PQ} = \langle 9-1, 4-1, 7-1 \rangle = \langle 8, 3, 6 \rangle = 8\hat{\mathbf{i}} + 3\hat{\mathbf{j}} + 6\hat{\mathbf{k}}.$$

Work is the dot product of force and displacement:

$$
\begin{aligned}
W &= \vec{\mathbf{F}} \cdot \overrightarrow{PQ} \\
&= (5\hat{\mathbf{i}} - 3\hat{\mathbf{j}} + \hat{\mathbf{k}}) \cdot (8\hat{\mathbf{i}} + 3\hat{\mathbf{j}} + 6\hat{\mathbf{k}}) \\
&= 5(8) + (-3)(3) + 1(6) \\
&= 37\,\text{N·m} \\
&= 37\,\text{J}
\end{aligned}
$$

? Exercise 11.3.9

A constant force of 30 lb is applied at an angle of 60° to pull a handcart 10 ft across the ground. What is the work done by this force?



Hint

Use the definition of work as the dot product of force and distance.

Answer

150 ft-lb

Key Concepts

- The dot product, or scalar product, of two vectors $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$ is $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = u_1 v_1 + u_2 v_2 + u_3 v_3$.
- The dot product satisfies the following properties:
  - $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = \vec{\mathbf{v}} \cdot \vec{\mathbf{u}}$
  - $\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} + \vec{\mathbf{w}}) = \vec{\mathbf{u}} \cdot \vec{\mathbf{v}} + \vec{\mathbf{u}} \cdot \vec{\mathbf{w}}$
  - $c(\vec{\mathbf{u}} \cdot \vec{\mathbf{v}}) = (c\vec{\mathbf{u}}) \cdot \vec{\mathbf{v}} = \vec{\mathbf{u}} \cdot (c\vec{\mathbf{v}})$
  - $\vec{\mathbf{v}} \cdot \vec{\mathbf{v}} = \|\vec{\mathbf{v}}\|^2$
- The dot product of two vectors can be expressed, alternatively, as $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = \|\vec{\mathbf{u}}\| \|\vec{\mathbf{v}}\| \cos\theta$. This form of the dot product is useful for finding the measure of the angle formed by two vectors.
- Vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ are orthogonal if $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = 0$.
- The angles formed by a nonzero vector and the coordinate axes are called the *direction angles* for the vector. The cosines of these angles are known as the *direction cosines.*
- The vector projection of $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$ is the vector $\text{proj}_{\vec{\mathbf{u}}} \vec{\mathbf{v}} = \dfrac{\vec{\mathbf{u}} \cdot \vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|^2} \vec{\mathbf{u}}$. The magnitude of this vector is known as the *scalar projection* of $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$, given by $\text{comp}_{\vec{\mathbf{u}}} \vec{\mathbf{v}} = \dfrac{\vec{\mathbf{u}} \cdot \vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|}$.
- Work is done when a force is applied to an object, causing displacement. When the force is represented by the vector $\vec{\mathbf{F}}$ and the displacement is represented by the vector $\vec{\mathbf{s}}$, then the work done W is given by the formula $W = \vec{\mathbf{F}} \cdot \vec{\mathbf{s}} = \| \vec{\mathbf{F}} \| \, \| \vec{\mathbf{s}} \| \cos\theta$.

Key Equations

- **Dot product of $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$**

  $$\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = u_1 v_1 + u_2 v_2 + u_3 v_3 = \|\vec{\mathbf{u}}\| \|\vec{\mathbf{v}}\| \cos\theta$$

- **Cosine of the angle formed by $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$**

  $$\cos\theta = \frac{\vec{\mathbf{u}} \cdot \vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\| \|\vec{\mathbf{v}}\|}$$

- **Vector projection of $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$**

  $$\text{proj}_{\vec{\mathbf{u}}} \vec{\mathbf{v}} = \frac{\vec{\mathbf{u}} \cdot \vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|^2} \vec{\mathbf{u}}$$

- **Scalar projection of $\vec{\mathbf{v}}$ onto $\vec{\mathbf{u}}$**

  $$\text{comp}_{\vec{\mathbf{u}}} \vec{\mathbf{v}} = \frac{\vec{\mathbf{u}} \cdot \vec{\mathbf{v}}}{\|\vec{\mathbf{u}}\|}$$

- **Work done by a force $\vec{\mathbf{F}}$ to move an object through displacement vector $\overrightarrow{PQ}$**

  $$W = \vec{\mathbf{F}} \cdot \overrightarrow{PQ} = \| \vec{\mathbf{F}} \| \| \overrightarrow{PQ} \| \cos\theta$$

Glossary

direction angles
the angles formed by a nonzero vector and the coordinate axes

direction cosines
the cosines of the angles formed by a nonzero vector and the coordinate axes

dot product or scalar product
$\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = u_1 v_1 + u_2 v_2 + u_3 v_3$ where $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$

scalar projection
the magnitude of the vector projection of a vector

orthogonal vectors
vectors that form a right angle when placed in standard position

vector projection
the component of a vector that follows a given direction

work done by a force
work is generally thought of as the amount of energy it takes to move an object; if we represent an applied force by a vector $\vec{\mathbf{F}}$ and the displacement of an object by a vector $\vec{\mathbf{s}}$, then the work done by the force is the dot product of $\vec{\mathbf{F}}$ and $\vec{\mathbf{s}}$.

- Gilbert Strang (MIT) and Edwin “Jed” Herman (Harvey Mudd) with many contributing authors. This content by OpenStax is licensed with a CC-BY-SA-NC 4.0 license. Download for free at http://cnx.org.
- edited for vector notation by Paul Seeburger

11.3: The Dot Product is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **12.3: The Dot Product** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

11.4: The Cross Product

Learning Objectives

- Calculate the cross product of two given vectors.
- Use determinants to calculate a cross product.
- Find a vector orthogonal to two given vectors.
- Determine areas and volumes by using the cross product.
- Calculate the torque of a given force and position vector.

Imagine a mechanic turning a wrench to tighten a bolt. The mechanic applies a force at the end of the wrench. This creates rotation, or torque, which tightens the bolt. We can use vectors to represent the force applied by the mechanic, and the distance (radius) from the bolt to the end of the wrench. Then, we can represent torque by a vector oriented along the axis of rotation. Note that the torque vector is orthogonal to both the force vector and the radius vector.

In this section, we develop an operation called the **cross product**, which allows us to find a vector orthogonal to two given vectors. Calculating torque is an important application of cross products, and we examine torque in more detail later in the section.

The Cross Product and Its Properties

The dot product is a multiplication of two vectors that results in a scalar. In this section, we introduce a product of two vectors that generates a third vector orthogonal to the first two. Consider how we might find such a vector. Let $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$ be nonzero vectors. We want to find a vector $\vec{\mathbf{w}} = \langle w_1, w_2, w_3 \rangle$ orthogonal to both $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$—that is, we want to find $\vec{\mathbf{w}}$ such that $\vec{\mathbf{u}} \cdot \vec{\mathbf{w}} = 0$ and $\vec{\mathbf{v}} \cdot \vec{\mathbf{w}} = 0$. Therefore, w_1, w_2, and w_3 must satisfy

$$u_1 w_1 + u_2 w_2 + u_3 w_3 = 0 \tag{11.4.1}$$

$$v_1 w_1 + v_2 w_2 + v_3 w_3 = 0. \tag{11.4.2}$$

If we multiply the top equation by v_3 and the bottom equation by u_3 and subtract, we can eliminate the variable w_3, which gives

$$(u_1 v_3 - v_1 u_3) w_1 + (u_2 v_3 - v_2 u_3) w_2 = 0.$$

If we select

$$\begin{aligned} w_1 &= u_2 v_3 - u_3 v_2 \\ w_2 &= -(u_1 v_3 - u_3 v_1), \end{aligned}$$

we get a possible solution vector. Substituting these values back into the original equations (Equations 11.4.1 and 11.4.2) gives

$$w_3 = u_1 v_2 - u_2 v_1.$$

That is, vector

$$\vec{\mathbf{w}} = \langle u_2 v_3 - u_3 v_2, -(u_1 v_3 - u_3 v_1), u_1 v_2 - u_2 v_1 \rangle$$

is orthogonal to both $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$, which leads us to define the following operation, called the *cross product*.

Definition: Cross Product

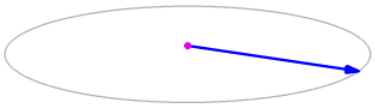
Let $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$. Then, the *cross product* $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ is vector

$$\begin{aligned} \vec{\mathbf{u}} \times \vec{\mathbf{v}} &= (u_2 v_3 - u_3 v_2)\hat{\mathbf{i}} - (u_1 v_3 - u_3 v_1)\hat{\mathbf{j}} + (u_1 v_2 - u_2 v_1)\hat{\mathbf{k}} \\ &= \langle u_2 v_3 - u_3 v_2, -(u_1 v_3 - u_3 v_1), u_1 v_2 - u_2 v_1 \rangle. \end{aligned} \tag{11.4.3}$$

From the way we have developed $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$, it should be clear that the cross product is orthogonal to both $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$. However, it never hurts to check. To show that $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ is orthogonal to $\vec{\mathbf{u}}$, we calculate the dot product of $\vec{\mathbf{u}}$ and $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$.

$$\begin{aligned}\vec{\mathbf{u}}\cdot(\vec{\mathbf{u}}\times\vec{\mathbf{v}}) &= \langle u_1,u_2,u_3\rangle\cdot\langle u_2v_3-u_3v_2,-u_1v_3+u_3v_1,u_1v_2-u_2v_1\rangle\\ &= u_1(u_2v_3-u_3v_2)+u_2(-u_1v_3+u_3v_1)+u_3(u_1v_2-u_2v_1)\\ &= u_1u_2v_3-u_1u_3v_2-u_1u_2v_3+u_2u_3v_1+u_1u_3v_2-u_2u_3v_1\\ &= (u_1u_2v_3-u_1u_2v_3)+(-u_1u_3v_2+u_1u_3v_2)+(u_2u_3v_1-u_2u_3v_1)\\ &= 0\end{aligned}$$

In a similar manner, we can show that the cross product is also orthogonal to $\vec{\mathbf{v}}$.



The cross product $\vec{\mathbf{a}}\times\vec{\mathbf{b}}$ (vertical, in pink) changes as the angle between the vectors $\vec{\mathbf{a}}$ (blue) and $\vec{\mathbf{b}}$ (red) changes. The cross product (purple) is always perpendicular to both vectors, and has magnitude zero when the vectors are parallel and maximum magnitude $\|\vec{\mathbf{a}}\|\|\vec{\mathbf{b}}\|$ when they are perpendicular. (Public Domain; LucasVB).

Example 11.4.1: Finding a Cross Product

Let $\vec{\mathbf{p}}=\langle -1,2,5\rangle$ and $\vec{\mathbf{q}}=\langle 4,0,-3\rangle$ (Figure 11.4.1). Find $\vec{\mathbf{p}}\times\vec{\mathbf{q}}$.

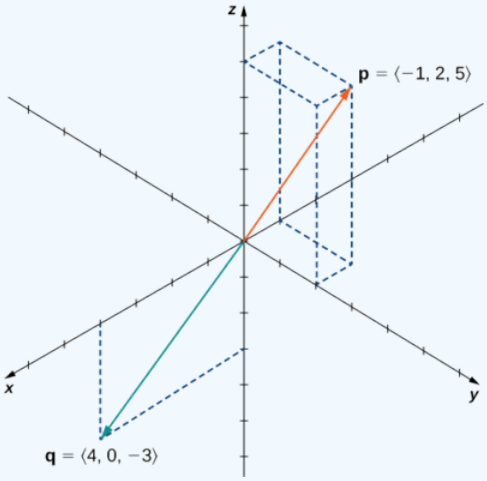


Figure 11.4.1: Finding a cross product to two given vectors.

Solution

Substitute the components of the vectors into Equation 11.4.3:

$$\begin{aligned}\vec{\mathbf{p}}\times\vec{\mathbf{q}} &= \langle -1,2,5\rangle\times\langle 4,0,-3\rangle\\ &= \langle p_2q_3-p_3q_2,-(p_1q_3-p_3q_1),p_1q_2-p_2q_1\rangle\\ &= \langle 2(-3)-5(0),-(-1)(-3)+5(4),(-1)(0)-2(4)\rangle\\ &= \langle -6,17,-8\rangle.\end{aligned}$$

Exercise 11.4.1

Find $\vec{\mathbf{p}} \times \vec{\mathbf{q}}$ for $\vec{\mathbf{p}} = \langle 5, 1, 2 \rangle$ and $\vec{\mathbf{q}} = \langle -2, 0, 1 \rangle$. Express the answer using standard unit vectors.

Hint

Use the formula $\vec{\mathbf{u}} \times \vec{\mathbf{v}} = (u_2 v_3 - u_3 v_2)\hat{\mathbf{i}} - (u_1 v_3 - u_3 v_1)\hat{\mathbf{j}} + (u_1 v_2 - u_2 v_1)\hat{\mathbf{k}}$.

Answer

$\vec{\mathbf{p}} \times \vec{\mathbf{q}} = \hat{\mathbf{i}} - 9\hat{\mathbf{j}} + 2\hat{\mathbf{k}}$

Although it may not be obvious from Equation 11.4.3, the direction of $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ is given by the right-hand rule. If we hold the right hand out with the fingers pointing in the direction of $\vec{\mathbf{u}}$, then curl the fingers toward vector $\vec{\mathbf{v}}$, the thumb points in the direction of the cross product, as shown in Figure 11.4.2

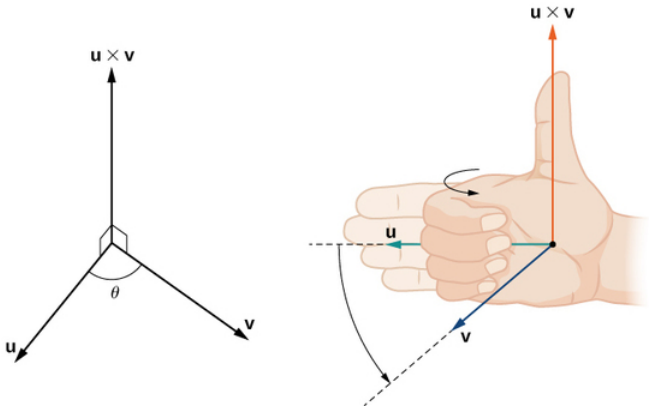


Figure 11.4.2: The direction of $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ is determined by the right-hand rule.

Notice what this means for the direction of $\vec{\mathbf{v}} \times \vec{\mathbf{u}}$. If we apply the right-hand rule to $\vec{\mathbf{v}} \times \vec{\mathbf{u}}$, we start with our fingers pointed in the direction of $\vec{\mathbf{v}}$, then curl our fingers toward the vector $\vec{\mathbf{u}}$. In this case, the thumb points in the opposite direction of $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$. (Try it!)

Example 11.4.2: Anticommutativity of the Cross Product

Let $\vec{\mathbf{u}} = \langle 0, 2, 1 \rangle$ and $\vec{\mathbf{v}} = \langle 3, -1, 0 \rangle$. Calculate $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ and $\vec{\mathbf{v}} \times \vec{\mathbf{u}}$ and graph them.

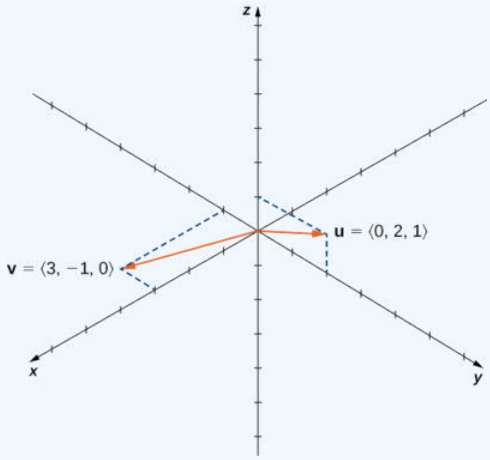


Figure 11.4.3: Are the cross products $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ and $\vec{\mathbf{v}} \times \vec{\mathbf{u}}$ in the same direction?

Solution

We have

$$\vec{\mathbf{u}} \times \vec{\mathbf{v}} = \langle (0+1), -(0-3), (0-6) \rangle = \langle 1, 3, -6 \rangle$$

$$\vec{\mathbf{v}} \times \vec{\mathbf{u}} = \langle (-1-0), -(3-0), (6-0) \rangle = \langle -1, -3, 6 \rangle.$$

We see that, in this case, $\vec{\mathbf{u}} \times \vec{\mathbf{v}} = -(\vec{\mathbf{v}} \times \vec{\mathbf{u}})$ (Figure 11.4.4). We prove this in general later in this section.

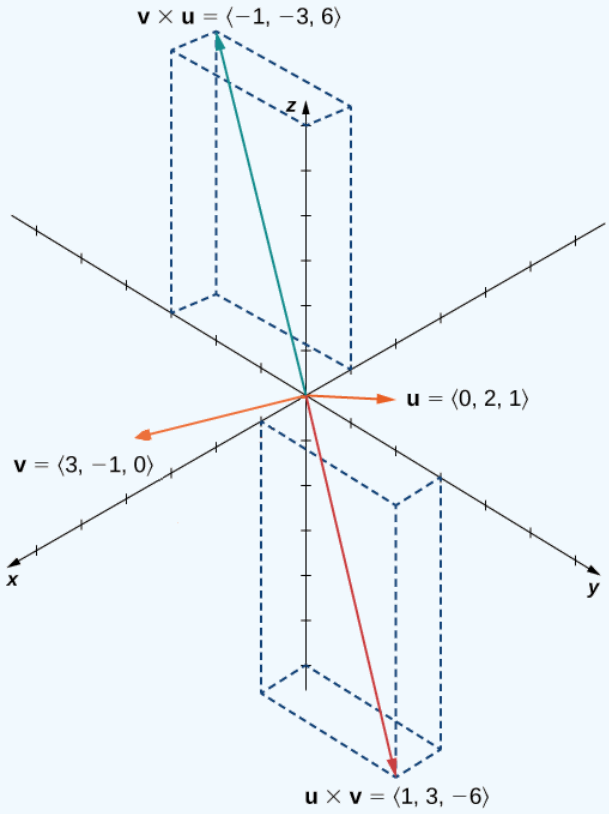


Figure 11.4.4: The cross products $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ and $\vec{\mathbf{v}} \times \vec{\mathbf{u}}$ are both orthogonal to $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$, but in opposite directions.

? Exercise 11.4.2

Suppose vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ lie in the xy-plane (the z-component of each vector is zero). Now suppose the x- and y-components of $\vec{\mathbf{u}}$ and the y-component of $\vec{\mathbf{v}}$ are all positive, whereas the x-component of $\vec{\mathbf{v}}$ is negative. Assuming the coordinate axes are oriented in the usual positions, in which direction does $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ point?

Hint

Remember the right-hand rule (Figure 11.4.2).

Answer

Up (the positive z-direction)

The cross products of the standard unit vectors $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$ can be useful for simplifying some calculations, so let's consider these cross products. A straightforward application of the definition shows that

$$\hat{\mathbf{i}} \times \hat{\mathbf{i}} = \hat{\mathbf{j}} \times \hat{\mathbf{j}} = \hat{\mathbf{k}} \times \hat{\mathbf{k}} = \vec{\mathbf{0}}.$$

(The cross product of two vectors is a vector, so each of these products results in the zero vector, not the scalar 0.) It's up to you to verify the calculations on your own.

Furthermore, because the cross product of two vectors is orthogonal to each of these vectors, we know that the cross product of $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ is parallel to $\hat{\mathbf{k}}$. Similarly, the vector product of $\hat{\mathbf{i}}$ and $\hat{\mathbf{k}}$ is parallel to $\hat{\mathbf{j}}$, and the vector product of $\hat{\mathbf{j}}$ and $\hat{\mathbf{k}}$ is parallel to $\hat{\mathbf{i}}$.

We can use the right-hand rule to determine the direction of each product. Then we have

$$\hat{\mathbf{i}} \times \hat{\mathbf{j}} = \hat{\mathbf{k}}$$
$$\hat{\mathbf{j}} \times \hat{\mathbf{i}} = -\hat{\mathbf{k}}$$

$$\hat{\mathbf{j}} \times \hat{\mathbf{k}} = \hat{\mathbf{i}}$$
$$\hat{\mathbf{k}} \times \hat{\mathbf{j}} = -\hat{\mathbf{i}}$$

$$\hat{\mathbf{k}} \times \hat{\mathbf{i}} = \hat{\mathbf{j}}$$
$$\hat{\mathbf{i}} \times \hat{\mathbf{k}} = -\hat{\mathbf{j}}.$$

These formulas come in handy later.

Example 11.4.3: Cross Product of Standard Unit Vectors

Find $\hat{\mathbf{i}} \times (\hat{\mathbf{j}} \times \hat{\mathbf{k}})$.

Solution

We know that $\hat{\mathbf{j}} \times \hat{\mathbf{k}} = \hat{\mathbf{i}}$. Therefore, $\hat{\mathbf{i}} \times (\hat{\mathbf{j}} \times \hat{\mathbf{k}}) = \hat{\mathbf{i}} \times \hat{\mathbf{i}} = \vec{\mathbf{0}}$.

Exercise 11.4.3

Find $(\hat{\mathbf{i}} \times \hat{\mathbf{j}}) \times (\hat{\mathbf{k}} \times \hat{\mathbf{i}})$.

Hint

Remember the right-hand rule (Figure 11.4.2).

Answer

$-\hat{\mathbf{i}}$

As we have seen, the dot product is often called the *scalar product* because it results in a scalar. The cross product results in a vector, so it is sometimes called the **vector product.** These operations are both versions of vector multiplication, but they have very different properties and applications. Let's explore some properties of the cross product. We prove only a few of them. Proofs of the other properties are left as exercises.

Properties of the Cross Product

Let $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$ be vectors in space, and let c be a scalar.

i. Anticommutative property:

$$\vec{\mathbf{u}} \times \vec{\mathbf{v}} = -(\vec{\mathbf{v}} \times \vec{\mathbf{u}})$$

ii. Distributive property:

$$\vec{\mathbf{u}} \times (\vec{\mathbf{v}} + \vec{\mathbf{w}}) = \vec{\mathbf{u}} \times \vec{\mathbf{v}} + \vec{\mathbf{u}} \times \vec{\mathbf{w}}$$

iii. Multiplication by a constant:

$$c(\vec{\mathbf{u}} \times \vec{\mathbf{v}}) = (c\vec{\mathbf{u}}) \times \vec{\mathbf{v}} = \vec{\mathbf{u}} \times (c\vec{\mathbf{v}})$$

iv. Cross product of the zero vector:

$$\vec{\mathbf{u}} \times \vec{\mathbf{0}} = \vec{\mathbf{0}} \times \vec{\mathbf{u}} = \vec{\mathbf{0}}$$

v. Cross product of a vector with itself:

$$\vec{\mathbf{v}} \times \vec{\mathbf{v}} = \vec{\mathbf{0}}$$

vi. Triple scalar product:

$$\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) = (\vec{\mathbf{u}} \times \vec{\mathbf{v}}) \cdot \vec{\mathbf{w}}$$

vii. Triple cross product:

$$\vec{\mathbf{u}} \times (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) = (\vec{\mathbf{u}} \cdot \vec{\mathbf{w}})\, \vec{\mathbf{v}} - (\vec{\mathbf{u}} \cdot \vec{\mathbf{v}})\, \vec{\mathbf{w}}$$

Proof

For property i, we want to show $\vec{\mathbf{u}} \times \vec{\mathbf{v}} = -(\vec{\mathbf{v}} \times \vec{\mathbf{u}})$. We have

$$\begin{aligned} \vec{\mathbf{u}} \times \vec{\mathbf{v}} &= \langle u_1, u_2, u_3 \rangle \times \langle v_1, v_2, v_3 \rangle \\ &= \langle u_2 v_3 - u_3 v_2, -u_1 v_3 + u_3 v_1, u_1 v_2 - u_2 v_1 \rangle \\ &= -\langle u_3 v_2 - u_2 v_3, -u_3 v_1 + u_1 v_3, u_2 v_1 - u_1 v_2 \rangle \\ &= -\langle v_1, v_2, v_3 \rangle \times \langle u_1, u_2, u_3 \rangle \\ &= -(\vec{\mathbf{v}} \times \vec{\mathbf{u}}). \end{aligned}$$

Unlike most operations we've seen, the cross product is not commutative. This makes sense if we think about the right-hand rule.

For property $iv.$, this follows directly from the definition of the cross product. We have

$$\vec{\mathbf{u}} \times \vec{\mathbf{0}} = \langle u_2(0) - u_3(0), -(u_1(0) - u_3(0)), u_1(0) - u_2(0) \rangle = \langle 0, 0, 0 \rangle = \vec{\mathbf{0}}.$$

Then, by property i., $\vec{\mathbf{0}} \times \vec{\mathbf{u}} = \vec{\mathbf{0}}$ as well. Remember that the dot product of a vector and the zero vector is the *scalar* 0, whereas the cross product of a vector with the zero vector is the *vector* $\vec{\mathbf{0}}$.

Property $vi.$ looks like the associative property, but note the change in operations:

$$\begin{aligned} \vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) &= u \cdot \langle v_2 w_3 - v_3 w_2, -v_1 w_3 + v_3 w_1, v_1 w_2 - v_2 w_1 \rangle \\ &= u_1 (v_2 w_3 - v_3 w_2) + u_2 (-v_1 w_3 + v_3 w_1) + u_3 (v_1 w_2 - v_2 w_1) \\ &= u_1 v_2 w_3 - u_1 v_3 w_2 - u_2 v_1 w_3 + u_2 v_3 w_1 + u_3 v_1 w_2 - u_3 v_2 w_1 \\ &= (u_2 v_3 - u_3 v_2) w_1 + (u_3 v_1 - u_1 v_3) w_2 + (u_1 v_2 - u_2 v_1) w_3 \\ &= \langle u_2 v_3 - u_3 v_2, u_3 v_1 - u_1 v_3, u_1 v_2 - u_2 v_1 \rangle \cdot \langle w_1, w_2, w_3 \rangle = (\vec{\mathbf{u}} \times \vec{\mathbf{v}}) \cdot \vec{\mathbf{w}}. \end{aligned}$$

□

Example 11.4.4: Using the Properties of the Cross Product

Use the cross product properties to calculate $(2\hat{\mathbf{i}} \times 3\hat{\mathbf{j}}) \times \hat{\mathbf{j}}$.

Solution

$$\begin{aligned} (2\hat{\mathbf{i}} \times 3\hat{\mathbf{j}}) \times \hat{\mathbf{j}} &= 2(\hat{\mathbf{i}} \times 3\hat{\mathbf{j}}) \times \hat{\mathbf{j}} \\ &= 2(3)(\hat{\mathbf{i}} \times \hat{\mathbf{j}}) \times \hat{\mathbf{j}} \\ &= (6\hat{\mathbf{k}}) \times \hat{\mathbf{j}} \\ &= 6(\hat{\mathbf{k}} \times \hat{\mathbf{j}}) \\ &= 6(-\hat{\mathbf{i}}) = -6\hat{\mathbf{i}}. \end{aligned}$$

Exercise 11.4.4

Use the properties of the cross product to calculate $(\hat{\mathbf{i}} \times \hat{\mathbf{k}}) \times (\hat{\mathbf{k}} \times \hat{\mathbf{j}})$.

Hint

$\vec{\mathbf{u}} \times \vec{\mathbf{v}} = -(\vec{\mathbf{v}} \times \vec{\mathbf{u}})$

Answer

$-\hat{\mathbf{k}}$

So far in this section, we have been concerned with the direction of the vector $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$, but we have not discussed its magnitude. It turns out there is a simple expression for the magnitude of $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ involving the magnitudes of $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$, and the sine of the angle between them.

Magnitude of the Cross Product

Let $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ be vectors, and let θ be the angle between them. Then, $\|\vec{\mathbf{u}} \times \vec{\mathbf{v}}\| = \|\vec{\mathbf{u}}\| \cdot \|\vec{\mathbf{v}}\| \cdot \sin\theta$.

Proof

Let $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$ be vectors, and let θ denote the angle between them. Then

$$\begin{aligned}
\|\vec{\mathbf{u}} \times \vec{\mathbf{v}}\|^2 &= (u_2v_3 - u_3v_2)^2 + (u_3v_1 - u_1v_3)^2 + (u_1v_2 - u_2v_1)^2 \\
&= u_2^2v_3^2 - 2u_2u_3v_2v_3 + u_3^2v_2^2 + u_3^2v_1^2 - 2u_1u_3v_1v_3 + u_1^2v_3^2 + u_1^2v_2^2 - 2u_1u_2v_1v_2 + u_2^2v_1^2 \\
&= u_1^2v_1^2 + u_1^2v_2^2 + u_1^2v_3^2 + u_2^2v_1^2 + u_2^2v_2^2 + u_2^2v_3^2 + u_3^2v_1^2 + u_3^2v_2^2 + u_3^2v_3^2 \\
&\quad - (u_1^2v_1^2 + u_2^2v_2^2 + u_3^2v_3^2 + 2u_1u_2v_1v_2 + 2u_1u_3v_1v_3 + 2u_2u_3v_2v_3) \\
&= (u_1^2 + u_2^2 + u_3^2)(v_1^2 + v_2^2 + v_3^2) - (u_1v_1 + u_2v_2 + u_3v_3)^2 \\
&= \|\vec{\mathbf{u}}\|^2\|\vec{\mathbf{v}}\|^2 - (\vec{\mathbf{u}} \cdot \vec{\mathbf{v}})^2 \\
&= \|\vec{\mathbf{u}}\|^2\|\vec{\mathbf{v}}\|^2 - \|\vec{\mathbf{u}}\|^2\|\vec{\mathbf{v}}\|^2 \cos^2\theta \\
&= \|\vec{\mathbf{u}}\|^2\|\vec{\mathbf{v}}\|^2(1 - \cos^2\theta) \\
&= \|\vec{\mathbf{u}}\|^2\|\vec{\mathbf{v}}\|^2(\sin^2\theta).
\end{aligned}$$

Taking square roots and noting that $\sqrt{\sin^2\theta} = \sin\theta$ for $0 \leq \theta \leq 180^\circ$, we have the desired result:

$$\|\vec{\mathbf{u}} \times \vec{\mathbf{v}}\| = \|\vec{\mathbf{u}}\|\|\vec{\mathbf{v}}\| \sin\theta.$$

□

This definition of the cross product allows us to visualize or interpret the product geometrically. It is clear, for example, that the cross product is defined only for vectors in three dimensions, not for vectors in two dimensions. In two dimensions, it is impossible to generate a vector simultaneously orthogonal to two nonparallel vectors.

Example 11.4.5: Calculating the Cross Product

Use "Magnitude of the Cross Product" to find the magnitude of the cross product of $\vec{\mathbf{u}} = \langle 0, 4, 0 \rangle$ and $\vec{\mathbf{v}} = \langle 0, 0, -3 \rangle$.

Solution

We have

$$\begin{aligned}\|\vec{\mathbf{u}}\times\vec{\mathbf{v}}\| &= \|\vec{\mathbf{u}}\|\cdot\|\vec{\mathbf{v}}\|\cdot\sin\theta \\ &= \sqrt{0^2+4^2+0^2}\cdot\sqrt{0^2+0^2+(-3)^2}\cdot\sin\frac{\pi}{2} \\ &= 4(3)(1)=12\end{aligned}$$

? Exercise 11.4.5

Use "Magnitude of the Cross Product" to find the magnitude of $\vec{\mathbf{u}}\times\vec{\mathbf{v}}$, where $\vec{\mathbf{u}}=\langle -8,0,0\rangle$ and $\vec{\mathbf{v}}=\langle 0,2,0\rangle$.

Hint

Vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ are orthogonal.

Answer

16

Determinants and the Cross Product

Using Equation 11.4.3 to find the cross product of two vectors is straightforward, and it presents the cross product in the useful component form. The formula, however, is complicated and difficult to remember. Fortunately, we have an alternative. We can calculate the cross product of two vectors using determinant notation.

A 2×2 determinant is defined by

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1b_2 - b_1a_2.$$

For example,

$$\begin{vmatrix} 3 & -2 \\ 5 & 1 \end{vmatrix} = 3(1)-5(-2)=3+10=13.$$

A 3×3 determinant is defined in terms of 2×2 determinants as follows:

$$\begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = a_1\begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} - a_2\begin{vmatrix} b_1 & b_3 \\ c_1 & c_3 \end{vmatrix} + a_3\begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix}. \tag{11.4.4}$$

Equation 11.4.4 is referred to as the expansion of the determinant along the first row. Notice that the multipliers of each of the 2×2 determinants on the right side of this expression are the entries in the first row of the 3×3 determinant. Furthermore, each of the 2×2 determinants contains the entries from the 3×3 determinant that would remain if you crossed out the row and column containing the multiplier. Thus, for the first term on the right, a_1 is the multiplier, and the 2×2 determinant contains the entries that remain if you cross out the first row and first column of the 3×3 determinant. Similarly, for the second term, the multiplier is a_2, and the 2×2 determinant contains the entries that remain if you cross out the first row and second column of the 3×3 determinant. Notice, however, that the coefficient of the second term is negative. The third term can be calculated in similar fashion.

✓ Example 11.4.6: Using Expansion Along the First Row to Compute a 3×3 Determinant

Evaluate the determinant $\begin{vmatrix} 2 & 5 & -1 \\ -1 & 1 & 3 \\ -2 & 3 & 4 \end{vmatrix}$.

Solution

We have

$$\begin{vmatrix} 2 & 5 & -1 \\ -1 & 1 & 3 \\ -2 & 3 & 4 \end{vmatrix} = 2\begin{vmatrix} 1 & 3 \\ 3 & 4 \end{vmatrix} - 5\begin{vmatrix} -1 & 3 \\ -2 & 4 \end{vmatrix} - 1\begin{vmatrix} -1 & 1 \\ -2 & 3 \end{vmatrix}$$
$$= 2(4-9) - 5(-4+6) - 1(-3+2)$$
$$= 2(-5) - 5(2) - 1(-1) = -10 - 10 + 1$$
$$= -19$$

? Exercise 11.4.6

Evaluate the determinant $\begin{vmatrix} 1 & -2 & -1 \\ 3 & 2 & -3 \\ 1 & 5 & 4 \end{vmatrix}$.

Hint

Expand along the first row. Don't forget the second term is negative!

Answer

40

Technically, determinants are defined only in terms of arrays of real numbers. However, the determinant notation provides a useful mnemonic device for the cross product formula.

Rule: Cross Product Calculated by a Determinant

Let $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$ be vectors. Then the cross product $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ is given by

$$\vec{\mathbf{u}} \times \vec{\mathbf{v}} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = \begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix} \hat{\mathbf{i}} - \begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix} \hat{\mathbf{j}} + \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \hat{\mathbf{k}}.$$

✓ Example 11.4.7: Using Determinant Notation to find $\vec{\mathbf{p}} \times \vec{\mathbf{q}}$

Let $\vec{\mathbf{p}} = \langle -1, 2, 5 \rangle$ and $\vec{\mathbf{q}} = \langle 4, 0, -3 \rangle$. Find $\vec{\mathbf{p}} \times \vec{\mathbf{q}}$.

Solution

We set up our determinant by putting the standard unit vectors across the first row, the components of $\vec{\mathbf{u}}$ in the second row, and the components of $\vec{\mathbf{v}}$ in the third row. Then, we have

$$\vec{\mathbf{p}} \times \vec{\mathbf{q}} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ -1 & 2 & 5 \\ 4 & 0 & -3 \end{vmatrix} = \begin{vmatrix} 2 & 5 \\ 0 & -3 \end{vmatrix} \hat{\mathbf{i}} - \begin{vmatrix} -1 & 5 \\ 4 & -3 \end{vmatrix} \hat{\mathbf{j}} + \begin{vmatrix} -1 & 2 \\ 4 & 0 \end{vmatrix} \hat{\mathbf{k}}$$
$$= (-6-0)\hat{\mathbf{i}} - (3-20)\hat{\mathbf{j}} + (0-8)\hat{\mathbf{k}}$$
$$= -6\hat{\mathbf{i}} + 17\hat{\mathbf{j}} - 8\hat{\mathbf{k}}.$$

Notice that this answer confirms the calculation of the cross product in Example 11.4.1.

? Exercise 11.4.7

Use determinant notation to find $\vec{\mathbf{a}} \times \vec{\mathbf{b}}$, where $\vec{\mathbf{a}} = \langle 8, 2, 3 \rangle$ and $\vec{\mathbf{b}} = \langle -1, 0, 4 \rangle$.

Hint

Calculate the determinant $\begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ 8 & 2 & 3 \\ -1 & 0 & 4 \end{vmatrix}$.

Answer

$\vec{\mathbf{a}} \times \vec{\mathbf{b}} = 8\hat{\mathbf{i}} - 35\hat{\mathbf{j}} + 2\hat{\mathbf{k}}$

Using the Cross Product

The cross product is very useful for several types of calculations, including finding a vector orthogonal to two given vectors, computing areas of triangles and parallelograms, and even determining the volume of the three-dimensional geometric shape made of parallelograms known as a *parallelepiped*. The following examples illustrate these calculations.

Example 11.4.8: Finding a Unit Vector Orthogonal to Two Given Vectors

Let $\vec{\mathbf{a}} = \langle 5, 2, -1 \rangle$ and $\vec{\mathbf{b}} = \langle 0, -1, 4 \rangle$. Find a unit vector orthogonal to both $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$.

Solution

The cross product $\vec{\mathbf{a}} \times \vec{\mathbf{b}}$ is orthogonal to both vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$. We can calculate it with a determinant:

$$\begin{aligned} \vec{\mathbf{a}} \times \vec{\mathbf{b}} &= \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ 5 & 2 & -1 \\ 0 & -1 & 4 \end{vmatrix} = \begin{vmatrix} 2 & -1 \\ -1 & 4 \end{vmatrix} \hat{\mathbf{i}} - \begin{vmatrix} 5 & -1 \\ 0 & 4 \end{vmatrix} \hat{\mathbf{j}} + \begin{vmatrix} 5 & 2 \\ 0 & -1 \end{vmatrix} \hat{\mathbf{k}} \\ &= (8-1)\hat{\mathbf{i}} - (20-0)\hat{\mathbf{j}} + (-5-0)\hat{\mathbf{k}} \\ &= 7\hat{\mathbf{i}} - 20\hat{\mathbf{j}} - 5\hat{\mathbf{k}}. \end{aligned}$$

Normalize this vector to find a unit vector in the same direction:

$$\|\vec{\mathbf{a}} \times \vec{\mathbf{b}}\| = \sqrt{(7)^2 + (-20)^2 + (-5)^2} = \sqrt{474}\,.$$

Thus, $\left\langle \frac{7}{\sqrt{474}}, \frac{-20}{\sqrt{474}}, \frac{-5}{\sqrt{474}} \right\rangle$ is a unit vector orthogonal to $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$.

Simplified, this vector becomes $\left\langle \frac{7\sqrt{474}}{474}, \frac{-10\sqrt{474}}{237}, \frac{-5\sqrt{474}}{474} \right\rangle$.

Exercise 11.4.8

Find a unit vector orthogonal to both $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$, where $\vec{\mathbf{a}} = \langle 4, 0, 3 \rangle$ and $\vec{\mathbf{b}} = \langle 1, 1, 4 \rangle$.

Hint

Normalize the cross product.

Answer

$\left\langle \frac{-3}{\sqrt{194}}, \frac{-13}{\sqrt{194}}, \frac{4}{\sqrt{194}} \right\rangle$ or, simplified as $\left\langle \frac{-3\sqrt{194}}{194}, \frac{-13\sqrt{194}}{194}, \frac{2\sqrt{194}}{97} \right\rangle$

To use the cross product for calculating areas, we state and prove the following theorem.

Area of a Parallelogram

If we locate vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ such that they form adjacent sides of a parallelogram, then the area of the parallelogram is given by $\|\vec{\mathbf{u}} \times \vec{\mathbf{v}}\|$ (Figure 11.4.5).

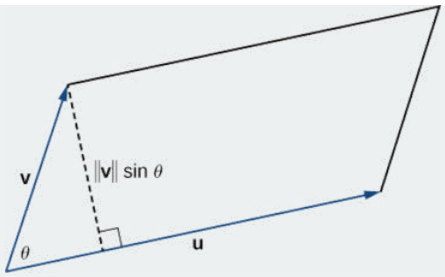


Figure 11.4.5: The parallelogram with adjacent sides $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ has base $\|\vec{\mathbf{u}}\|$ and height $\|\vec{\mathbf{v}}\| \sin\theta$.

Proof

We show that the magnitude of the cross product is equal to the base times height of the parallelogram.

$$\begin{aligned} \text{Area of a parallelogram} &= \text{base} \times \text{height} \\ &= \|\vec{\mathbf{u}}\|(\|\vec{\mathbf{v}}\| \sin\theta) \\ &= \|\vec{\mathbf{u}} \times \vec{\mathbf{v}}\| \end{aligned}$$

□

Example 11.4.9: Finding the Area of a Triangle

Let $P = (1, 0, 0), Q = (0, 1, 0)$, and $R = (0, 0, 1)$ be the vertices of a triangle (Figure 11.4.6). Find its area.

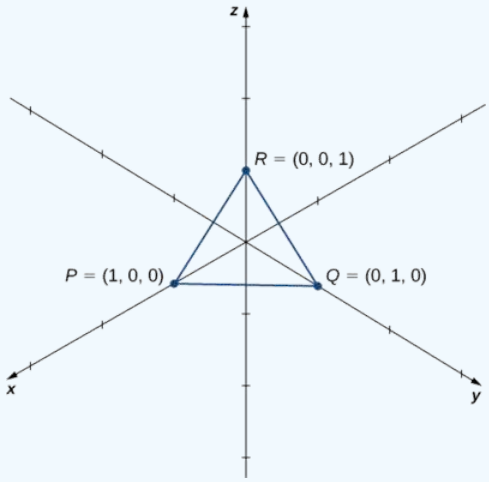


Figure 11.4.6: Finding the area of a triangle by using the cross product.

Solution

We have $\overrightarrow{PQ} = \langle 0-1, 1-0, 0-0\rangle = \langle -1, 1, 0\rangle$ and $\overrightarrow{PR} = \langle 0-1, 0-0, 1-0\rangle = \langle -1, 0, 1\rangle$. The area of the parallelogram with adjacent sides $\overrightarrow{PQ}$ and $\overrightarrow{PR}$ is given by $\left\|\overrightarrow{PQ} \times \overrightarrow{PR}\right\|$:

$$\overrightarrow{PQ} \times \overrightarrow{PR} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{vmatrix}$$
$$= (1-0)\hat{\mathbf{i}} - (-1-0)\hat{\mathbf{j}} + (0-(-1))\hat{\mathbf{k}}$$
$$= \hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}$$

$$\left\| \overrightarrow{PQ} \times \overrightarrow{PR} \right\| = \| \langle 1, 1, 1 \rangle \|$$
$$= \sqrt{1^2 + 1^2 + 1^2}$$
$$= \sqrt{3}.$$

The area of ΔPQR is half the area of the parallelogram or $\sqrt{3}/2$ units^2.

Exercise 11.4.9

Find the area of the parallelogram $PQRS$ with vertices $P(1, 1, 0)$, $Q(7, 1, 0)$, $R(9, 4, 2)$, and $S(3, 4, 2)$.

Hint

Sketch the parallelogram and identify two vectors that form adjacent sides of the parallelogram.

Answer

$6\sqrt{13}$ units^2

The Triple Scalar Product

Because the cross product of two vectors is a vector, it is possible to combine the dot product and the cross product. The dot product of a vector with the cross product of two other vectors is called the **triple scalar product** because the result is a scalar.

Definition: Triple Scalar Product

The triple scalar product of vectors $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$ is

$$\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}).$$

Calculating a Triple Scalar Product

The triple scalar product of vectors

$$\vec{\mathbf{u}} = u_1 \hat{\mathbf{i}} + u_2 \hat{\mathbf{j}} + u_3 \hat{\mathbf{k}}$$
$$\vec{\mathbf{v}} = v_1 \hat{\mathbf{i}} + v_2 \hat{\mathbf{j}} + v_3 \hat{\mathbf{k}}$$

and

$$\vec{\mathbf{w}} = w_1 \hat{\mathbf{i}} + w_2 \hat{\mathbf{j}} + w_3 \hat{\mathbf{k}}$$

is the determinant of the 3×3 matrix formed by the components of the vectors:

$$\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}. \tag{11.4.5}$$

Proof

The calculation is straightforward.

$$\begin{aligned}\vec{\mathbf{u}}\cdot(\vec{\mathbf{v}}\times\vec{\mathbf{w}}) &= \langle u_1,u_2,u_3\rangle\cdot\langle v_2w_3-v_3w_2,-v_1w_3+v_3w_1,v_1w_2-v_2w_1\rangle\\ &= u_1(v_2w_3-v_3w_2)+u_2(-v_1w_3+v_3w_1)+u_3(v_1w_2-v_2w_1)\\ &= u_1(v_2w_3-v_3w_2)-u_2(v_1w_3-v_3w_1)+u_3(v_1w_2-v_2w_1)\\ &= \begin{vmatrix} u_1 & u_2 & u_3\\ v_1 & v_2 & v_3\\ w_1 & w_2 & w_3\end{vmatrix}.\end{aligned}$$

□

Example 11.4.10: Calculating the Triple Scalar Product

Let $\vec{\mathbf{u}}=\langle 1,3,5\rangle$, $\vec{\mathbf{v}}=\langle 2,-1,0\rangle$ and $\vec{\mathbf{w}}=\langle -3,0,-1\rangle$. Calculate the triple scalar product $\vec{\mathbf{u}}\cdot(\vec{\mathbf{v}}\times\vec{\mathbf{w}})$.

Solution

Apply Equation 11.4.5 directly:

$$\begin{aligned}\vec{\mathbf{u}}\cdot(\vec{\mathbf{v}}\times\vec{\mathbf{w}}) &= \begin{vmatrix} 1 & 3 & 5\\ 2 & -1 & 0\\ -3 & 0 & -1\end{vmatrix}\\ &= 1\begin{vmatrix} -1 & 0\\ 0 & -1\end{vmatrix}-3\begin{vmatrix} 2 & 0\\ -3 & -1\end{vmatrix}+5\begin{vmatrix} 2 & -1\\ -3 & 0\end{vmatrix}\\ &= (1-0)-3(-2-0)+5(0-3)\\ &= 1+6-15=-8.\end{aligned}$$

Exercise 11.4.10

Calculate the triple scalar product $\vec{\mathbf{a}}\cdot(\vec{\mathbf{b}}\times\vec{\mathbf{c}})$, where $\vec{\mathbf{a}}=\langle 2,-4,1\rangle$, $\vec{\mathbf{b}}=\langle 0,3,-1\rangle$, and $\vec{\mathbf{c}}=\langle 5,-3,3\rangle$.

Hint

Place the vectors as the rows of a 3×3 matrix, then calculate the determinant.

Answer

17

When we create a matrix from three vectors, we must be careful about the order in which we list the vectors. If we list them in a matrix in one order and then rearrange the rows, the absolute value of the determinant remains unchanged. However, each time two rows switch places, the determinant changes sign:

$$\begin{vmatrix} a_1 & a_2 & a_3\\ b_1 & b_2 & b_3\\ c_1 & c_2 & c_3\end{vmatrix}=d \qquad \begin{vmatrix} b_1 & b_2 & b_3\\ a_1 & a_2 & a_3\\ c_1 & c_2 & c_3\end{vmatrix}=-d \qquad \begin{vmatrix} b_1 & b_2 & b_3\\ c_1 & c_2 & c_3\\ a_1 & a_2 & a_3\end{vmatrix}=d \qquad \begin{vmatrix} c_1 & c_2 & c_3\\ b_1 & b_2 & b_3\\ a_1 & a_2 & a_3\end{vmatrix}=-d$$

Verifying this fact is straightforward, but rather messy. Let's take a look at this with an example:

$$\begin{aligned}\begin{vmatrix} 1 & 2 & 1\\ -2 & 0 & 3\\ 4 & 1 & -1\end{vmatrix} &= \begin{vmatrix} 0 & 3\\ 1 & -1\end{vmatrix}-2\begin{vmatrix} -2 & 3\\ 4 & -1\end{vmatrix}+\begin{vmatrix} -2 & 0\\ 4 & 1\end{vmatrix}\\ &= (0-3)-2(2-12)+(-2-0)\\ &= -3+20-2=15.\end{aligned}$$

Switching the top two rows we have

$$\begin{vmatrix} -2 & 0 & 3 \\ 1 & 2 & 1 \\ 4 & 1 & -1 \end{vmatrix} = -2\begin{vmatrix} 2 & 1 \\ 1 & -1 \end{vmatrix} + 3\begin{vmatrix} 1 & 2 \\ 4 & 1 \end{vmatrix}$$
$$= -2(-2-1) + 3(1-8)$$
$$= 6 - 21 = -15.$$

Rearranging vectors in the triple products is equivalent to reordering the rows in the matrix of the determinant. Let $\vec{\mathbf{u}} = u_1\,\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}}$, $\vec{\mathbf{v}} = v_1\,\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}}$, and $\vec{\mathbf{w}} = w_1\,\hat{\mathbf{i}} + w_2\hat{\mathbf{j}} + w_3\hat{\mathbf{k}}$. Applying Calculating a Triple Scalar Product, we have

$$\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}$$

and

$$\vec{\mathbf{u}} \cdot (\vec{\mathbf{w}} \times \vec{\mathbf{v}}) = \begin{vmatrix} u_1 & u_2 & u_3 \\ w_1 & w_2 & w_3 \\ v_1 & v_2 & v_3 \end{vmatrix}.$$

We can obtain the determinant for calculating $\vec{\mathbf{u}} \cdot (\vec{\mathbf{w}} \times \vec{\mathbf{v}})$ by switching the bottom two rows of $\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}})$. Therefore, $\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) = -\vec{\mathbf{u}} \cdot (\vec{\mathbf{w}} \times \vec{\mathbf{v}})$.

Following this reasoning and exploring the different ways we can interchange variables in the triple scalar product lead to the following identities:

$$\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) = -\vec{\mathbf{u}} \cdot (\vec{\mathbf{w}} \times \vec{\mathbf{v}}) \tag{11.4.6}$$

$$\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) = \vec{\mathbf{v}} \cdot (\vec{\mathbf{w}} \times \vec{\mathbf{u}}) = \vec{\mathbf{w}} \cdot (\vec{\mathbf{u}} \times \vec{\mathbf{v}}). \tag{11.4.7}$$

Let $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ be two vectors in standard position. If $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ are not scalar multiples of each other, then these vectors form adjacent sides of a parallelogram. We saw in Area of a Parallelogram that the area of this parallelogram is $\|\vec{\mathbf{u}} \times \vec{\mathbf{v}}\|$. Now suppose we add a third vector $\vec{\mathbf{w}}$ that does not lie in the same plane as $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ but still shares the same initial point. Then these vectors form three edges of a **parallelepiped**, a three-dimensional prism with six faces that are each parallelograms, as shown in Figure 11.4.7. The volume of this prism is the product of the figure's height and the area of its base. The triple scalar product of $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$ provides a simple method for calculating the volume of the parallelepiped defined by these vectors.

Volume of a Parallelepiped

The volume of a parallelepiped with adjacent edges given by the vectors $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$ is the absolute value of the triple scalar product (Figure 11.4.7):

$$V = |\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}})|.$$

Note that, as the name indicates, the triple scalar product produces a scalar. The volume formula just presented uses the absolute value of a scalar quantity.

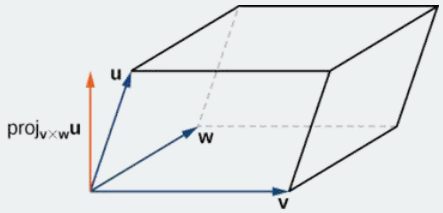


Figure 11.4.7: The height of the parallelepiped is given by $\|\text{proj}_{\vec{\mathbf{v}}\times\vec{\mathbf{w}}}\vec{\mathbf{u}}\|$.

Proof

The area of the base of the parallelepiped is given by $\|\vec{\mathbf{v}} \times \vec{\mathbf{w}}\|$. The height of the figure is given by $\|\text{proj}_{\vec{\mathbf{v}} \times \vec{\mathbf{w}}} \vec{\mathbf{u}}\|$. The volume of the parallelepiped is the product of the height and the area of the base, so we have

$$\begin{aligned} V &= \| \text{proj}_{\vec{\mathbf{v}} \times \vec{\mathbf{w}}} \vec{\mathbf{u}} \| \, \|\vec{\mathbf{v}} \times \vec{\mathbf{w}}\| \\ &= \left| \frac{\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}})}{\|\vec{\mathbf{v}} \times \vec{\mathbf{w}}\|} \right| \|\vec{\mathbf{v}} \times \vec{\mathbf{w}}\| \\ &= |\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}})|. \end{aligned}$$

□

Example 11.4.11: Calculating the Volume of a Parallelepiped

Let $\vec{\mathbf{u}} = \langle -1, -2, 1 \rangle$, $\vec{\mathbf{v}} = \langle 4, 3, 2 \rangle$, and $\vec{\mathbf{w}} = \langle 0, -5, -2 \rangle$. Find the volume of the parallelepiped with adjacent edges $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$ (Figure 11.4.8).

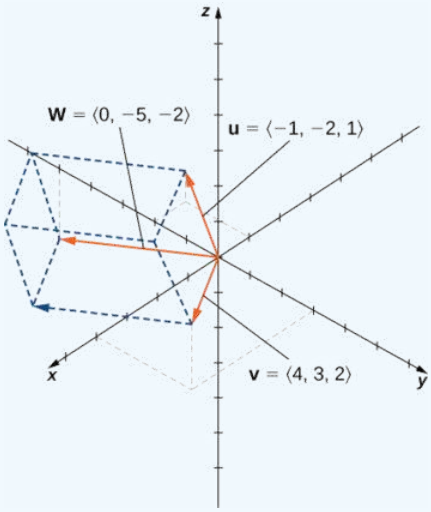


Figure 11.4.8

Solution

We have

$$\begin{aligned} \vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) &= \begin{vmatrix} -1 & -2 & 1 \\ 4 & 3 & 2 \\ 0 & -5 & -2 \end{vmatrix} \\ &= (-1) \begin{vmatrix} 3 & 2 \\ -5 & -2 \end{vmatrix} + 2 \begin{vmatrix} 4 & 2 \\ 0 & -2 \end{vmatrix} + \begin{vmatrix} 4 & 3 \\ 0 & -5 \end{vmatrix} \\ &= (-1)(-6+10) + 2(-8-0) + (-20-0) \\ &= -4 - 16 - 20 \\ &= -40. \end{aligned}$$

Thus, the volume of the parallelepiped is $|-40| = 40$ units^3

Exercise 11.4.11

Find the volume of the parallelepiped formed by the vectors $\vec{\mathbf{a}} = 3\hat{\mathbf{i}} + 4\hat{\mathbf{j}} - \hat{\mathbf{k}}$, $\vec{\mathbf{b}} = 2\hat{\mathbf{i}} - \hat{\mathbf{j}} - \hat{\mathbf{k}}$, and $\vec{\mathbf{c}} = 3\hat{\mathbf{j}} + \hat{\mathbf{k}}$.

Hint

Calculate the triple scalar product by finding a determinant.

Answer

8 units^3

Applications of the Cross Product

The cross product appears in many practical applications in mathematics, physics, and engineering. Let's examine some of these applications here, including the idea of torque, with which we began this section. Other applications show up in later chapters, particularly in our study of vector fields such as gravitational and electromagnetic fields (Introduction to Vector Calculus).

Example 11.4.12: Using the Triple Scalar Product

Use the triple scalar product to show that vectors $\vec{\mathbf{u}} = \langle 2, 0, 5\rangle$, $\vec{\mathbf{v}} = \langle 2, 2, 4\rangle$, and $\vec{\mathbf{w}} = \langle 1, -1, 3\rangle$ are coplanar—that is, show that these vectors lie in the same plane.

Solution

Start by calculating the triple scalar product to find the volume of the parallelepiped defined by $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$:

$$\begin{aligned}\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) &= \begin{vmatrix} 2 & 0 & 5 \\ 2 & 2 & 4 \\ 1 & -1 & 3 \end{vmatrix} \\ &= [2(2)(3) + (0)(4)(1) + 5(2)(-1)] - [5(2)(1) + (2)(4)(-1) + (0)(2)(3)] \\ &= 2 - 2 = 0.\end{aligned}$$

The volume of the parallelepiped is 0 units^3, so one of the dimensions must be zero. Therefore, the three vectors all lie in the same plane.

Exercise 11.4.12

Are the vectors $\vec{\mathbf{a}} = \hat{\mathbf{i}} + \hat{\mathbf{j}} - \hat{\mathbf{k}}$, $\vec{\mathbf{b}} = \hat{\mathbf{i}} - \hat{\mathbf{j}} + \hat{\mathbf{k}}$, and $\vec{\mathbf{c}} = \hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}$ coplanar?

Hint

Calculate the triple scalar product.

Answer

No, the triple scalar product is $-4 \neq 0$, so the three vectors form the adjacent edges of a parallelepiped. They are not coplanar.

Example 11.4.13: Finding an Orthogonal Vector

Only a single plane can pass through any set of three noncolinear points. Find a vector orthogonal to the plane containing points $P = (9, -3, -2)$, $Q = (1, 3, 0)$, and $R = (-2, 5, 0)$.

Solution

The plane must contain vectors $\overrightarrow{PQ}$ and $\overrightarrow{QR}$:

$$\overrightarrow{PQ} = \langle 1 - 9, 3 - (-3), 0 - (-2)\rangle = \langle -8, 6, 2\rangle$$

$$\overrightarrow{QR} = \langle -2 - 1, 5 - 3, 0 - 0\rangle = \langle -3, 2, 0\rangle.$$

The cross product $\overrightarrow{PQ} \times \overrightarrow{QR}$ produces a vector orthogonal to both $\overrightarrow{PQ}$ and $\overrightarrow{QR}$. Therefore, the cross product is orthogonal to the plane that contains these two vectors:

$$\overrightarrow{PQ} \times \overrightarrow{QR} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ -8 & 6 & 2 \\ -3 & 2 & 0 \end{vmatrix}$$
$$= 0\hat{\mathbf{i}} - 6\hat{\mathbf{j}} - 16\hat{\mathbf{k}} - (-18\hat{\mathbf{k}} + 4\hat{\mathbf{i}} + 0\hat{\mathbf{j}})$$
$$= -4\hat{\mathbf{i}} - 6\hat{\mathbf{j}} + 2\hat{\mathbf{k}}.$$

We have seen how to use the triple scalar product and how to find a vector orthogonal to a plane. Now we apply the cross product to real-world situations.

Sometimes a force causes an object to rotate. For example, turning a screwdriver or a wrench creates this kind of rotational effect, called torque.

Definition: Torque

Torque, $\vec{\tau}$ (the Greek letter tau), measures the tendency of a force to produce rotation about an axis of rotation. Let $\vec{\mathbf{r}}$ be a vector with an initial point located on the axis of rotation and with a terminal point located at the point where the force is applied, and let vector $\vec{\mathbf{F}}$ represent the force. Then torque is equal to the cross product of $\vec{\mathbf{r}}$ and $\vec{\mathbf{F}}$:

$$\vec{\tau} = \vec{\mathbf{r}} \times \vec{\mathbf{F}}.$$

See Figure 11.4.9.

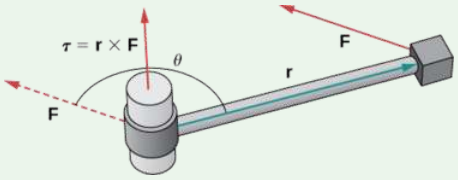


Figure 11.4.9: Torque measures how a force causes an object to rotate.

Think about using a wrench to tighten a bolt. The torque τ applied to the bolt depends on how hard we push the wrench (force) and how far up the handle we apply the force (distance). The torque increases with a greater force on the wrench at a greater distance from the bolt. Common units of torque are the newton-meter or foot-pound. Although torque is dimensionally equivalent to work (it has the same units), the two concepts are distinct. Torque is used specifically in the context of rotation, whereas work typically involves motion along a line.

Example 11.4.14: Evaluating Torque

A bolt is tightened by applying a force of 6 N to a 0.15-m wrench (Figure 11.4.10). The angle between the wrench and the force vector is 40°. Find the magnitude of the torque about the center of the bolt. Round the answer to two decimal places.

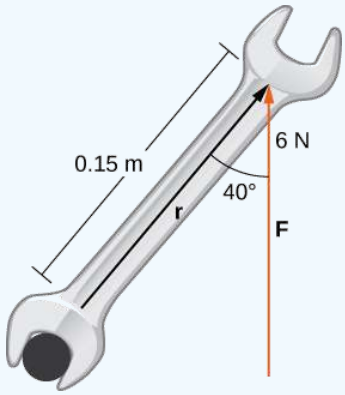


Figure 11.4.10: Torque describes the twisting action of the wrench.

Solution:

Substitute the given information into the equation defining torque:

$$\begin{aligned}\|\vec{\boldsymbol{\tau}}\| &= \|\vec{\mathbf{r}} \times \vec{\mathbf{F}}\| \\ &= \|\vec{\mathbf{r}}\| \, \|\vec{\mathbf{F}}\| \sin\theta \\ &= (0.15\,\text{m})(6\,\text{N}) \sin 40^\circ \\ &\approx 0.58\,\text{N·m}.\end{aligned}$$

? Exercise 11.4.14

Calculate the force required to produce 15 N·m torque at an angle of 30° from a 150-cm rod.

Hint

$\|\vec{\boldsymbol{\tau}}\| = 15$ N·m and $\|\vec{\mathbf{r}}\| = 1.5$ m

Answer

20 N

Key Concepts

- The cross product $\vec{\mathbf{u}} \times \vec{\mathbf{v}}$ of two vectors $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$ is a vector orthogonal to both $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$. Its length is given by $\|\vec{\mathbf{u}} \times \vec{\mathbf{v}}\| = \|\vec{\mathbf{u}}\| \cdot \|\vec{\mathbf{v}}\| \cdot \sin\theta$, where θ is the angle between $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$. Its direction is given by the right-hand rule.
- The algebraic formula for calculating the cross product of two vectors,

  $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$, is

  $\vec{\mathbf{u}} \times \vec{\mathbf{v}} = (u_2 v_3 - u_3 v_2)\hat{\mathbf{i}} - (u_1 v_3 - u_3 v_1)\hat{\mathbf{j}} + (u_1 v_2 - u_2 v_1)\hat{\mathbf{k}}$.
- The cross product satisfies the following properties for vectors $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$, and scalar c:

  $$\vec{\mathbf{u}} \times \vec{\mathbf{v}} = -(\vec{\mathbf{v}} \times \vec{\mathbf{u}})$$
  $$\vec{\mathbf{u}} \times (\vec{\mathbf{v}} + \vec{\mathbf{w}}) = \vec{\mathbf{u}} \times \vec{\mathbf{v}} + \vec{\mathbf{u}} \times \vec{\mathbf{w}}$$
  $$c(\vec{\mathbf{u}} \times \vec{\mathbf{v}}) = (c\vec{\mathbf{u}}) \times \vec{\mathbf{v}} = \vec{\mathbf{u}} \times (c\vec{\mathbf{v}})$$
  $$\vec{\mathbf{u}} \times \vec{\mathbf{0}} = \vec{\mathbf{0}} \times \vec{\mathbf{u}} = \vec{\mathbf{0}}$$
  $$\vec{\mathbf{v}} \times \vec{\mathbf{v}} = \vec{\mathbf{0}}$$
  $$\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}}) = (\vec{\mathbf{u}} \times \vec{\mathbf{v}}) \cdot \vec{\mathbf{w}}$$
- The cross product of vectors $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$ is the determinant $\begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$
- If vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ form adjacent sides of a parallelogram, then the area of the parallelogram is given by $\|\vec{\mathbf{u}} \times \vec{\mathbf{v}}\|$.
- The triple scalar product of vectors $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$ is $\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}})$.
- The volume of a parallelepiped with adjacent edges given by vectors $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$ is $V = |\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}})|$.
- If the triple scalar product of vectors $\vec{\mathbf{u}}$, $\vec{\mathbf{v}}$, and $\vec{\mathbf{w}}$ is zero, then the vectors are coplanar. The converse is also true: If the vectors are coplanar, then their triple scalar product is zero.
- The cross product can be used to identify a vector orthogonal to two given vectors or to a plane.
- Torque $\vec{\boldsymbol{\tau}}$ measures the tendency of a force to produce rotation about an axis of rotation. If force $\vec{\mathbf{F}}$ is acting at a distance (displacement) $\vec{\mathbf{r}}$ from the axis, then torque is equal to the cross product of $\vec{\mathbf{r}}$ and $\vec{\mathbf{F}}$: $\vec{\boldsymbol{\tau}} = \vec{\mathbf{r}} \times \vec{\mathbf{F}}$.

Key Equations

- **The cross product of two vectors in terms of the unit vectors**

$$\vec{\mathbf{u}} \times \vec{\mathbf{v}} = (u_2 v_3 - u_3 v_2)\hat{\mathbf{i}} - (u_1 v_3 - u_3 v_1)\hat{\mathbf{j}} + (u_1 v_2 - u_2 v_1)\hat{\mathbf{k}}$$

Glossary

cross product

$\vec{\mathbf{u}} \times \vec{\mathbf{v}} = (u_2v_3 - u_3v_2)\hat{\mathbf{i}} - (u_1v_3 - u_3v_1)\hat{\mathbf{j}} + (u_1v_2 - u_2v_1)\hat{\mathbf{k}}$, where $\vec{\mathbf{u}} = \langle u_1, u_2, u_3 \rangle$ and $\vec{\mathbf{v}} = \langle v_1, v_2, v_3 \rangle$

determinant

a real number associated with a square matrix

parallelepiped

a three-dimensional prism with six faces that are parallelograms

torque

the effect of a force that causes an object to rotate

triple scalar product

the dot product of a vector with the cross product of two other vectors: $\vec{\mathbf{u}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{w}})$

vector product

the cross product of two vectors

11.4: The Cross Product is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **12.4: The Cross Product** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

11.5: Lines and Planes in Space

Learning Objectives

- Write the vector, parametric, and symmetric equations of a line through a given point in a given direction, and a line through two given points.
- Find the distance from a point to a given line.
- Write the vector and scalar equations of a plane through a given point with a given normal.
- Find the distance from a point to a given plane.
- Find the angle between two planes.

By now, we are familiar with writing equations that describe a line in two dimensions. To write an equation for a line, we must know two points on the line, or we must know the direction of the line and at least one point through which the line passes. In two dimensions, we use the concept of slope to describe the orientation, or direction, of a line. In three dimensions, we describe the direction of a line using a vector parallel to the line. In this section, we examine how to use equations to describe lines and planes in space.

Equations for a Line in Space

Let's first explore what it means for two vectors to be parallel. Recall that parallel vectors must have the same or opposite directions. If two nonzero vectors, $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$, are parallel, we claim there must be a scalar, k, such that $\vec{\mathbf{u}} = k\vec{\mathbf{v}}$. If $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ have the same direction, simply choose

$$k = \frac{\|\vec{\mathbf{u}}\|}{\|\vec{\mathbf{v}}\|}.$$

If $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ have opposite directions, choose

$$k = -\frac{\|\vec{\mathbf{u}}\|}{\|\vec{\mathbf{v}}\|}.$$

Note that the converse holds as well. If $\vec{\mathbf{u}} = k\vec{\mathbf{v}}$ for some scalar k, then either $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ have the same direction $(k > 0)$ or opposite directions $(k < 0)$, so $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ are parallel. Therefore, two nonzero vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ are parallel if and only if $\vec{\mathbf{u}} = k\vec{\mathbf{v}}$ for some scalar k. By convention, the zero vector $\vec{\mathbf{0}}$ is considered to be parallel to all vectors.

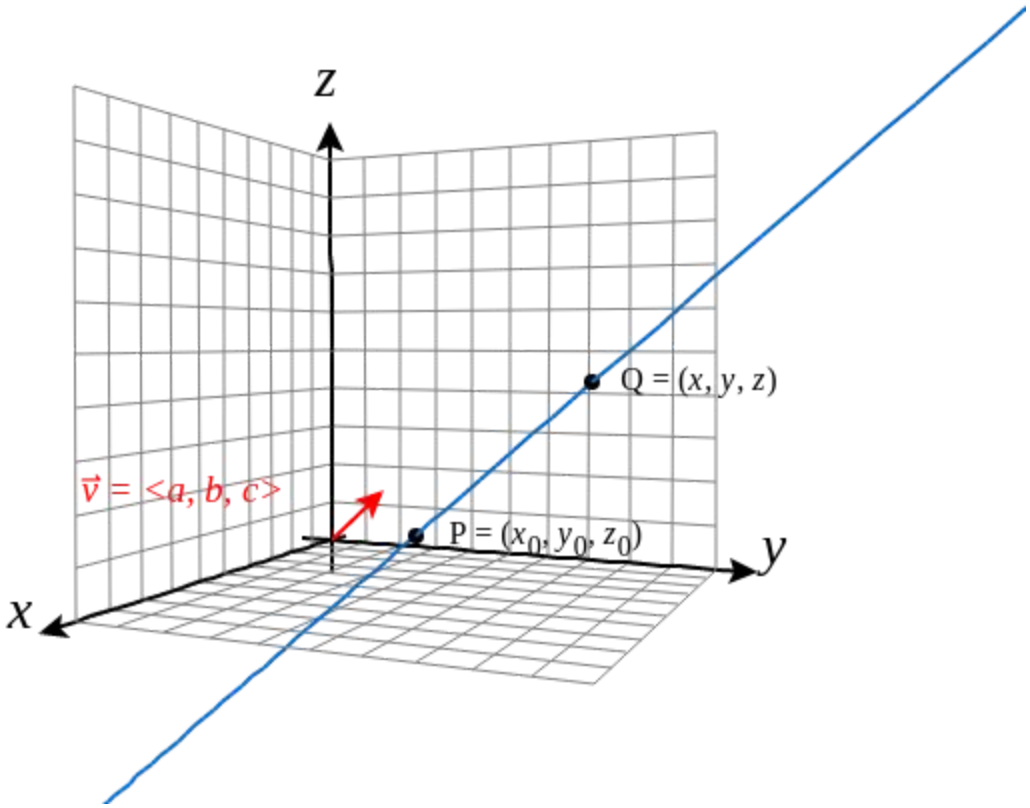


Figure 11.5.1: Vector $\vec{\mathbf{v}}$ is the direction vector for $\overrightarrow{PQ}$.

As in two dimensions, we can describe a line in space using a point on the line and the direction of the line, or a parallel vector, which we call the **direction vector** (Figure 11.5.1). Let L be a line in space passing through point $P(x_0, y_0, z_0)$. Let $\vec{\mathbf{v}} = \langle a, b, c \rangle$ be a vector parallel to L. Then, for any point on line $Q(x, y, z)$, we know that $\overrightarrow{PQ}$ is parallel to $\vec{\mathbf{v}}$. Thus, as we just discussed, there is a scalar, t, such that $\overrightarrow{PQ} = t\vec{\mathbf{v}}$, which gives

$$\begin{aligned} \overrightarrow{PQ} &= t\vec{\mathbf{v}} \\ \langle x - x_0, y - y_0, z - z_0 \rangle &= t\langle a, b, c \rangle \\ \langle x - x_0, y - y_0, z - z_0 \rangle &= \langle ta, tb, tc \rangle. \end{aligned} \tag{11.5.1}$$

Using vector operations, we can rewrite Equation 11.5.1

$$\begin{aligned} \langle x - x_0, y - y_0, z - z_0 \rangle &= \langle ta, tb, tc \rangle \\ \langle x, y, z \rangle - \langle x_0, y_0, z_0 \rangle &= t\langle a, b, c \rangle \\ \underbrace{\langle x, y, z \rangle}_{\vec{\mathbf{r}}} &= \underbrace{\langle x_0, y_0, z_0 \rangle}_{\vec{\mathbf{r}}_0} + t\underbrace{\langle a, b, c \rangle}_{\vec{\mathbf{v}}}. \end{aligned}$$

Setting $\vec{\mathbf{r}} = \langle x, y, z \rangle$ and $\vec{\mathbf{r}}_0 = \langle x_0, y_0, z_0 \rangle$, we now have the **vector equation of a line:**

$$\vec{\mathbf{r}} = \vec{\mathbf{r}}_0 + t\vec{\mathbf{v}}. \tag{11.5.2}$$

Equating components, Equation 11.5.2 shows that the following equations are simultaneously true: $x - x_0 = ta, y - y_0 = tb,$ and $z - z_0 = tc$. If we solve each of these equations for the component variables $x, y,$ and z, we get a set of equations in which each

variable is defined in terms of the parameter t and that, together, describe the line. This set of three equations forms a set of **parametric equations of a line:**

$$x = x_0 + ta$$
$$y = y_0 + tb$$
$$z = z_0 + tc.$$

If we solve each of the equations for t assuming a, b, and c are nonzero, we get a different description of the same line:

$$\frac{x - x_0}{a} = t$$
$$\frac{y - y_0}{b} = t$$
$$\frac{z - z_0}{c} = t.$$

Because each expression equals t, they all have the same value. We can set them equal to each other to create **symmetric equations of a line:**

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c}.$$

We summarize the results in the following theorem.

Theorem: Parametric and Symmetric Equations of a Line

A line L parallel to vector $\vec{\mathbf{v}} = \langle a, b, c \rangle$ and passing through point $P(x_0, y_0, z_0)$ can be described by the following parametric equations:

$$x = x_0 + ta, y = y_0 + tb,$$

and

$$z = z_0 + tc.$$

If the constants a, b, and c are all nonzero, then L can be described by the symmetric equation of the line:

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c}.$$

The parametric equations of a line are not unique. Using a different parallel vector or a different point on the line leads to a different, equivalent representation. Each set of parametric equations leads to a related set of symmetric equations, so it follows that a symmetric equation of a line is not unique either.

Example 11.5.1: Equations of a Line in Space

Find parametric and symmetric equations of the line passing through points $(1, 4, -2)$ and $(-3, 5, 0)$.

Solution

First, identify a vector parallel to the line:

$$\vec{\mathbf{v}} = \langle -3 - 1, 5 - 4, 0 - (-2) \rangle = \langle -4, 1, 2 \rangle.$$

Use either of the given points on the line to complete the parametric equations:

$$x = 1 - 4t$$
$$y = 4 + t,$$

and

$$z = -2 + 2t.$$

Solve each equation for t to create the symmetric equation of the line:

$$\frac{x-1}{-4} = y-4 = \frac{z+2}{2}.$$

Exercise 11.5.1

Find parametric and symmetric equations of the line passing through points $(1,-3,2)$ and $(5,-2,8)$.

Hint:

Start by finding a vector parallel to the line.

Answer

Possible set of parametric equations: $x = 1+4t,\ y = -3+t,\ z = 2+6t;$ related set of symmetric equations:

$$\frac{x-1}{4} = y+3 = \frac{z-2}{6}$$

Sometimes we don't want the equation of a whole line, just a line segment. In this case, we limit the values of our parameter t. For example, let $P(x_0,y_0,z_0)$ and $Q(x_1,y_1,z_1)$ be points on a line, and let $\vec{\mathbf{p}} = \langle x_0,y_0,z_0\rangle$ and $\vec{\mathbf{q}} = \langle x_1,y_1,z_1\rangle$ be the associated position vectors. In addition, let $\vec{\mathbf{r}} = \langle x,y,z\rangle$. We want to find a vector equation for the line segment between P and Q. Using P as our known point on the line, and $\overrightarrow{PQ} = \langle x_1-x_0, y_1-y_0, z_1-z_0\rangle$ as the direction vector equation, Equation 11.5.2 gives

$$\vec{\mathbf{r}} = \vec{\mathbf{p}} + t(\overrightarrow{PQ}). \tag{11.5.3}$$

Equation 11.5.3 can be expanded using properties of vectors:

$$\begin{aligned}
\vec{\mathbf{r}} &= \vec{\mathbf{p}} + t(\overrightarrow{PQ}) \\
&= \langle x_0,y_0,z_0\rangle + t\langle x_1-x_0, y_1-y_0, z_1-z_0\rangle \\
&= \langle x_0,y_0,z_0\rangle + t(\langle x_1,y_1,z_1\rangle - \langle x_0,y_0,z_0\rangle) \\
&= \langle x_0,y_0,z_0\rangle + t\langle x_1,y_1,z_1\rangle - t\langle x_0,y_0,z_0\rangle \\
&= (1-t)\langle x_0,y_0,z_0\rangle + t\langle x_1,y_1,z_1\rangle \\
&= (1-t)\vec{\mathbf{p}} + t\vec{\mathbf{q}}.
\end{aligned}$$

Thus, the vector equation of the line passing through P and Q is

$$\vec{\mathbf{r}} = (1-t)\vec{\mathbf{p}} + t\vec{\mathbf{q}}.$$

Remember that we did not want the equation of the whole line, just the line segment between P and Q. Notice that when $t=0$, we have $\vec{\mathbf{r}} = \vec{\mathbf{p}}$, and when $t=1$, we have $\vec{\mathbf{r}} = \vec{\mathbf{q}}$. Therefore, the vector equation of the line segment between P and Q is

$$\vec{\mathbf{r}} = (1-t)\vec{\mathbf{p}} + t\vec{\mathbf{q}}, 0 \le t \le 1.$$

Going back to Equation 11.5.2, we can also find parametric equations for this line segment. We have

$$\begin{aligned}
\vec{\mathbf{r}} &= \vec{\mathbf{p}} + t(\overrightarrow{PQ}) \\
\langle x,y,z\rangle &= \langle x_0,y_0,z_0\rangle + t\langle x_1-x_0, y_1-y_0, z_1-z_0\rangle \\
&= \langle x_0+t(x_1-x_0), y_0+t(y_1-y_0), z_0+t(z_1-z_0)\rangle.
\end{aligned}$$

Then, the parametric equations are

$$\begin{aligned}
x &= x_0 + t(x_1-x_0) \\
y &= y_0 + t(y_1-y_0) \\
z &= z_0 + t(z_1-z_0), \quad 0 \le t \le 1.
\end{aligned} \tag{11.5.4}$$

Example 11.5.2: Parametric Equations of a Line Segment

Find parametric equations of the line segment between the points $P(2,1,4)$ and $Q(3,-1,3)$.

Solution

Start with the parametric equations for a line (Equations 11.5.4) and work with each component separately:

$$\begin{aligned} x &= x_0 + t(x_1 - x_0) \\ &= 2 + t(3-2) \\ &= 2 + t, \\ y &= y_0 + t(y_1 - y_0) \\ &= 1 + t(-1-1) \\ &= 1 - 2t, \end{aligned}$$

and

$$\begin{aligned} z &= z_0 + t(z_1 - z_0) \\ &= 4 + t(3-4) \\ &= 4 - t. \end{aligned}$$

Therefore, the parametric equations for the line segment are

$$\begin{aligned} x &= 2 + t \\ y &= 1 - 2t \\ z &= 4 - t, \quad 0 \le t \le 1. \end{aligned}$$

Exercise 11.5.2

Find parametric equations of the line segment between points $P(-1,3,6)$ and $Q(-8,2,4)$.

Answer

$x = -1 - 7t,\ y = 3 - t,\ z = 6 - 2t, \quad 0 \le t \le 1$

Distance between a Point and a Line

We already know how to calculate the distance between two points in space. We now expand this definition to describe the distance between a point and a line in space. Several real-world contexts exist when it is important to be able to calculate these distances. When building a home, for example, builders must consider "setback" requirements, when structures or fixtures have to be a certain distance from the property line. Air travel offers another example. Airlines are concerned about the distances between populated areas and proposed flight paths.

Let L be a line in the plane and let M be any point not on the line. Then, we define distance d from M to L as the length of line segment $\overline{MP}$, where P is a point on L such that $\overline{MP}$ is perpendicular to L (Figure 11.5.2).

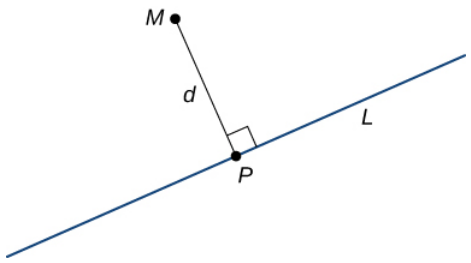


Figure 11.5.2: The distance from point M to line L is the length of $\overline{MP}$.

When we're looking for the distance between a line and a point in space, Figure 11.5.2 still applies. We still define the distance as the length of the perpendicular line segment connecting the point to the line. In space, however, there is no clear way to know which point on the line creates such a perpendicular line segment, so we select an arbitrary point on the line and use properties of vectors to calculate the distance. Therefore, let P be an arbitrary point on line L and let $\vec{\mathbf{v}}$ be a direction vector for L (Figure 11.5.3).

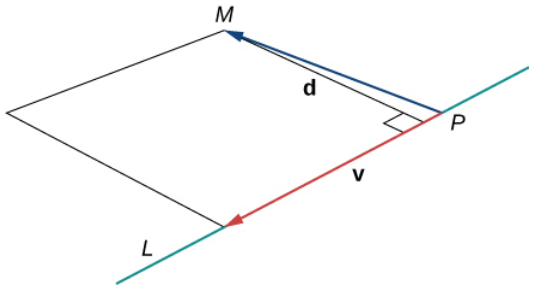


Figure 11.5.3: Vectors $\overrightarrow{PM}$ and $\vec{\mathbf{v}}$ form two sides of a parallelogram with base $\|\vec{\mathbf{v}}\|$ and height d, which is the distance between a line and a point in space.

Vectors $\overrightarrow{PM}$ and $\vec{\mathbf{v}}$ form two sides of a parallelogram with area $\|\overrightarrow{PM} \times \vec{\mathbf{v}}\|$. Using a formula from geometry, the area of this parallelogram can also be calculated as the product of its base and height:

$$\|\overrightarrow{PM} \times \vec{\mathbf{v}}\| = \|\vec{\mathbf{v}}\| d.$$

We can use this formula to find a general formula for the distance between a line in space and any point not on the line.

Distance from a Point to a Line

Let L be a line in space passing through point P with direction vector $\vec{\mathbf{v}}$. If M is any point not on L, then the distance from M to L is

$$d = \frac{\|\overrightarrow{PM} \times \vec{\mathbf{v}}\|}{\|\vec{\mathbf{v}}\|}.$$

Example 11.5.3: Calculating the Distance from a Point to a Line

Find the distance between the point $M = (1, 1, 3)$ and line $\dfrac{x-3}{4} = \dfrac{y+1}{2} = z - 3$.

Solution:

From the symmetric equations of the line, we know that vector $\vec{\mathbf{v}} = \langle 4, 2, 1 \rangle$ is a direction vector for the line. Setting the symmetric equations of the line equal to zero, we see that point $P(3, -1, 3)$ lies on the line. Then,

$$\begin{aligned} \overrightarrow{PM} &= \langle 1-3, 1-(-1), 3-3 \rangle \\ &= \langle -2, 2, 0 \rangle. \end{aligned}$$

To calculate the distance, we need to find $\overrightarrow{PM} \times \vec{\mathbf{v}}$:

$$\begin{aligned} \overrightarrow{PM} \times \vec{\mathbf{v}} &= \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ -2 & 2 & 0 \\ 4 & 2 & 1 \end{vmatrix} \\ &= (2-0)\hat{\mathbf{i}} - (-2-0)\hat{\mathbf{j}} + (-4-8)\hat{\mathbf{k}} \\ &= 2\hat{\mathbf{i}} + 2\hat{\mathbf{j}} - 12\hat{\mathbf{k}}. \end{aligned}$$

Therefore, the distance between the point and the line is (Figure 11.5.4)

$$d = \frac{\|\overrightarrow{PM} \times \vec{\mathbf{v}}\|}{\|\vec{\mathbf{v}}\|}$$
$$= \frac{\sqrt{2^2 + 2^2 + 12^2}}{\sqrt{4^2 + 2^2 + 1^2}}$$
$$= \frac{2\sqrt{38}}{\sqrt{21}}$$
$$= \frac{2\sqrt{798}}{21} \text{ units}$$

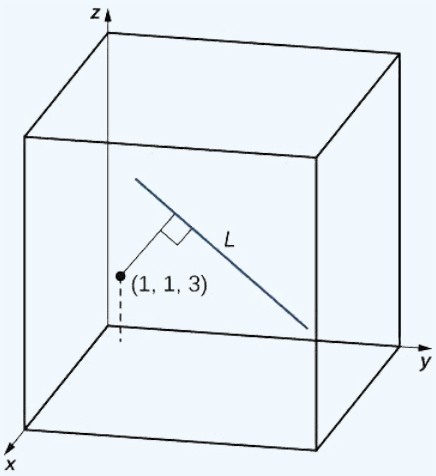


Figure 11.5.4: Point $(1, 1, 3)$ is approximately 2.7 units from the line with symmetric equations $\frac{x-3}{4} = \frac{y+1}{2} = z - 3$.

? Exercise 11.5.3

Find the distance between point $(0, 3, 6)$ and the line with parametric equations $x = 1 - t,\ y = 1 + 2t,\ z = 5 + 3t$.

Hint

Find a vector with initial point $(0, 3, 6)$ and a terminal point on the line, and then find a direction vector for the line.

Answer

$$\sqrt{\frac{10}{7}} = \frac{\sqrt{70}}{7} \text{ units}$$

Relationships between Lines

Given two lines in the two-dimensional plane, the lines are equal, they are parallel but not equal, or they intersect in a single point. In three dimensions, a fourth case is possible. If two lines in space are not parallel, but do not intersect, then the lines are said to be **skew lines** (Figure 11.5.5).

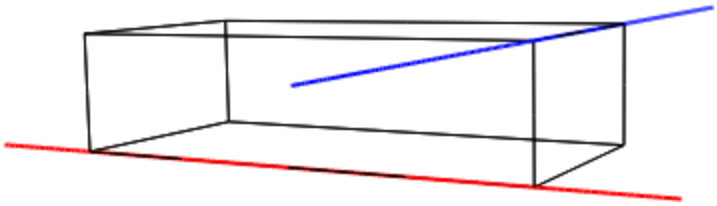


Figure 11.5.5: In three dimensions, it is possible that two lines do not cross, even when they have different directions.

To classify lines as parallel but not equal, equal, intersecting, or skew, we need to know two things: whether the direction vectors are parallel and whether the lines share a point (Figure 11.5.6).

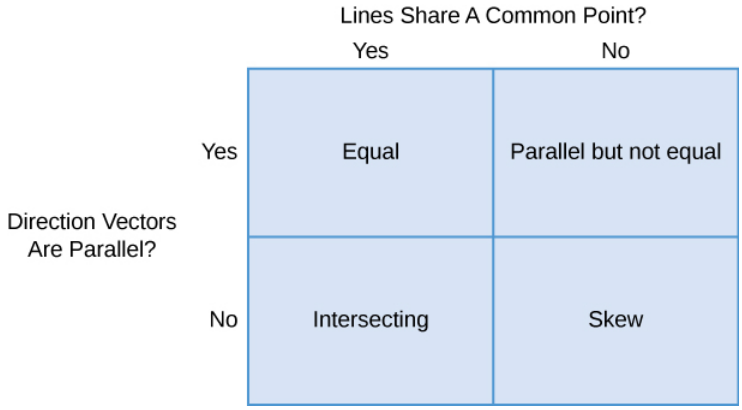


Figure 11.5.6: Determine the relationship between two lines based on whether their direction vectors are parallel and whether they share a point.

Example 11.5.4: Classifying Lines in Space

For each pair of lines, determine whether the lines are equal, parallel but not equal, skew, or intersecting.

a.

- $L_1: \ x = 2s - 1, \ y = s - 1, \ z = s - 4$

- $L_2: \ x = t - 3, \ y = 3t + 8, \ z = 5 - 2t$

b.

- $L_1: \ x = -y = z$
- $L_2: \ \dfrac{x-3}{2} = y = z - 2$

c.

- $L_1: \ x = 6s - 1, \ y = -2s, \ z = 3s + 1$
- $L_2: \ \dfrac{x-4}{6} = \dfrac{y+3}{-2} = \dfrac{z-1}{3}$

Solution

a. Line L_1 has direction vector $\vec{\mathbf{v}}_1 = \langle 2, 1, 1 \rangle$; line L_2 has direction vector $\vec{\mathbf{v}}_2 = \langle 1, 3, -2 \rangle$. Because the direction vectors are not parallel vectors, the lines are either intersecting or skew. To determine whether the lines intersect, we see if there is a point, (x, y, z), that lies on both lines. To find this point, we use the parametric equations to create a system of equalities:

$$2s - 1 = t - 3;$$

$$s - 1 = 3t + 8;$$

$$s - 4 = 5 - 2t.$$

By the first equation, $t = 2s + 2$. Substituting into the second equation yields

$$s - 1 = 3(2s + 2) + 8$$

$$s - 1 = 6s + 6 + 8$$

$$5s = -15$$

$$s = -3.$$

Substitution into the third equation, however, yields a contradiction:

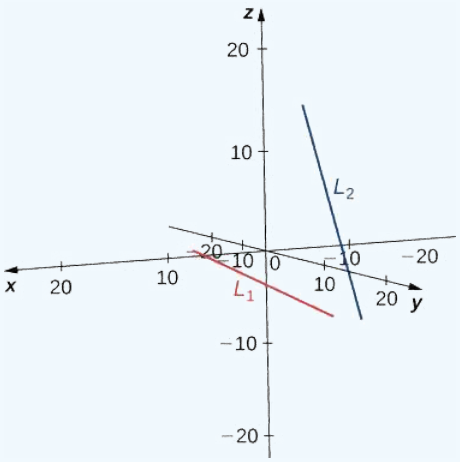
$$s - 4 = 5 - 2(2s + 2)$$

$$s - 4 = 5 - 4s - 4$$

$$5s = 5$$

$$s = 1.$$

There is no single point that satisfies the parametric equations for L_1 and L_2 simultaneously. These lines do not intersect, so they are skew (see the following figure).

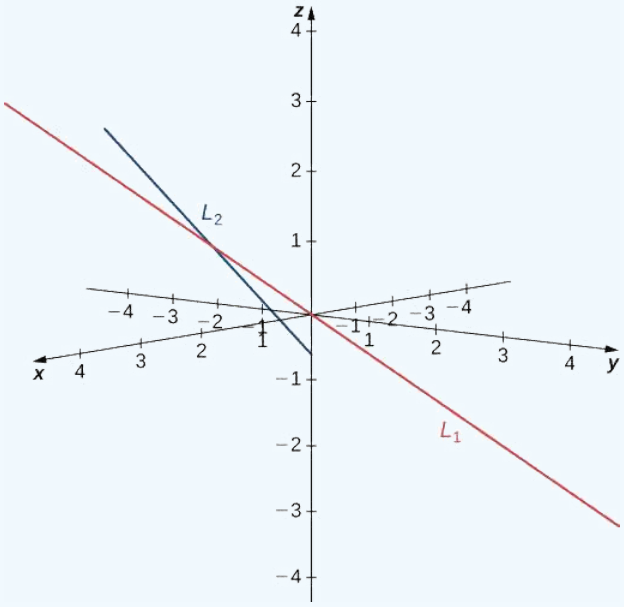


b. Line L_1 has direction vector $\vec{\mathbf{v}}_1 = \langle 1, -1, 1 \rangle$ and passes through the origin, $(0, 0, 0)$. Line L_2 has a different direction vector, $\vec{\mathbf{v}}_2 = \langle 2, 1, 1 \rangle$, so these lines are not parallel or equal. Let r represent the parameter for line L_1 and let s represent the

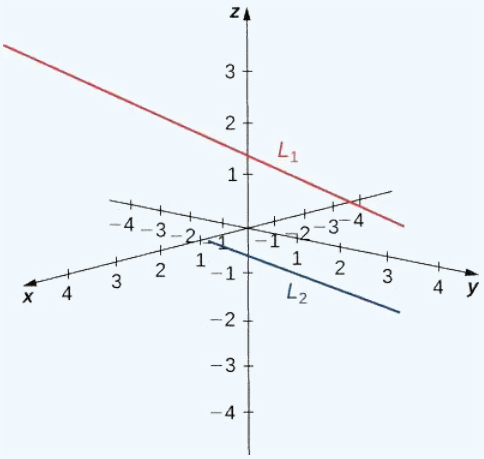
parameter for L_2:

$$\begin{array}{ll} \text{Line } L_1: & \text{Line } L_2: \\ x = r & x = 2s + 3 \\ y = -r & y = s \\ z = r & z = s + 2 \end{array}$$

Solve the system of equations to find $r = 1$ and $s = -1$. If we need to find the point of intersection, we can substitute these parameters into the original equations to get $(1, -1, 1)$ (see the following figure).



c. Lines L_1 and L_2 have equivalent direction vectors: $\vec{\mathbf{v}} = \langle 6, -2, 3 \rangle$. These two lines are parallel (see the following figure).



Exercise 11.5.4

Describe the relationship between the lines with the following parametric equations:

$$x = 1 - 4t,\ y = 3 + t,\ z = 8 - 6t$$

$$x = 2 + 3s,\ y = 2s,\ z = -1 - 3s.$$

Hint

Start by identifying direction vectors for each line. Is one a multiple of the other?

Answer

These lines are skew because their direction vectors are not parallel and there is no point (x, y, z) that lies on both lines.

Equations for a Plane

We know that a line is determined by two points. In other words, for any two distinct points, there is exactly one line that passes through those points, whether in two dimensions or three. Similarly, given any three points that do not all lie on the same line, there is a unique plane that passes through these points. Just as a line is determined by two points, a plane is determined by three.

This may be the simplest way to characterize a plane, but we can use other descriptions as well. For example, given two distinct, intersecting lines, there is exactly one plane containing both lines. A plane is also determined by a line and any point that does not lie on the line. These characterizations arise naturally from the idea that a plane is determined by three points. Perhaps the most surprising characterization of a plane is actually the most useful.

Imagine a pair of orthogonal vectors that share an initial point. Visualize grabbing one of the vectors and twisting it. As you twist, the other vector spins around and sweeps out a plane. Here, we describe that concept mathematically. Let $\vec{\mathbf{n}} = \langle a, b, c \rangle$ be a vector and $P = (x_0, y_0, z_0)$ be a point. Then the set of all points $Q = (x, y, z)$ such that $\overrightarrow{PQ}$ is orthogonal to $\vec{\mathbf{n}}$ forms a plane (Figure 11.5.7). We say that $\vec{\mathbf{n}}$ is a **normal vector**, or perpendicular to the plane. Remember, the dot product of orthogonal vectors is zero. This fact generates the **vector equation of a plane**:

$$\vec{\mathbf{n}} \cdot \overrightarrow{PQ} = 0.$$

Rewriting this equation provides additional ways to describe the plane:

$$\begin{aligned} \vec{\mathbf{n}} \cdot \overrightarrow{PQ} &= 0 \\ \langle a, b, c \rangle \cdot \langle x - x_0, y - y_0, z - z_0 \rangle &= 0 \\ a(x - x_0) + b(y - y_0) + c(z - z_0) &= 0. \end{aligned}$$

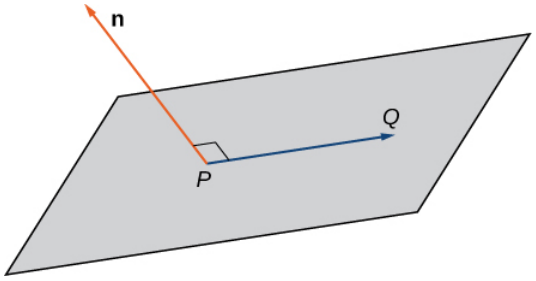


Figure 11.5.7: Given a point P and vector $\vec{\mathbf{n}}$, the set of all points Q with $\overrightarrow{PQ}$ orthogonal to $\vec{\mathbf{n}}$ forms a plane.

Definition: Scalar Equation of a Plane

Given a point P and vector $\vec{\mathbf{n}}$, the set of all points Q satisfying the equation $\vec{\mathbf{n}} \cdot \overrightarrow{PQ} = 0$ forms a plane. The equation

$$\vec{\mathbf{n}} \cdot \overrightarrow{PQ} = 0$$

is known as the *vector equation of a plane*.

The *scalar equation of a plane* (sometimes also called the *standard equation of a plane*) containing point $P = (x_0, y_0, z_0)$ with normal vector $\vec{n} = \langle a, b, c \rangle$ is

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0.$$

This equation can be expressed as $ax + by + cz + d = 0$, where $d = -ax_0 - by_0 - cz_0$. This form of the equation is sometimes called the *general form of the equation of a plane.*

As described earlier in this section, any three points that do not all lie on the same line determine a plane. Given three such points, we can find an equation for the plane containing these points.

Example 11.5.5: Writing an Equation of a Plane Given Three Points in the Plane

Write an equation for the plane containing points $P=(1,1,-2), Q=(0,2,1)$, and $R=(-1,-1,0)$ in both standard and general forms.

Solution

To write an equation for a plane, we must find a normal vector for the plane. We start by identifying two vectors in the plane:

$$\begin{aligned}\overrightarrow{PQ} &= \langle 0-1, 2-1, 1-(-2)\rangle \\ &= \langle -1, 1, 3\rangle \\ \overrightarrow{QR} &= \langle -1-0, -1-2, 0-1\rangle \\ &= \langle -1, -3, -1\rangle.\end{aligned}$$

The cross product $\overrightarrow{PQ}\times\overrightarrow{QR}$ is orthogonal to both $\overrightarrow{PQ}$ and $\overrightarrow{QR}$, so it is normal to the plane that contains these two vectors:

$$\begin{aligned}\vec{\mathbf{n}} &= \overrightarrow{PQ}\times\overrightarrow{QR} \\ &= \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ -1 & 1 & 3 \\ -1 & -3 & -1 \end{vmatrix} \\ &= (-1+9)\hat{\mathbf{i}} - (1+3)\hat{\mathbf{j}} + (3+1)\hat{\mathbf{k}} \\ &= 8\hat{\mathbf{i}} - 4\hat{\mathbf{j}} + 4\hat{\mathbf{k}}.\end{aligned}$$

Thus, $\vec{\mathbf{n}} = \langle 8, -4, 4\rangle$, and we can choose any of the three given points to write an equation of the plane:

$$\begin{aligned}8(x-1) - 4(y-1) + 4(z+2) &= 0 \\ 8x - 4y + 4z + 4 &= 0.\end{aligned}$$

The scalar equations of a plane vary depending on the normal vector and point chosen.

Example 11.5.6: Writing an Equation for a Plane Given a Point and a Line

Find an equation of the plane that passes through point $(1,4,3)$ and contains the line given by $x=\dfrac{y-1}{2}=z+1$.

Solution

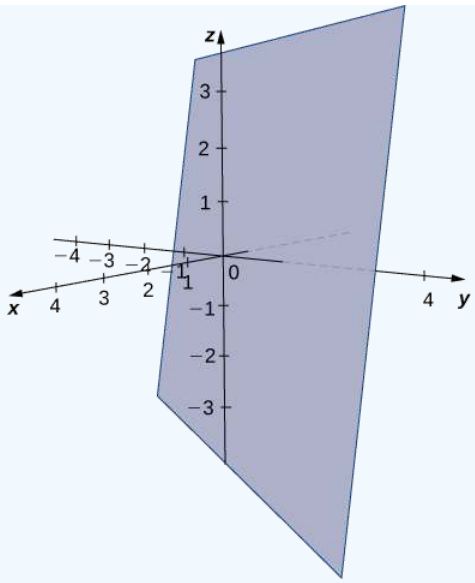
Symmetric equations describe the line that passes through point $(0,1,-1)$ parallel to vector $\vec{\mathbf{v}}_1=\langle 1,2,1\rangle$ (see the following figure). Use this point and the given point, $(1,4,3)$, to identify a second vector parallel to the plane:

$$\vec{\mathbf{v}}_2 = \langle 1-0, 4-1, 3-(-1)\rangle = \langle 1,3,4\rangle.$$

Use the cross product of these vectors to identify a normal vector for the plane:

$$\begin{aligned}\vec{\mathbf{n}} &= \vec{\mathbf{v}}_1\times\vec{\mathbf{v}}_2 \\ &= \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ 1 & 2 & 1 \\ 1 & 3 & 4 \end{vmatrix} \\ &= (8-3)\hat{\mathbf{i}} - (4-1)\hat{\mathbf{j}} + (3-2)\hat{\mathbf{k}} \\ &= 5\hat{\mathbf{i}} - 3\hat{\mathbf{j}} + \hat{\mathbf{k}}.\end{aligned}$$

The scalar equations for the plane are $5x-3(y-1)+(z+1)=0$ and $5x-3y+z+4=0$.



Exercise 11.5.6

Find an equation of the plane containing the lines L_1 and L_2:

$$L_1 : \ x = -y = z$$

$$L_2 : \ \frac{x-3}{2} = y = z-2.$$

Hint

Hint: The cross product of the lines' direction vectors gives a normal vector for the plane.

Answer

$$-2(x-1)+(y+1)+3(z-1)=0$$

or

$$-2x+y+3z=0$$

Now that we can write an equation for a plane, we can use the equation to find the distance d between a point P and the plane. It is defined as the shortest possible distance from P to a point on the plane.

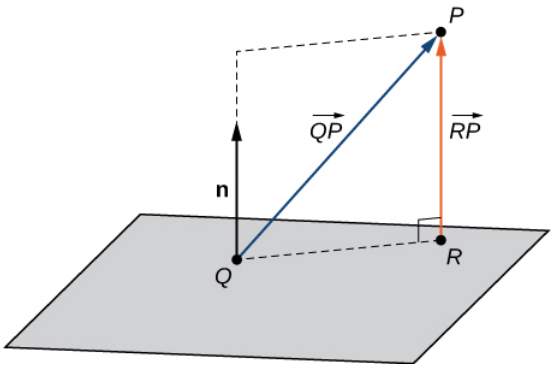


Figure 11.5.8: We want to find the shortest distance from point P to the plane. Let point R be the point in the plane such that, for any other point in the plane Q, $\|\overrightarrow{RP}\| < \|\overrightarrow{QP}\|$.

Just as we find the two-dimensional distance between a point and a line by calculating the length of a line segment perpendicular to the line, we find the three-dimensional distance between a point and a plane by calculating the length of a line segment

perpendicular to the plane. Let R be the point in the plane such that $\overrightarrow{RP}$ is orthogonal to the plane, and let Q be an arbitrary point in the plane. Then the projection of vector $\overrightarrow{QP}$ onto the normal vector describes vector $\overrightarrow{RP}$, as shown in Figure 11.5.8.

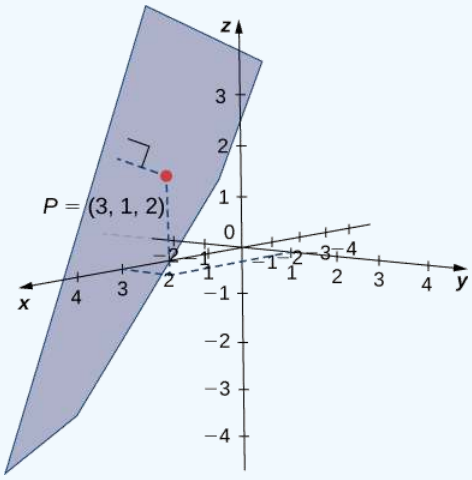
The Distance between a Plane and a Point

Suppose a plane with normal vector $\vec{\mathbf{n}}$ passes through point Q. The distance d from the plane to a point P not in the plane is given by

$$d = \|\text{proj}_{\vec{\mathbf{n}}}\, \overrightarrow{QP}\| = |\, \text{comp}_{\vec{\mathbf{n}}}\, \overrightarrow{QP}\,| = \frac{\left|\overrightarrow{QP} \cdot \vec{\mathbf{n}}\right|}{\|\vec{\mathbf{n}}\|}. \tag{11.5.5}$$

Example 11.5.7: Distance between a Point and a Plane

Find the distance between point $P = (3, 1, 2)$ and the plane given by $x - 2y + z = 5$ (see the following figure).



Solution

The coefficients of the plane's equation provide a normal vector for the plane: $\vec{\mathbf{n}} = \langle 1, -2, 1\rangle$. To find vector $\overrightarrow{QP}$, we need a point in the plane. Any point will work, so set $y = z = 0$ to see that point $Q = (5, 0, 0)$ lies in the plane. Find the component form of the vector from Q to P:

$$\overrightarrow{QP} = \langle 3 - 5, 1 - 0, 2 - 0\rangle = \langle -2, 1, 2\rangle.$$

Apply the distance formula from Equation 11.5.5:

$$\begin{aligned} d &= \frac{\left|\overrightarrow{QP} \cdot \vec{\mathbf{n}}\right|}{\|\vec{\mathbf{n}}\|} \\ &= \frac{|\langle -2, 1, 2\rangle \cdot \langle 1, -2, 1\rangle|}{\sqrt{1^2 + (-2)^2 + 1^2}} \\ &= \frac{|-2 - 2 + 2|}{\sqrt{6}} \\ &= \frac{2}{\sqrt{6}} = \frac{\sqrt{6}}{3} \text{ units.} \end{aligned}$$

Exercise 11.5.7

Find the distance between point $P = (5, -1, 0)$ and the plane given by $4x + 2y - z = 3$.

Hint

Point $(0, 0, -3)$ lies on the plane.

Answer

$$\frac{15}{\sqrt{21}} = \frac{5\sqrt{21}}{7} \text{ units}$$

Parallel and Intersecting Planes

We have discussed the various possible relationships between two lines in two dimensions and three dimensions. When we describe the relationship between two planes in space, we have only two possibilities: the two distinct planes are parallel or they intersect. When two planes are parallel, their normal vectors are parallel. When two planes intersect, the intersection is a line (Figure 11.5.9).

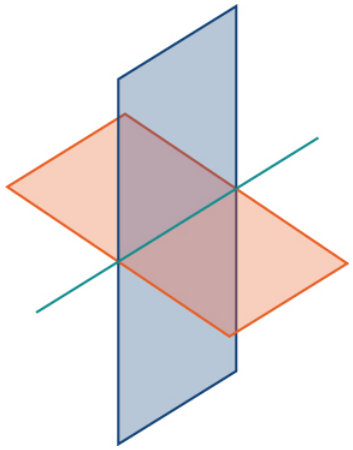
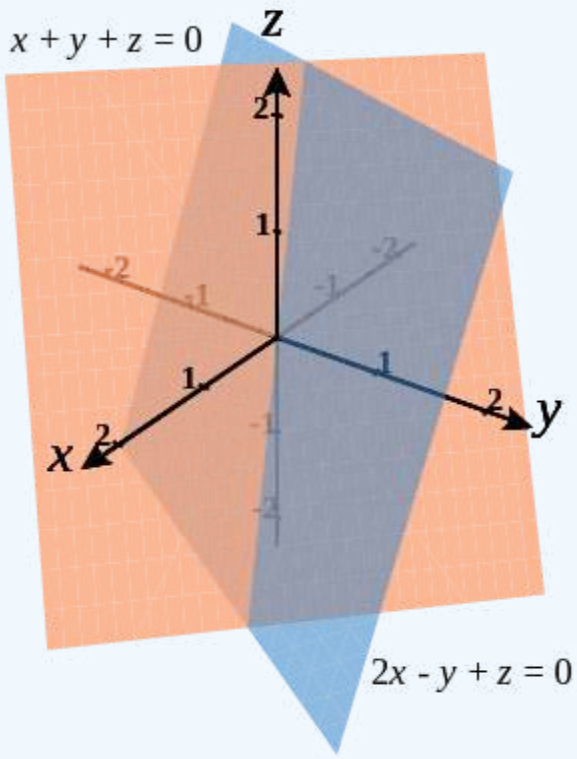


Figure 11.5.9: The intersection of two nonparallel planes is always a line.

We can use the equations of the two planes to find parametric equations for the line of intersection.

Example 11.5.8: Finding the Line of Intersection for Two Planes

Find parametric and symmetric equations for the line formed by the intersection of the planes given by $x + y + z = 0$ and $2x - y + z = 0$ (see the following figure).



Solution

Note that the two planes have nonparallel normals, so the planes intersect. Further, the origin satisfies each equation, so we know the line of intersection passes through the origin. Add the plane equations so we can eliminate one of the variables, in this case, y:

$$\begin{array}{r} x+y+z=0 \\ 2x-y+z=0 \\ \hline 3x+2z=0\,. \end{array}$$

This gives us $x=-\frac{2}{3}z$. We substitute this value into the first equation to express y in terms of z:

$$\begin{array}{r} x+y+z=0 \\ -\frac{2}{3}z+y+z=0 \\ y+\frac{1}{3}z=0 \\ y=-\frac{1}{3}z \end{array}.$$

We now have the first two variables, x and y, in terms of the third variable, z. Now we define z in terms of t. To eliminate the need for fractions, we choose to define the parameter t as $t=-\frac{1}{3}z$. Then, $z=-3t$. Substituting the parametric

representation of z back into the other two equations, we see that the parametric equations for the line of intersection are $x = 2t,\ y = t,\ z = -3t$. The symmetric equations for the line are $\dfrac{x}{2} = y = \dfrac{z}{-3}$.

? Exercise 11.5.8

Find parametric equations for the line formed by the intersection of planes $x + y - z = 3$ and $3x - y + 3z = 5$.

Hint

Add the two equations, then express z in terms of x. Then, express y in terms of x.

Answer

$x = t,\ y = 7 - 3t,\ z = 4 - 2t$

In addition to finding the equation of the line of intersection between two planes, we may need to find the angle formed by the intersection of two planes. For example, builders constructing a house need to know the angle where different sections of the roof meet to know whether the roof will look good and drain properly. We can use normal vectors to calculate the angle between the two planes. We can do this because the angle between the normal vectors is the same as the angle between the planes. Figure 11.5.10 shows why this is true.

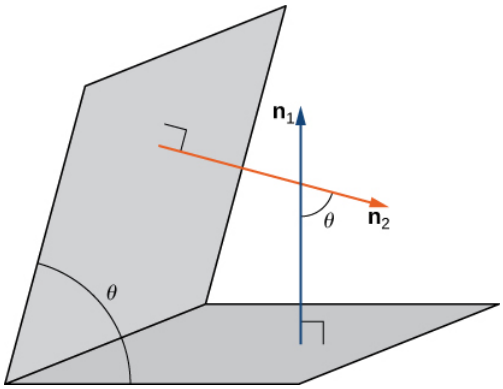


Figure 11.5.10: The angle between two planes has the same measure as the angle between the normal vectors for the planes.

We can find the measure of the angle θ between two intersecting planes by first finding the cosine of the angle, using the following equation:

$$\cos\theta = \frac{|\vec{\mathbf{n}}_1 \cdot \vec{\mathbf{n}}_2|}{\|\vec{\mathbf{n}}_1\|\|\vec{\mathbf{n}}_2\|}.$$

We can then use the angle to determine whether two planes are parallel or orthogonal or if they intersect at some other angle.

✓ Example 11.5.9: Finding the Angle between Two Planes

Determine whether each pair of planes is parallel, orthogonal, or neither. If the planes are intersecting, but not orthogonal, find the measure of the angle between them. Give the answer in radians and round to two decimal places.

a. $x + 2y - z = 8$ and $2x + 4y - 2z = 10$
b. $2x - 3y + 2z = 3$ and $6x + 2y - 3z = 1$
c. $x + y + z = 4$ and $x - 3y + 5z = 1$

Solution:

a. The normal vectors for these planes are $\vec{\mathbf{n}}_1 = \langle 1, 2, -1\rangle$ and $\vec{\mathbf{n}}_2 = \langle 2, 4, -2\rangle$. These two vectors are scalar multiples of each other. The normal vectors are parallel, so the planes are parallel.
b. The normal vectors for these planes are $\vec{\mathbf{n}}_1 = \langle 2, -3, 2\rangle$ and $\vec{\mathbf{n}}_2 = \langle 6, 2, -3\rangle$. Taking the dot product of these vectors, we have

$$\begin{aligned}\vec{\mathbf{n}}_1 \cdot \vec{\mathbf{n}}_2 &= \langle 2,-3,2\rangle \cdot \langle 6,2,-3\rangle \\ &= 2(6)-3(2)+2(-3)=0.\end{aligned}$$

The normal vectors are orthogonal, so the corresponding planes are orthogonal as well.

c. The normal vectors for these planes are $\vec{\mathbf{n}}_1 = \langle 1,1,1\rangle$ and $\vec{\mathbf{n}}_2 = \langle 1,-3,5\rangle$:

$$\begin{aligned}\cos\theta &= \frac{|\vec{\mathbf{n}}_1 \cdot \vec{\mathbf{n}}_2|}{\|\vec{\mathbf{n}}_1\|\|\vec{\mathbf{n}}_2\|} \\ &= \frac{|\langle 1,1,1\rangle \cdot \langle 1,-3,5\rangle|}{\sqrt{1^2+1^2+1^2}\sqrt{1^2+(-3)^2+5^2}} \\ &= \frac{3}{\sqrt{105}}\end{aligned}$$

Then $\theta = \arccos\frac{3}{\sqrt{105}} \approx 1.27$ rad.

Thus the angle between the two planes is about 1.27 rad, or approximately 73°.

Exercise 11.5.9

Find the measure of the angle between planes $x+y-z=3$ and $3x-y+3z=5$. Give the answer in radians and round to two decimal places.

Hint

Use the coefficients of the variables in each equation to find a normal vector for each plane.

Answer

1.44 rad

When we find that two planes are parallel, we may need to find the distance between them. To find this distance, we simply select a point in one of the planes. The distance from this point to the other plane is the distance between the planes.

Previously, we introduced the formula for calculating this distance in Equation 11.5.5:

$$d = \frac{\overrightarrow{QP}\cdot\vec{\mathbf{n}}}{\|\vec{\mathbf{n}}\|},$$

where Q is a point on the plane, P is a point not on the plane, and $\vec{n}$ is the normal vector that passes through point Q. Consider the distance from point (x_0,y_0,z_0) to plane $ax+by+cz+k=0$. Let (x_1,y_1,z_1) be any point in the plane. Substituting into the formula yields

$$\begin{aligned}d &= \frac{|a(x_0-x_1)+b(y_0-y_1)+c(z_0-z_1)|}{\sqrt{a^2+b^2+c^2}} \\ &= \frac{|ax_0+by_0+cz_0+k|}{\sqrt{a^2+b^2+c^2}}.\end{aligned}$$

We state this result formally in the following theorem.

Distance from a Point to a Plane

Let $P(x_0,y_0,z_0)$ be a point. The distance from P to plane $ax+by+cz+k=0$ is given by

$$d = \frac{|ax_0+by_0+cz_0+k|}{\sqrt{a^2+b^2+c^2}}.$$

Example 11.5.10: Finding the Distance between Parallel Planes

Find the distance between the two parallel planes given by $2x+y-z=2$ and $2x+y-z=8$.

Solution

Point $(1,0,0)$ lies in the first plane. The desired distance, then, is

$$\begin{aligned} d &= \frac{|ax_0+by_0+cz_0+k|}{\sqrt{a^2+b^2+c^2}} \\ &= \frac{|2(1)+1(0)+(-1)(0)+(-8)|}{\sqrt{2^2+1^2+(-1)^2}} \\ &= \frac{6}{\sqrt{6}} = \sqrt{6}\,\text{units} \end{aligned}$$

Exercise 11.5.10:

Find the distance between parallel planes $5x-2y+z=6$ and $5x-2y+z=-3$.

Hint

Set $x=y=0$ to find a point on the first plane.

Answer

$$\frac{9}{\sqrt{30}} = \frac{3\sqrt{30}}{10}\,\text{units}$$

Distance between Two Skew Lines

Finding the distance from a point to a line or from a line to a plane seems like a pretty abstract procedure. But, if the lines represent pipes in a chemical plant or tubes in an oil refinery or roads at an intersection of highways, confirming that the distance between them meets specifications can be both important and awkward to measure. One way is to model the two pipes as lines, using the techniques in this chapter, and then calculate the distance between them. The calculation involves forming vectors along the directions of the lines and using both the cross product and the dot product.

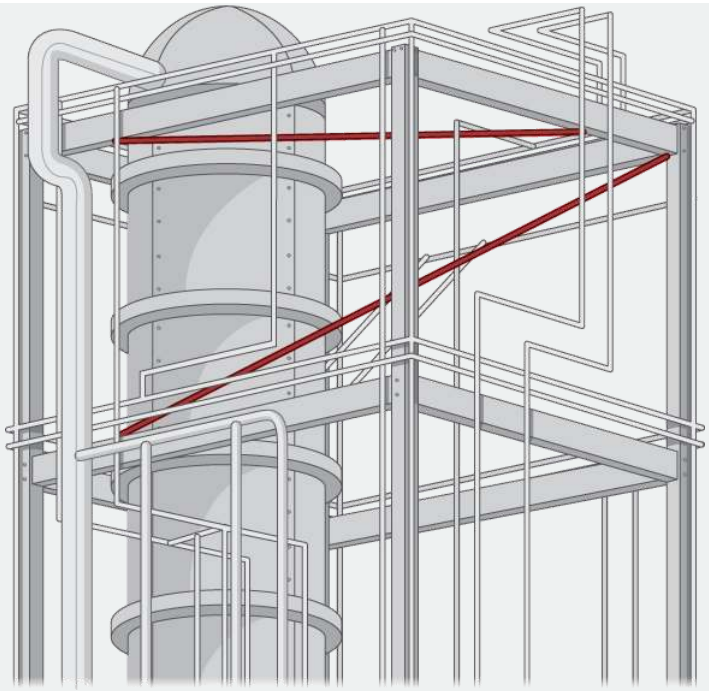


Figure 11.5.11: Industrial pipe installations often feature pipes running in different directions. How can we find the distance between two skew pipes?

The symmetric forms of two lines, L_1 and L_2, are

$$L_1 : \frac{x-x_1}{a_1} = \frac{y-y_1}{b_1} = \frac{z-z_1}{c_1}$$

$$L_2 : \frac{x-x_2}{a_2} = \frac{y-y_2}{b_2} = \frac{z-z_2}{c_2}.$$

You are to develop a formula for the distance d between these two lines, in terms of the values $a_1, b_1, c_1; a_2, b_2, c_2; x_1, y_1, z_1;$ and x_2, y_2, z_2. The distance between two lines is usually taken to mean the minimum distance, so this is the length of a line segment or the length of a vector that is perpendicular to both lines and intersects both lines.

1. First, write down two vectors, $\vec{\mathbf{v}}_1$ and $\vec{\mathbf{v}}_2$, that lie along L_1 and L_2, respectively.

2. Find the cross product of these two vectors and call it $\vec{\mathbf{N}}$. This vector is perpendicular to $\vec{\mathbf{v}}_1$ and $\vec{\mathbf{v}}_2$, and hence is perpendicular to both lines.

3. From vector $\vec{\mathbf{N}}$, form a unit vector $\vec{\mathbf{n}}$ in the same direction.

4. Use symmetric equations to find a convenient vector $\vec{\mathbf{v}}_{12}$ that lies between any two points, one on each line. Again, this can be done directly from the symmetric equations.

5. The dot product of two vectors is the magnitude of the projection of one vector onto the other—that is, $\vec{\mathbf{A}} \cdot \vec{\mathbf{B}} = \|\vec{\mathbf{A}}\| \|\vec{\mathbf{B}}\| \cos\theta$, where θ is the angle between the vectors. Using the dot product, find the projection of vector $\vec{\mathbf{v}}_{12}$ found in step 4 onto unit vector $\vec{\mathbf{n}}$ found in step 3. This projection is perpendicular to both lines, and hence its length must be the perpendicular distance d between them. Note that the value of d may be negative, depending on your choice of vector $\vec{\mathbf{v}}_{12}$ or the order of the cross product, so use absolute value signs around the numerator.

6. Check that your formula gives the correct distance of $|-25|/\sqrt{198} \approx 1.78$ between the following two lines:

$$L_1 : \frac{x-5}{2} = \frac{y-3}{4} = \frac{z-1}{3}$$

$$L_2 : \frac{x-6}{3} = \frac{y-1}{5} = \frac{z}{7}.$$

7. Is your general expression valid when the lines are parallel? If not, why not? (Hint: What do you know about the value of the cross product of two parallel vectors? Where would that result show up in your expression for d?)

8. Demonstrate that your expression for the distance is zero when the lines intersect. Recall that two lines intersect if they are not parallel and they are in the same plane. Hence, consider the direction of $\vec{\mathbf{n}}$ and $\vec{\mathbf{v}}_{12}$. What is the result of their dot product?

9. Consider the following application. Engineers at a refinery have determined they need to install support struts between many of the gas pipes to reduce damaging vibrations. To minimize cost, they plan to install these struts at the closest points between adjacent skewed pipes. Because they have detailed schematics of the structure, they are able to determine the correct lengths of the struts needed, and hence manufacture and distribute them to the installation crews without spending valuable time making measurements.

The rectangular frame structure has the dimensions $4.0 \times 15.0 \times 10.0\,\text{m}$ (height, width, and depth). One sector has a pipe entering the lower corner of the standard frame unit and exiting at the diametrically opposed corner (the one farthest away at the top); call this L_1. A second pipe enters and exits at the two different opposite lower corners; call this L_2 (Figure 11.5.12).

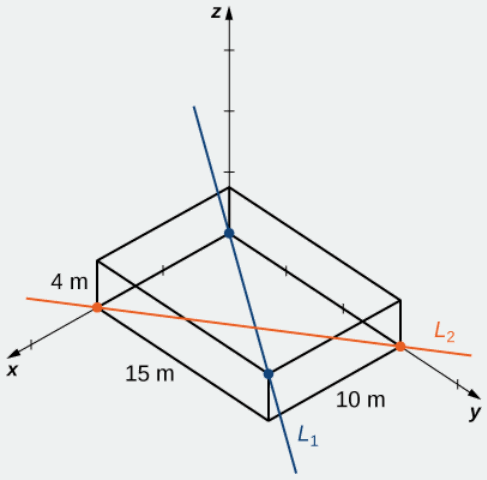


Figure 11.5.12: Two pipes cross through a standard frame unit.

Write down the vectors along the lines representing those pipes, find the cross product between them from which to create the unit vector $\vec{\mathbf{n}}$, define a vector that spans two points on each line, and finally determine the minimum distance between the lines. (Take the origin to be at the lower corner of the first pipe.) Similarly, you may also develop the symmetric equations for each line and substitute directly into your formula.

Key Concepts

- In three dimensions, the direction of a line is described by a direction vector. The vector equation of a line with direction vector $\vec{\mathbf{v}} = \langle a, b, c \rangle$ passing through point $P = (x_0, y_0, z_0)$ is $\vec{\mathbf{r}} = \vec{\mathbf{r}}_0 + t\vec{\mathbf{v}}$, where $\vec{\mathbf{r}}_0 = \langle x_0, y_0, z_0 \rangle$ is the position vector of point P. This equation can be rewritten to form the parametric equations of the line: $x = x_0 + ta, y = y_0 + tb$, and $z = z_0 + tc$. The line can also be described with the symmetric equations $\dfrac{x - x_0}{a} = \dfrac{y - y_0}{b} = \dfrac{z - z_0}{c}$.
- Let L be a line in space passing through point P with direction vector $\vec{\mathbf{v}}$. If Q is any point not on L, then the distance from Q to L is $d = \dfrac{\|\overrightarrow{PQ} \times \vec{\mathbf{v}}\|}{\|\vec{\mathbf{v}}\|}$.
- In three dimensions, two lines may be parallel but not equal, equal, intersecting, or skew.
- Given a point P and vector $\vec{\mathbf{n}}$, the set of all points Q satisfying equation $\vec{\mathbf{n}} \cdot \overrightarrow{PQ} = 0$ forms a plane. Equation $\vec{\mathbf{n}} \cdot \overrightarrow{PQ} = 0$ is known as the *vector equation of a plane.*
- The scalar equation of a plane containing point $P = (x_0, y_0, z_0)$ with normal vector $\vec{\mathbf{n}} = \langle a, b, c \rangle$ is $a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$. This equation can be expressed as $ax + by + cz + d = 0$, where $d = -ax_0 - by_0 - cz_0$. This form of the equation is sometimes called the general *form of the equation of a plane.*

- Suppose a plane with normal vector n passes through point Q. The distance D from the plane to point P not in the plane is given by

$$D = \|\text{proj}_{\vec{\mathbf{n}}}\overrightarrow{QP}\| = |\text{ comp}_{\vec{\mathbf{n}}}\overrightarrow{QP}| = \frac{\left|\overrightarrow{QP}\cdot\vec{\mathbf{n}}\right|}{\|\vec{\mathbf{n}}\|.}$$

- The normal vectors of parallel planes are parallel. When two planes intersect, they form a line.
- The measure of the angle θ between two intersecting planes can be found using the equation: $\cos\theta = \frac{|\vec{\mathbf{n}}_1 \cdot \vec{\mathbf{n}}_2|}{\|\vec{\mathbf{n}}_1\|\|\vec{\mathbf{n}}_2\|}$, where $\vec{\mathbf{n}}_1$ and $\vec{\mathbf{n}}_2$ are normal vectors to the planes.
- The distance D from point (x_0, y_0, z_0) to plane $ax + by + cz + d = 0$ is given by

$$D = \frac{|a(x_0 - x_1) + b(y_0 - y_1) + c(z_0 - z_1)|}{\sqrt{a^2 + b^2 + c^2}} = \frac{|ax_0 + by_0 + cz_0 + d|}{\sqrt{a^2 + b^2 + c^2}}$$

.

Key Equations

- **Vector Equation of a Line**

  $\vec{\mathbf{r}} = \vec{\mathbf{r}}_0 + t\vec{\mathbf{v}}$

- **Parametric Equations of a Line**

  $x = x_0 + ta,\ y = y_0 + tb,$ and $z = z_0 + tc$

- **Symmetric Equations of a Line**

  $\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c}$

- **Vector Equation of a Plane**

  $\vec{\mathbf{n}} \cdot \overrightarrow{PQ} = 0$

- **Scalar Equation of a Plane**

  $a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$

- **Distance between a Plane and a Point**

  $d = \|\text{proj}_{\vec{\mathbf{n}}}\overrightarrow{QP}\| = |\text{ comp}_{\vec{\mathbf{n}}}\overrightarrow{QP}| = \frac{\left|\overrightarrow{QP}\cdot\vec{\mathbf{n}}\right|}{\|\vec{\mathbf{n}}\|}$

Glossary

direction vector

a vector parallel to a line that is used to describe the direction, or orientation, of the line in space

general form of the equation of a plane

an equation in the form $ax + by + cz + d = 0,$ where $\vec{\mathbf{n}} = \langle a, b, c\rangle$ is a normal vector of the plane, $P = (x_0, y_0, z_0)$ is a point on the plane, and $d = -ax_0 - by_0 - cz_0$

normal vector

a vector perpendicular to a plane

parametric equations of a line

the set of equations $x = x_0 + ta, y = y_0 + tb,$ and $z = z_0 + tc$ describing the line with direction vector $v = \langle a, b, c\rangle$ passing through point (x_0, y_0, z_0)

scalar equation of a plane

the equation $a(x-x_0)+b(y-y_0)+c(z-z_0)=0$ used to describe a plane containing point $P=(x_0,y_0,z_0)$ with normal vector $n=\langle a,b,c\rangle$ or its alternate form $ax+by+cz+d=0$, where $d=-ax_0-by_0-cz_0$

skew lines

two lines that are not parallel but do not intersect

symmetric equations of a line

the equations $\frac{x-x_0}{a}=\frac{y-y_0}{b}=\frac{z-z_0}{c}$ describing the line with direction vector $v=\langle a,b,c\rangle$ passing through point (x_0,y_0,z_0)

vector equation of a line

the equation $\vec{\mathbf{r}}=\vec{\mathbf{r}}_0+t\vec{\mathbf{v}}$ used to describe a line with direction vector $\vec{\mathbf{v}}=\langle a,b,c\rangle$ passing through point $P=(x_0,y_0,z_0)$, where $\vec{\mathbf{r}}_0=\langle x_0,y_0,z_0\rangle$, is the position vector of point P

vector equation of a plane

the equation $\vec{\mathbf{n}}\cdot\overrightarrow{PQ}=0,$ where P is a given point in the plane, Q is any point in the plane, and $\vec{\mathbf{n}}$ is a normal vector of the plane

---

11.5: Lines and Planes in Space is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

11.6: Cylinders and Quadric Surfaces

Learning Objectives

- Convert from cylindrical to rectangular coordinates.
- Convert from rectangular to cylindrical coordinates.
- Convert from spherical to rectangular coordinates.
- Convert from rectangular to spherical coordinates.

The Cartesian coordinate system provides a straightforward way to describe the location of points in space. Some surfaces, however, can be difficult to model with equations based on the Cartesian system. This is a familiar problem; recall that in two dimensions, polar coordinates often provide a useful alternative system for describing the location of a point in the plane, particularly in cases involving circles. In this section, we look at two different ways of describing the location of points in space, both of them based on extensions of polar coordinates. As the name suggests, cylindrical coordinates are useful for dealing with problems involving cylinders, such as calculating the volume of a round water tank or the amount of oil flowing through a pipe. Similarly, spherical coordinates are useful for dealing with problems involving spheres, such as finding the volume of domed structures.

Cylindrical Coordinates

When we expanded the traditional Cartesian coordinate system from two dimensions to three, we simply added a new axis to model the third dimension. Starting with polar coordinates, we can follow this same process to create a new three-dimensional coordinate system, called the cylindrical coordinate system. In this way, cylindrical coordinates provide a natural extension of polar coordinates to three dimensions.

Definition: The Cylindrical Coordinate System

In the cylindrical coordinate system**,** a point in space (Figure $11.6.1$) is represented by the ordered triple (r, θ, z), where

- (r, θ) are the polar coordinates of the point's projection in the xy-plane
- z is the usual z-**coordinate** in the Cartesian coordinate system

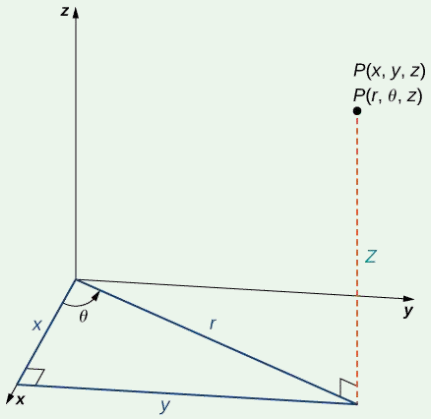


Figure 11.6.1: The right triangle lies in the xy-plane. The length of the hypotenuse is r and θ is the measure of the angle formed by the positive x-axis and the hypotenuse. The z-coordinate describes the location of the point above or below the xy-plane.

In the xy-plane, the right triangle shown in Figure $11.6.1$ provides the key to transformation between cylindrical and Cartesian, or rectangular, coordinates.

Conversion between Cylindrical and Cartesian Coordinates

The rectangular coordinates (x, y, z) and the cylindrical coordinates (r, θ, z) of a point are related as follows:

These equations are used to convert from cylindrical coordinates to rectangular coordinates.

- $x = r\cos\theta$
- $y = r\sin\theta$
- $z = z$

These equations are used to convert from rectangular coordinates to cylindrical coordinates

1. $r^2 = x^2 + y^2$
2. $\tan\theta = \dfrac{y}{x}$
3. $z = z$

As when we discussed conversion from rectangular coordinates to polar coordinates in two dimensions, it should be noted that the equation $\tan\theta = \dfrac{y}{x}$ has an infinite number of solutions. However, if we restrict θ to values between 0 and 2π, then we can find a unique solution based on the quadrant of the xy-plane in which original point (x, y, z) is located. Note that if $x = 0$, then the value of θ is either $\dfrac{\pi}{2}, \dfrac{3\pi}{2}$, or 0, depending on the value of y.

Notice that these equations are derived from properties of right triangles. To make this easy to see, consider point P in the xy-plane with rectangular coordinates $(x, y, 0)$ and with cylindrical coordinates $(r, \theta, 0)$, as shown in Figure 11.6.2

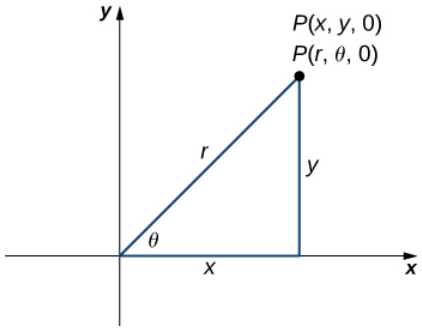


Figure 11.6.2: The Pythagorean theorem provides equation $r^2 = x^2 + y^2$. Right-triangle relationships tell us that $x = r\cos\theta, y = r\sin\theta,$ and $\tan\theta = y/x$.

Let's consider the differences between rectangular and cylindrical coordinates by looking at the surfaces generated when each of the coordinates is held constant. If c is a constant, then in rectangular coordinates, surfaces of the form $x = c, y = c$, or $z = c$ are all planes. Planes of these forms are parallel to the yz-plane, the xz-plane, and the xy-plane, respectively. When we convert to cylindrical coordinates, the z-coordinate does not change. Therefore, in cylindrical coordinates, surfaces of the form $z = c$ are planes parallel to the xy-plane. Now, let's think about surfaces of the form $r = c$. The points on these surfaces are at a fixed distance from the z-axis. In other words, these surfaces are vertical circular cylinders. Last, what about $\theta = c$? The points on a surface of the form $\theta = c$ are at a fixed angle from the x-axis, which gives us a half-plane that starts at the z-axis (Figures 11.6.3 and 11.6.4).

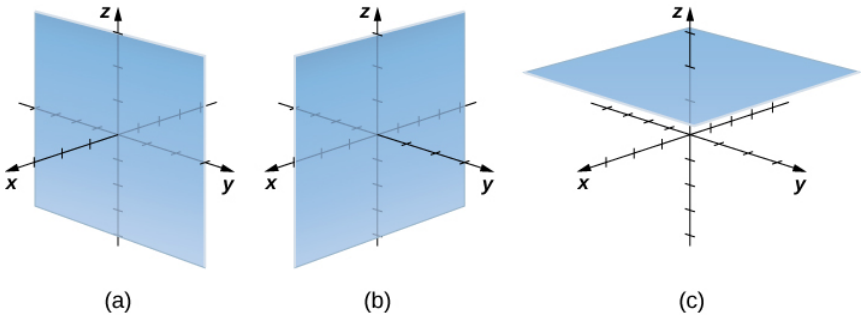


Figure 11.6.3: In rectangular coordinates, (a) surfaces of the form $x = c$ are planes parallel to the yz-plane, (b) surfaces of the form $y = c$ are planes parallel to the xz-plane, and (c) surfaces of the form $z = c$ are planes parallel to the xy-plane.

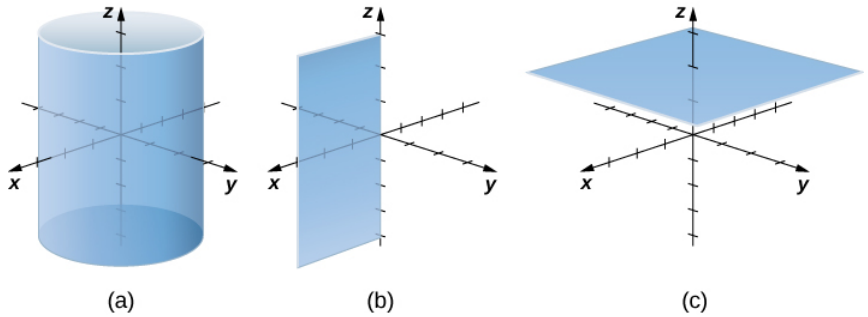


Figure 11.6.4: In cylindrical coordinates, (a) surfaces of the form $r = c$ are vertical cylinders of radius r, (b) surfaces of the form $\theta = c$ are half-planes at angle θ from the x-axis, and (c) surfaces of the form $z = c$ are planes parallel to the xy-plane.

Example 11.6.1: Converting from Cylindrical to Rectangular Coordinates

Plot the point with cylindrical coordinates $(4, \frac{2\pi}{3}, -2)$ and express its location in rectangular coordinates.

Solution

Conversion from cylindrical to rectangular coordinates requires a simple application of the equations listed in Conversion between Cylindrical and Cartesian Coordinates:

$$\begin{aligned} x &= r\cos\theta = 4\cos\frac{2\pi}{3} = -2 \\ y &= r\sin\theta = 4\sin\frac{2\pi}{3} = 2\sqrt{3} \\ z &= -2 \end{aligned}.$$

The point with cylindrical coordinates $(4, \frac{2\pi}{3}, -2)$ has rectangular coordinates $(-2, 2\sqrt{3}, -2)$ (Figure 11.6.5).

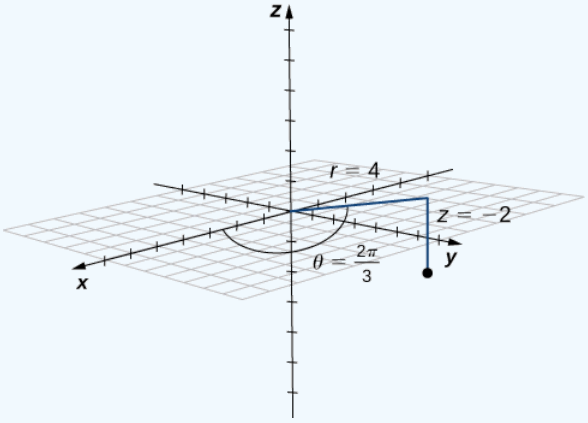


Figure 11.6.5: The projection of the point in the xy-plane is 4 units from the origin. The line from the origin to the point's projection forms an angle of $\frac{2\pi}{3}$ with the positive x-axis. The point lies 2 units below the xy-plane.

Exercise 11.6.1

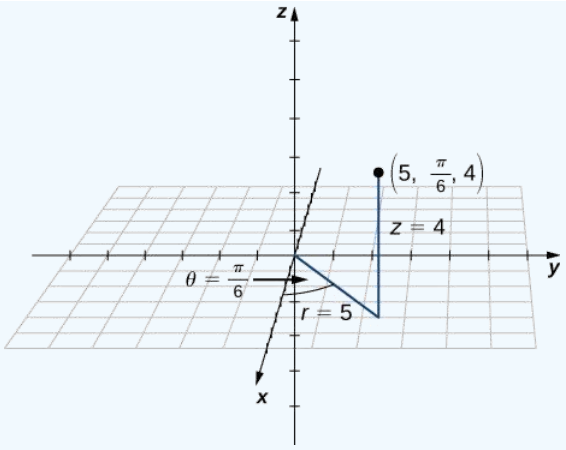
Point R has cylindrical coordinates $(5, \frac{\pi}{6}, 4)$. Plot R and describe its location in space using rectangular, or Cartesian, coordinates.

Hint

The first two components match the polar coordinates of the point in the xy-plane.

Answer

The rectangular coordinates of the point are $(\frac{5\sqrt{3}}{2}, \frac{5}{2}, 4)$.



If this process seems familiar, it is with good reason. This is exactly the same process that we followed in Introduction to Parametric Equations and Polar Coordinates to convert from polar coordinates to two-dimensional rectangular coordinates.

Example 11.6.2: Converting from Rectangular to Cylindrical Coordinates

Convert the rectangular coordinates $(1, -3, 5)$ to cylindrical coordinates.

Solution

Use the second set of equations from Conversion between Cylindrical and Cartesian Coordinates to translate from rectangular to cylindrical coordinates:

$$\begin{aligned} r^2 &= x^2 + y^2 \\ r &= \pm\sqrt{1^2 + (-3)^2} \\ &= \pm\sqrt{10}. \end{aligned}$$

We choose the positive square root, so $r = \sqrt{10}$.Now, we apply the formula to find θ. In this case, y is negative and x is positive, which means we must select the value of θ between $\dfrac{3\pi}{2}$ and 2π:

$$\begin{aligned} \tan\theta &= \frac{y}{x} & &= \frac{-3}{1} \\ \theta &= \arctan(-3) & &\approx 5.03 \text{ rad}. \end{aligned}$$

In this case, the z-coordinates are the same in both rectangular and cylindrical coordinates:

$$z = 5.$$

The point with rectangular coordinates $(1, -3, 5)$ has cylindrical coordinates approximately equal to $(\sqrt{10}, 5.03, 5)$.

Exercise 11.6.2

Convert point $(-8, 8, -7)$ from Cartesian coordinates to cylindrical coordinates.

Hint

$r^2 = x^2 + y^2$ and $\tan\theta = \frac{y}{x}$

Answer

$(8\sqrt{2}, \frac{3\pi}{4}, -7)$

The use of cylindrical coordinates is common in fields such as physics. Physicists studying electrical charges and the capacitors used to store these charges have discovered that these systems sometimes have a cylindrical symmetry. These systems have complicated modeling equations in the Cartesian coordinate system, which make them difficult to describe and analyze. The equations can often be expressed in more simple terms using cylindrical coordinates. For example, the cylinder described by equation $x^2+y^2=25$ in the Cartesian system can be represented by cylindrical equation $r=5$.

Example 11.6.3: Identifying Surfaces in the Cylindrical Coordinate System

Describe the surfaces with the given cylindrical equations.

a. $\theta=\dfrac{\pi}{4}$
b. $r^2+z^2=9$
c. $z=r$

Solution

a. When the angle θ is held constant while r and z are allowed to vary, the result is a half-plane (Figure 11.6.6).

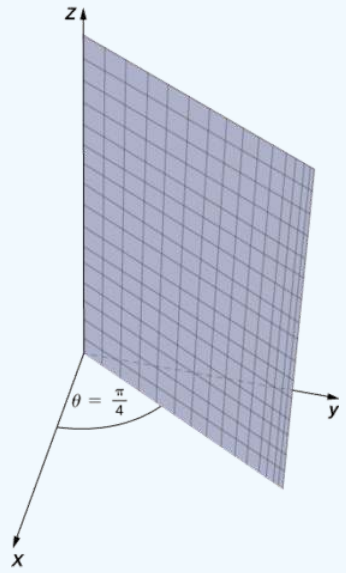


Figure 11.6.6: In polar coordinates, the equation $\theta=\pi/4$ describes the ray extending diagonally through the first quadrant. In three dimensions, this same equation describes a half-plane.

b. Substitute $r^2=x^2+y^2$ into equation $r^2+z^2=9$ to express the rectangular form of the equation: $x^2+y^2+z^2=9$. This equation describes a sphere centered at the origin with radius 3 (Figure 11.6.7).

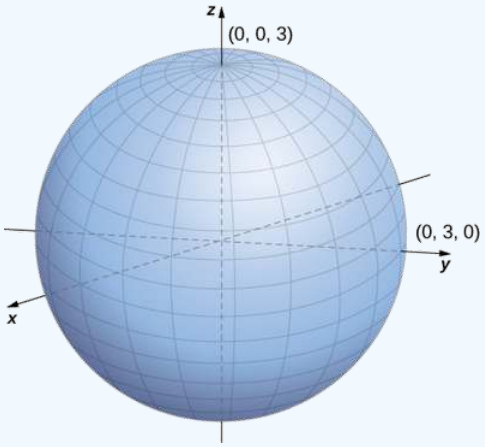


Figure 11.6.7: The sphere centered at the origin with radius 3 can be described by the cylindrical equation $r^2+z^2=9$.

c. To describe the surface defined by equation $z = r$, is it useful to examine traces parallel to the xy-plane. For example, the trace in plane $z = 1$ is circle $r = 1$, the trace in plane $z = 3$ is circle $r = 3$, and so on. Each trace is a circle. As the value of z increases, the radius of the circle also increases. The resulting surface is a cone (Figure 11.6.8).

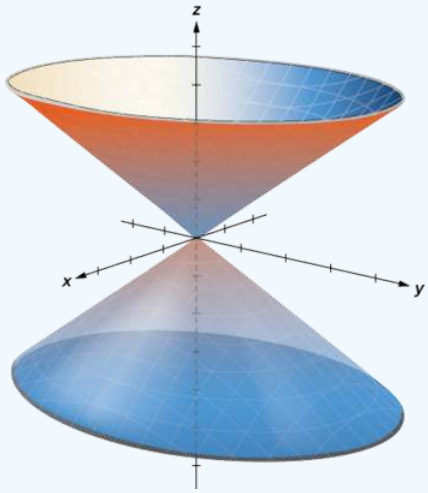


Figure 11.6.8: The traces in planes parallel to the xy-plane are circles. The radius of the circles increases as z increases.

? Exercise 11.6.3

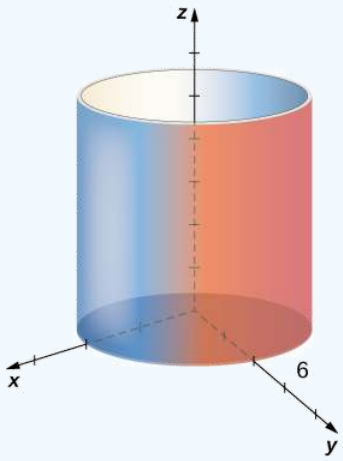
Describe the surface with cylindrical equation $r = 6$.

Hint

The θ and z components of points on the surface can take any value.

Answer

This surface is a cylinder with radius 6.



Spherical Coordinates

In the Cartesian coordinate system, the location of a point in space is described using an ordered triple in which each coordinate represents a distance. In the cylindrical coordinate system, the location of a point in space is described using two distances (r and z) and an angle measure (θ). In the spherical coordinate system, we again use an ordered triple to describe the location of a point in space. In this case, the triple describes one distance and two angles. Spherical coordinates make it simple to describe a sphere, just as cylindrical coordinates make it easy to describe a cylinder. Grid lines for spherical coordinates are based on angle measures, like those for polar coordinates.

Definition: spherical coordinate system

In the *spherical coordinate system*, a point P in space (Figure 11.6.9) is represented by the ordered triple (ρ, θ, φ) where

- ρ (the Greek letter rho) is the distance between P and the origin $(\rho \neq 0)$;
- θ is the same angle used to describe the location in cylindrical coordinates;
- φ (the Greek letter phi) is the angle formed by the positive z-axis and line segment $\overline{OP}$, where O is the origin and $0 \leq \varphi \leq \pi$.

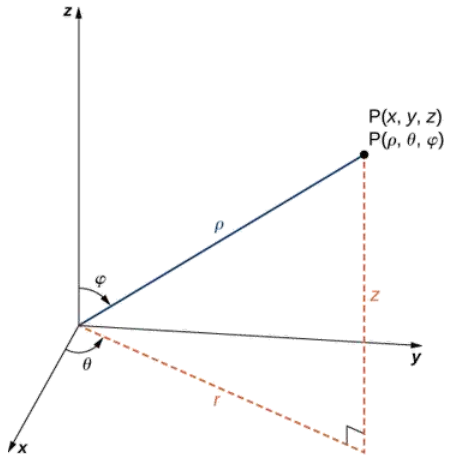


Figure 11.6.9: The relationship among spherical, rectangular, and cylindrical coordinates.

By convention, the origin is represented as $(0, 0, 0)$ in spherical coordinates.

HOW TO: Converting among Spherical, Cylindrical, and Rectangular Coordinates

Rectangular coordinates (x, y, z), cylindrical coordinates (r, θ, z), and spherical coordinates (ρ, θ, φ) of a point are related as follows:

Convert from spherical coordinates to rectangular coordinates

These equations are used to convert from spherical coordinates to rectangular coordinates.

- $x = \rho \sin\varphi \cos\theta$
- $y = \rho \sin\varphi \sin\theta$
- $z = \rho \cos\varphi$

Convert from rectangular coordinates to spherical coordinates

These equations are used to convert from rectangular coordinates to spherical coordinates.

- $\rho^2 = x^2 + y^2 + z^2$
- $\tan\theta = \dfrac{y}{x}$
- $\varphi = \arccos\left(\dfrac{z}{\sqrt{x^2 + y^2 + z^2}}\right)$.

Convert from spherical coordinates to cylindrical coordinates

These equations are used to convert from spherical coordinates to cylindrical coordinates.

- $r = \rho \sin\varphi$
- $\theta = \theta$
- $z = \rho \cos\varphi$

Convert from cylindrical coordinates to spherical coordinates

These equations are used to convert from cylindrical coordinates to spherical coordinates.

- $\rho = \sqrt{r^2 + z^2}$
- $\theta = \theta$
- $\varphi = \arccos\left(\dfrac{z}{\sqrt{r^2 + z^2}}\right)$

The formulas to convert from spherical coordinates to rectangular coordinates may seem complex, but they are straightforward applications of trigonometry. Looking at Figure 11.6.10, it is easy to see that $r = \rho \sin\varphi$. Then, looking at the triangle in the xy-plane with r as its hypotenuse, we have $x = r\cos\theta = \rho\sin\varphi\cos\theta$. The derivation of the formula for y is similar. Figure 11.6.10 also shows that $\rho^2 = r^2 + z^2 = x^2 + y^2 + z^2$ and $z = \rho\cos\varphi$. Solving this last equation for φ and then substituting $\rho = \sqrt{r^2 + z^2}$ (from the first equation) yields $\varphi = \arccos\left(\dfrac{z}{\sqrt{r^2 + z^2}}\right)$. Also, note that, as before, we must be careful when using the formula $\tan\theta = \dfrac{y}{x}$ to choose the correct value of θ.

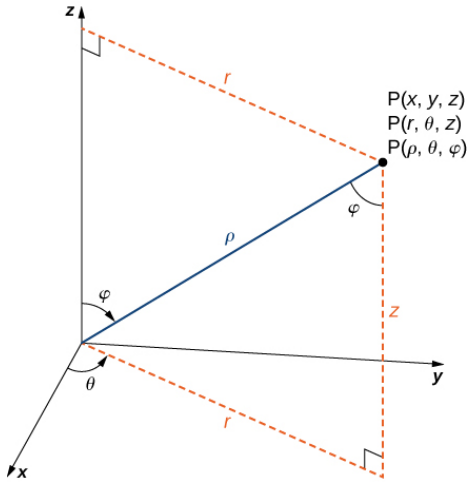


Figure 11.6.10: The equations that convert from one system to another are derived from right-triangle relationships.

As we did with cylindrical coordinates, let's consider the surfaces that are generated when each of the coordinates is held constant. Let c be a constant, and consider surfaces of the form $\rho = c$. Points on these surfaces are at a fixed distance from the origin and form a sphere. The coordinate θ in the spherical coordinate system is the same as in the cylindrical coordinate system, so surfaces of the form $\theta = c$ are half-planes, as before. Last, consider surfaces of the form $\varphi = c$. The points on these surfaces are at a fixed angle from the z-axis and form a half-cone (Figure 11.6.11).

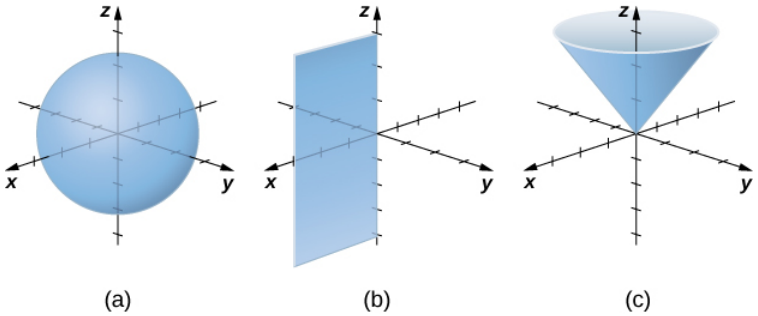


Figure 11.6.11: In spherical coordinates, surfaces of the form $\rho = c$ are spheres of radius ρ (a), surfaces of the form $\theta = c$ are half-planes at an angle θ from the x-axis (b), and surfaces of the form $\phi = c$ are half-cones at an angle ϕ from the z-axis (c).

Example 11.6.4: Converting from Spherical Coordinates

Plot the point with spherical coordinates $(8, \dfrac{\pi}{3}, \dfrac{\pi}{6})$ and express its location in both rectangular and cylindrical coordinates.

Solution

Use the equations in Converting among Spherical, Cylindrical, and Rectangular Coordinates to translate between spherical and cylindrical coordinates (Figure 11.6.12):

$$
\begin{aligned}
x &= \rho \sin\varphi \cos\theta \\
&= 8 \sin\left(\frac{\pi}{6}\right) \cos\left(\frac{\pi}{3}\right) \\
&= 8 \left(\frac{1}{2}\right) \frac{1}{2} \\
&= 2 \\
y &= \rho \sin\varphi \sin\theta \\
&= 8 \sin\left(\frac{\pi}{6}\right) \sin\left(\frac{\pi}{3}\right) \\
&= 8 \left(\frac{1}{2}\right) \frac{\sqrt{3}}{2} \\
&= 2\sqrt{3} \\
z &= \rho \cos\varphi \\
&= 8 \cos\left(\frac{\pi}{6}\right) \\
&= 8 \left(\frac{\sqrt{3}}{2}\right) \\
&= 4\sqrt{3}
\end{aligned}
$$

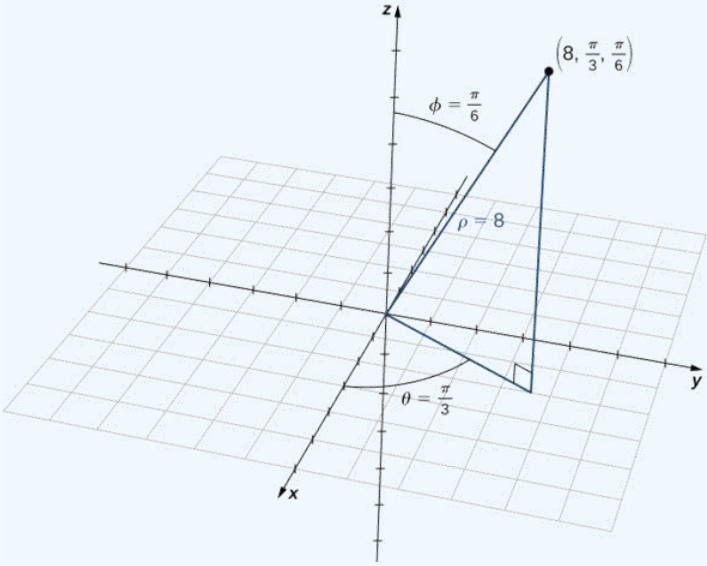


Figure 11.6.12: The projection of the point in the xy-plane is 4 units from the origin. The line from the origin to the point's projection forms an angle of $\pi/3$ with the positive x-axis. The point lies $4\sqrt{3}$ units above the xy-plane.

The point with spherical coordinates $(8, \frac{\pi}{3}, \frac{\pi}{6})$ has rectangular coordinates $(2, 2\sqrt{3}, 4\sqrt{3})$.

Finding the values in cylindrical coordinates is equally straightforward:

$$
\begin{aligned}
r &= \rho \sin \varphi \\
&= 8 \sin \frac{\pi}{6} \\
&= 4 \\
\theta &= \theta \\
z &= \rho \cos \varphi \\
&= 8 \cos \frac{\pi}{6} \\
&= 4\sqrt{3}.
\end{aligned}
$$

Thus, cylindrical coordinates for the point are $(4, \frac{\pi}{3}, 4\sqrt{3})$.

Exercise 11.6.4

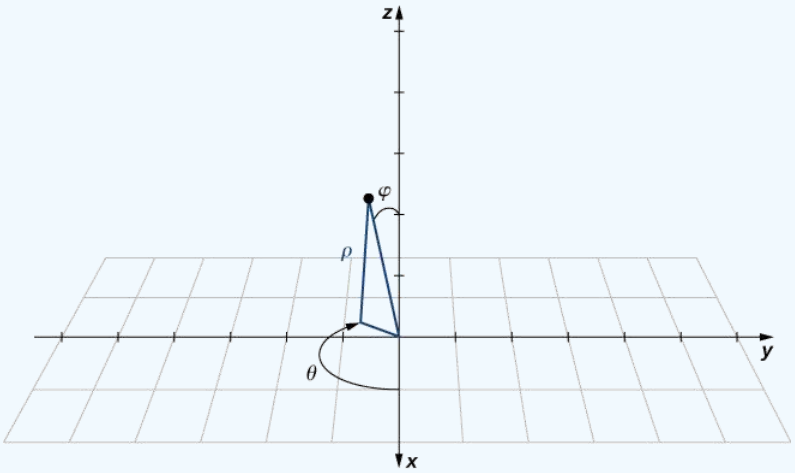
Plot the point with spherical coordinates $(2, -\frac{5\pi}{6}, \frac{\pi}{6})$ and describe its location in both rectangular and cylindrical coordinates.

Hint

Converting the coordinates first may help to find the location of the point in space more easily.

Answer

Cartesian: $(-\frac{\sqrt{3}}{2}, -\frac{1}{2}, \sqrt{3})$, cylindrical: $(1, -\frac{5\pi}{6}, \sqrt{3})$



Example 11.6.5: Converting from Rectangular Coordinates

Convert the rectangular coordinates $(-1, 1, \sqrt{6})$ to both spherical and cylindrical coordinates.

Solution

Start by converting from rectangular to spherical coordinates:

$$
\begin{aligned}
\rho^2 &= x^2 + y^2 + z^2 = (-1)^2 + 1^2 + (\sqrt{6})^2 = 8 \\
\tan \theta &= \frac{1}{-1} \\
\rho &= 2\sqrt{2} \text{ and } \theta = \arctan(-1) = \frac{3\pi}{4}.
\end{aligned}
$$

Because $(x, y) = (-1, 1)$, then the correct choice for θ is $\frac{3\pi}{4}$.

There are actually two ways to identify φ. We can use the equation $\varphi = \arccos\left(\frac{z}{\sqrt{x^2+y^2+z^2}}\right)$. A more simple approach, however, is to use equation $z = \rho\cos\varphi$. We know that $z = \sqrt{6}$ and $\rho = 2\sqrt{2}$, so

$$\sqrt{6} = 2\sqrt{2}\cos\varphi, \text{ so } \cos\varphi = \frac{\sqrt{6}}{2\sqrt{2}} = \frac{\sqrt{3}}{2}$$

and therefore $\varphi = \frac{\pi}{6}$. The spherical coordinates of the point are $(2\sqrt{2}, \frac{3\pi}{4}, \frac{\pi}{6})$.

To find the cylindrical coordinates for the point, we need only find r:

$$r = \rho\sin\varphi = 2\sqrt{2}\sin(\frac{\pi}{6}) = \sqrt{2}.$$

The cylindrical coordinates for the point are $(\sqrt{2}, \frac{3\pi}{4}, \sqrt{6})$.

Example 11.6.6: Identifying Surfaces in the Spherical Coordinate System

Describe the surfaces with the given spherical equations.

a. $\theta = \frac{\pi}{3}$
b. $\varphi = \frac{5\pi}{6}$
c. $\rho = 6$
d. $\rho = \sin\theta\sin\varphi$

Solution

a. The variable θ represents the measure of the same angle in both the cylindrical and spherical coordinate systems. Points with coordinates $(\rho, \frac{\pi}{3}, \varphi)$ lie on the plane that forms angle $\theta = \frac{\pi}{3}$ with the positive x-axis. Because $\rho > 0$, the surface described by equation $\theta = \frac{\pi}{3}$ is the half-plane shown in Figure 11.6.13

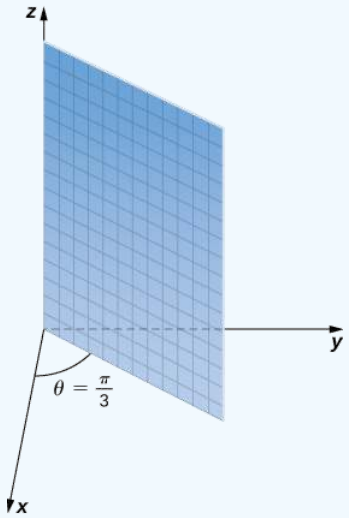


Figure 11.6.13: The surface described by equation $\theta = \frac{\pi}{3}$ is a half-plane.

b. Equation $\varphi = \frac{5\pi}{6}$ describes all points in the spherical coordinate system that lie on a line from the origin forming an angle measuring $\frac{5\pi}{6}$ rad with the positive z-axis. These points form a half-cone (Figure 11.6.14). Because there is only one value for φ that is measured from the positive z-axis, we do not get the full cone (with two pieces).

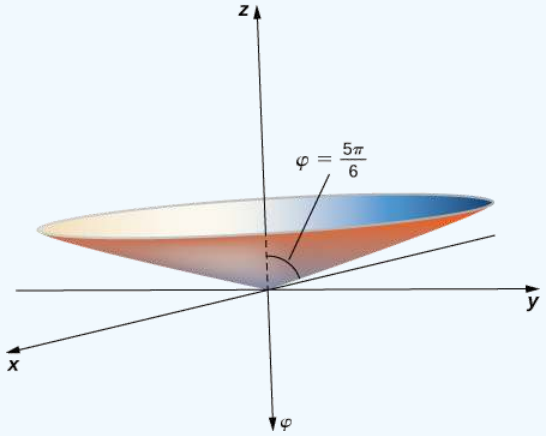


Figure 11.6.14: The equation $\varphi = \dfrac{5\pi}{6}$ describes a cone.

To find the equation in rectangular coordinates, use equation $\varphi = \arccos\left(\dfrac{z}{\sqrt{x^2+y^2+z^2}}\right)$.

$$\begin{aligned} \frac{5\pi}{6} &= \arccos\left(\frac{z}{\sqrt{x^2+y^2+z^2}}\right) \\ \cos\frac{5\pi}{6} &= \frac{z}{\sqrt{x^2+y^2+z^2}} \\ -\frac{\sqrt{3}}{2} &= \frac{z}{\sqrt{x^2+y^2+z^2}} \\ \frac{3}{4} &= \frac{z^2}{x^2+y^2+z^2} \\ \frac{3x^2}{4}+\frac{3y^2}{4}+\frac{3z^2}{4} &= z^2 \\ \frac{3x^2}{4}+\frac{3y^2}{4}-\frac{z^2}{4} &= 0. \end{aligned}$$

This is the equation of a cone centered on the z-axis.

c. Equation $\rho = 6$ describes the set of all points 6 units away from the origin—a sphere with radius 6 (Figure 11.6.15).

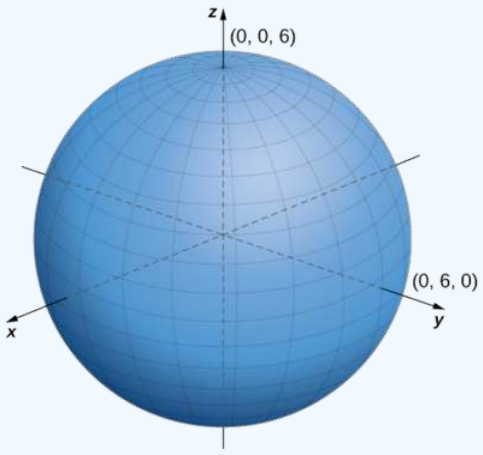


Figure 11.6.15: Equation $\rho = 6$ describes a sphere with radius 6.

d. To identify this surface, convert the equation from spherical to rectangular coordinates, using equations $y = \rho\sin\varphi\sin\theta$ and $\rho^2 = x^2+y^2+z^2$:

$\rho = \sin\theta \sin\varphi$

$\rho^2 = \rho \sin\theta \sin\varphi$ Multiply both sides of the equation by ρ.

$x^2 + y^2 + z^2 = y$ Substitute rectangular variables using the equations above.

$x^2 + y^2 - y + z^2 = 0$ Subtract y from both sides of the equation.

$x^2 + y^2 - y + \frac{1}{4} + z^2 = \frac{1}{4}$ Complete the square.

$x^2 + \left(y - \frac{1}{2}\right)^2 + z^2 = \frac{1}{4}$. Rewrite the middle terms as a perfect square.

The equation describes a sphere centered at point $(0, \frac{1}{2}, 0)$ with radius $\frac{1}{2}$.

? Exercise 11.6.5

Describe the surfaces defined by the following equations.

a. $\rho = 13$
b. $\theta = \frac{2\pi}{3}$
c. $\varphi = \frac{\pi}{4}$

Hint

Think about what each component represents and what it means to hold that component constant.

Answer a

This is the set of all points 13 units from the origin. This set forms a sphere with radius 13.

Answer b

This set of points forms a half plane. The angle between the half plane and the positive x-axis is $\theta = \frac{2\pi}{3}$.

Answer c

Let P be a point on this surface. The position vector of this point forms an angle of $\varphi = \frac{\pi}{4}$ with the positive z-axis, which means that points closer to the origin are closer to the axis. These points form a half-cone.

Spherical coordinates are useful in analyzing systems that have some degree of symmetry about a point, such as the volume of the space inside a domed stadium or wind speeds in a planet's atmosphere. A sphere that has Cartesian equation $x^2 + y^2 + z^2 = c^2$ has the simple equation $\rho = c$ in spherical coordinates.

In geography, latitude and longitude are used to describe locations on Earth's surface, as shown in Figure $11.6.16$ Although the shape of Earth is not a perfect sphere, we use spherical coordinates to communicate the locations of points on Earth. Let's assume Earth has the shape of a sphere with radius 4000 mi. We express angle measures in degrees rather than radians because latitude and longitude are measured in degrees.

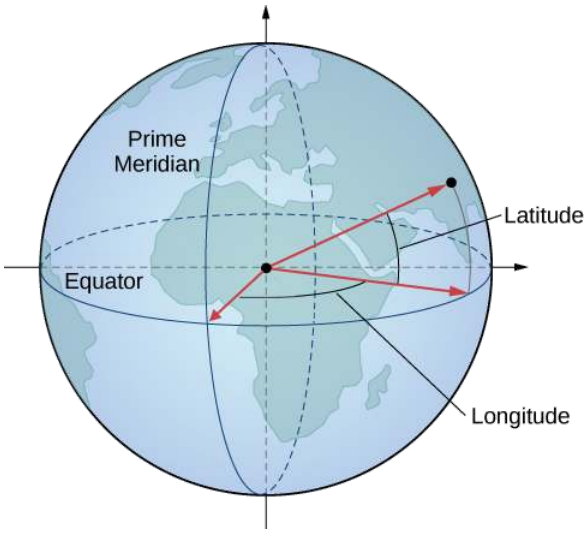


Figure 11.6.16: In the latitude–longitude system, angles describe the location of a point on Earth relative to the equator and the prime meridian.

Let the center of Earth be the center of the sphere, with the ray from the center through the North Pole representing the positive z-axis. The prime meridian represents the trace of the surface as it intersects the xz-plane. The equator is the trace of the sphere intersecting the xy-plane.

Example 11.6.7: Converting Latitude and Longitude to Spherical Coordinates

The latitude of Columbus, Ohio, is 40° N and the longitude is 83° W, which means that Columbus is 40° north of the equator. Imagine a ray from the center of Earth through Columbus and a ray from the center of Earth through the equator directly south of Columbus. The measure of the angle formed by the rays is 40°. In the same way, measuring from the prime meridian, Columbus lies 83° to the west. Express the location of Columbus in spherical coordinates.

Solution

The radius of Earth is 4000mi, so $\rho = 4000$. The intersection of the prime meridian and the equator lies on the positive x-axis. Movement to the west is then described with negative angle measures, which shows that $\theta = -83^\circ$, Because Columbus lies 40° north of the equator, it lies 50° south of the North Pole, so $\varphi = 50^\circ$. In spherical coordinates, Columbus lies at point $(4000, -83^\circ, 50^\circ)$.

Exercise 11.6.6

Sydney, Australia is at $34^\circ S$ and $151^\circ E$. Express Sydney's location in spherical coordinates.

Hint

Because Sydney lies south of the equator, we need to add 90° to find the angle measured from the positive z-axis.

Answer

$(4000, 151^\circ, 124^\circ)$

Cylindrical and spherical coordinates give us the flexibility to select a coordinate system appropriate to the problem at hand. A thoughtful choice of coordinate system can make a problem much easier to solve, whereas a poor choice can lead to unnecessarily complex calculations. In the following example, we examine several different problems and discuss how to select the best coordinate system for each one.

Example 11.6.8: Choosing the Best Coordinate System

In each of the following situations, we determine which coordinate system is most appropriate and describe how we would orient the coordinate axes. There could be more than one right answer for how the axes should be oriented, but we select an orientation that makes sense in the context of the problem. Note: There is not enough information to set up or solve these problems; we simply select the coordinate system (Figure 11.6.17).

a. Find the center of gravity of a bowling ball.
b. Determine the velocity of a submarine subjected to an ocean current.
c. Calculate the pressure in a conical water tank.
d. Find the volume of oil flowing through a pipeline.
e. Determine the amount of leather required to make a football.



(a) (b) (c)



(d) (e)

Figure 11.6.17: (credit: (a) modification of work by scl hua, Wikimedia, (b) modification of work by DVIDSHUB, Flickr, (c) modification of work by Michael Malak, Wikimedia, (d) modification of work by Sean Mack, Wikimedia, (e) modification of work by Elvert Barnes, Flickr)

Solution

a. Clearly, a bowling ball is a sphere, so spherical coordinates would probably work best here. The origin should be located at the physical center of the ball. There is no obvious choice for how the x-, y- and z-axes should be oriented. Bowling balls normally have a weight block in the center. One possible choice is to align the z-axis with the axis of symmetry of the weight block.
b. A submarine generally moves in a straight line. There is no rotational or spherical symmetry that applies in this situation, so rectangular coordinates are a good choice. The z-axis should probably point upward. The x- and y-axes could be aligned to point east and north, respectively. The origin should be some convenient physical location, such as the starting position of the submarine or the location of a particular port.

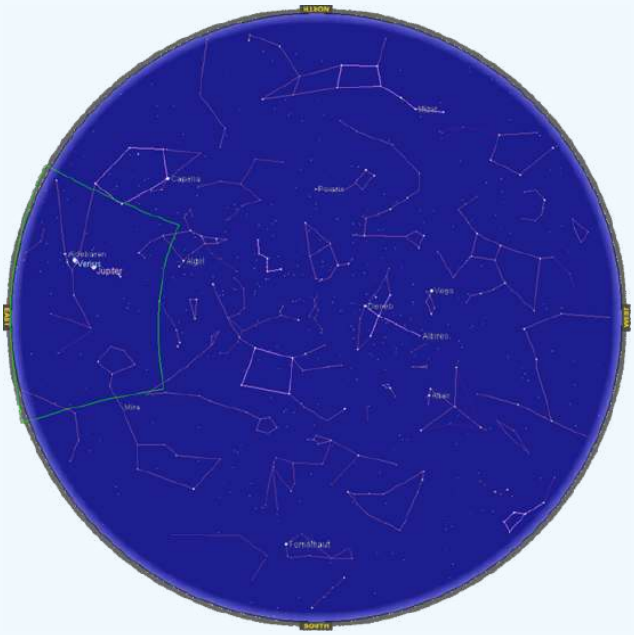
c. A cone has several kinds of symmetry. In cylindrical coordinates, a cone can be represented by equation $z = kr,$ where k is a constant. In spherical coordinates, we have seen that surfaces of the form $\varphi = c$ are half-cones. Last, in rectangular coordinates, elliptic cones are quadric surfaces and can be represented by equations of the form $z^2 = \frac{x^2}{a^2} + \frac{y^2}{b^2}.$ In this case, we could choose any of the three. However, the equation for the surface is more complicated in rectangular coordinates than in the other two systems, so we might want to avoid that choice. In addition, we are talking about a water tank, and the depth of the water might come into play at some point in our calculations, so it might be nice to have a component that represents height and depth directly. Based on this reasoning, cylindrical coordinates might be the best choice. Choose the z-axis to align with the axis of the cone. The orientation of the other two axes is arbitrary. The origin should be the bottom point of the cone.

d. A pipeline is a cylinder, so cylindrical coordinates would be best the best choice. In this case, however, we would likely choose to orient our z-axis with the center axis of the pipeline. The x-axis could be chosen to point straight downward or to some other logical direction. The origin should be chosen based on the problem statement. Note that this puts the z-axis in a horizontal orientation, which is a little different from what we usually do. It may make sense to choose an unusual orientation for the axes if it makes sense for the problem.

e. A football has rotational symmetry about a central axis, so cylindrical coordinates would work best. The z-axis should align with the axis of the ball. The origin could be the center of the ball or perhaps one of the ends. The position of the x-axis is arbitrary.

? Exercise 11.6.7

Which coordinate system is most appropriate for creating a star map, as viewed from Earth (see the following figure)?



How should we orient the coordinate axes?

Hint

What kinds of symmetry are present in this situation?

Answer

Spherical coordinates with the origin located at the center of the earth, the z-axis aligned with the North Pole, and the x-axis aligned with the prime meridian

Key Concepts

- In the cylindrical coordinate system, a point in space is represented by the ordered triple (r, θ, z), where (r, θ) represents the polar coordinates of the point's projection in the xy-plane and z represents the point's projection onto the z-axis.
- To convert a point from cylindrical coordinates to Cartesian coordinates, use equations $x = r\cos\theta, y = r\sin\theta,$ and $z = z$.
- To convert a point from Cartesian coordinates to cylindrical coordinates, use equations $r^2 = x^2 + y^2, \tan\theta = \dfrac{y}{x},$ and $z = z$.
- In the spherical coordinate system, a point P in space is represented by the ordered triple (ρ, θ, φ), where ρ is the distance between P and the origin $(\rho \neq 0), \theta$ is the same angle used to describe the location in cylindrical coordinates, and φ is the angle formed by the positive z-axis and line segment $\overline{OP}$, where O is the origin and $0 \leq \varphi \leq \pi$.
- To convert a point from spherical coordinates to Cartesian coordinates, use equations $x = \rho\sin\varphi\cos\theta, y = \rho\sin\varphi\sin\theta,$ and $z = \rho\cos\varphi$.
- To convert a point from Cartesian coordinates to spherical coordinates, use equations $\rho^2 = x^2 + y^2 + z^2, \tan\theta = \dfrac{y}{x},$ and $\varphi = \arccos\left(\dfrac{z}{\sqrt{x^2 + y^2 + z^2}}\right)$.
- To convert a point from spherical coordinates to cylindrical coordinates, use equations $r = \rho\sin\varphi, \theta = \theta,$ and $z = \rho\cos\varphi$.
- To convert a point from cylindrical coordinates to spherical coordinates, use equations $\rho = \sqrt{r^2 + z^2}, \theta = \theta,$ and $\varphi = \arccos\left(\dfrac{z}{\sqrt{r^2 + z^2}}\right)$.

Glossary

cylindrical coordinate system

a way to describe a location in space with an ordered triple (r, θ, z), where (r, θ) represents the polar coordinates of the point's projection in the xy-plane, and z represents the point's projection onto the z-axis

spherical coordinate system

a way to describe a location in space with an ordered triple (ρ, θ, φ), where ρ is the distance between P and the origin $(\rho \neq 0), \theta$ is the same angle used to describe the location in cylindrical coordinates, and φ is the angle formed by the positive z-axis and line segment $\bar{OP}$, where O is the origin and $0 \leq \varphi \leq \pi$

- Gilbert Strang (MIT) and Edwin "Jed" Herman (Harvey Mudd) with many contributing authors. This content by OpenStax is licensed with a CC-BY-SA-NC 4.0 license. Download for free at http://cnx.org.
- Paul Seeburger edited the LaTeX on the page

11.6: Cylinders and Quadric Surfaces is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **12.7: Cylindrical and Spherical Coordinates** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

CHAPTER OVERVIEW

12: Vector-Valued Functions and Motion in Space

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

12.1: Curves in Space and Their Tangents
12.2: Integrals of Vector Functions; Projectile Motion
12.3: Arc Length in Space
12.4: Curvature and Normal Vectors of a Curve
12.5: Tangential and Normal Components of Acceleration
12.6: Velocity and Acceleration in Polar Coordinates

12: Vector-Valued Functions and Motion in Space is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

12.1: Curves in Space and Their Tangents

Learning Objectives

- Write the general equation of a vector-valued function in component form and unit-vector form.
- Recognize parametric equations for a space curve.
- Describe the shape of a helix and write its equation.
- Define the limit of a vector-valued function.

Our study of vector-valued functions combines ideas from our earlier examination of single-variable calculus with our description of vectors in three dimensions from the preceding chapter. In this section, we extend concepts from earlier chapters and also examine new ideas concerning curves in three-dimensional space. These definitions and theorems support the presentation of material in the rest of this chapter and also in the remaining chapters of the text.

Definition of a Vector-Valued Function

Our first step in studying the calculus of vector-valued functions is to define what exactly a vector-valued function is. We can then look at graphs of vector-valued functions and see how they define curves in both two and three dimensions.

Definition: Vector-valued Functions

A vector-valued function is a function of the form

$$\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} \quad \text{or} \quad \vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\,\hat{\mathbf{k}},$$

where the component functions f, g, and h, are real-valued functions of the parameter t. Vector-valued functions are also written in the form

$$\vec{\mathbf{r}}(t) = \langle f(t),\, g(t)\rangle \quad \text{or} \quad \vec{\mathbf{r}}(t) = \langle f(t),\, g(t),\, h(t)\rangle.$$

In both cases, the first form of the function defines a two-dimensional vector-valued function; the second form describes a three-dimensional vector-valued function.

The parameter t can lie between two real numbers: $a \le t \le b$. Another possibility is that the value of t might take on all real numbers. Last, the component functions themselves may have domain restrictions that enforce restrictions on the value of t. We often use t as a parameter because t can represent time.

Example 12.1.1: Evaluating Vector-Valued Functions and Determining Domains

For each of the following vector-valued functions, evaluate $\vec{\mathbf{r}}(0)$, $\vec{\mathbf{r}}\left(\frac{\pi}{2}\right)$, and $\vec{\mathbf{r}}\left(\frac{2\pi}{3}\right)$. Do any of these functions have domain restrictions?

a. $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} + 3\sin t\,\hat{\mathbf{j}}$
b. $\vec{\mathbf{r}}(t) = 3\tan t\,\hat{\mathbf{i}} + 4\sec t\,\hat{\mathbf{j}} + 5t\,\hat{\mathbf{k}}$

Solution

a. To calculate each of the function values, substitute the appropriate value of t into the function:

$$\begin{aligned}\vec{\mathbf{r}}(0) &= 4\cos(0)\hat{\mathbf{i}} + 3\sin(0)\hat{\mathbf{j}} \\ &= 4\hat{\mathbf{i}} + 0\hat{\mathbf{j}} = 4\hat{\mathbf{i}} \\ \vec{\mathbf{r}}\left(\frac{\pi}{2}\right) &= 4\cos\left(\frac{\pi}{2}\right)\hat{\mathbf{i}} + 3\sin\left(\frac{\pi}{2}\right)\hat{\mathbf{j}} \\ &= 0\hat{\mathbf{i}} + 3\hat{\mathbf{j}} = 3\hat{\mathbf{j}} \\ \vec{\mathbf{r}}\left(\frac{2\pi}{3}\right) &= 4\cos\left(\frac{2\pi}{3}\right)\hat{\mathbf{i}} + 3\sin\left(\frac{2\pi}{3}\right)\hat{\mathbf{j}} \\ &= 4\left(-\tfrac{1}{2}\right)\hat{\mathbf{i}} + 3\left(\tfrac{\sqrt{3}}{2}\right)\hat{\mathbf{j}} = -2\hat{\mathbf{i}} + \tfrac{3\sqrt{3}}{2}\hat{\mathbf{j}}\end{aligned}$$

To determine whether this function has any domain restrictions, consider the component functions separately. The first component function is $f(t) = 4\cos t$ and the second component function is $g(t) = 3\sin t$. Neither of these functions has a domain restriction, so the domain of $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} + 3\sin t\,\hat{\mathbf{j}}$ is all real numbers.

b. To calculate each of the function values, substitute the appropriate value of t into the function:

$$\begin{aligned}\vec{\mathbf{r}}(0) &= 3\tan(0)\hat{\mathbf{i}} + 4\sec(0)\hat{\mathbf{j}} + 5(0)\hat{\mathbf{k}} \\ &= 0\hat{\mathbf{i}} + 4j + 0\hat{\mathbf{k}} = 4\hat{\mathbf{j}} \\ \vec{\mathbf{r}}\left(\frac{\pi}{2}\right) &= 3\tan\left(\frac{\pi}{2}\right)\hat{\mathbf{i}} + 4\sec\left(\frac{\pi}{2}\right)\hat{\mathbf{j}} + 5\left(\frac{\pi}{2}\right)\hat{\mathbf{k}}, \text{ which does not exist} \\ \vec{\mathbf{r}}\left(\frac{2\pi}{3}\right) &= 3\tan\left(\frac{2\pi}{3}\right)\hat{\mathbf{i}} + 4\sec\left(\frac{2\pi}{3}\right)\hat{\mathbf{j}} + 5\left(\frac{2\pi}{3}\right)\hat{\mathbf{k}} \\ &= 3(-\sqrt{3})\hat{\mathbf{i}} + 4(-2)\hat{\mathbf{j}} + \frac{10\pi}{3}\hat{\mathbf{k}} \\ &= (-3\sqrt{3})\hat{\mathbf{i}} - 8\hat{\mathbf{j}} + \frac{10\pi}{3}\hat{\mathbf{k}}\end{aligned}$$

To determine whether this function has any domain restrictions, consider the component functions separately. The first component function is $f(t) = 3\tan t$, the second component function is $g(t) = 4\sec t$, and the third component function is $h(t) = 5t$. The first two functions are not defined for odd multiples of $\frac{\pi}{2}$, so the function is not defined for odd multiples of $\frac{\pi}{2}$. Therefore,

$$\mathrm{D}_{\vec{\mathbf{r}}} = \left\{t \,|\, t \neq \frac{(2n+1)\pi}{2}\right\},$$

where n is any integer.

? Exercise 12.1.1

For the vector-valued function $\vec{\mathbf{r}}(t) = (t^2 - 3t)\,\hat{\mathbf{i}} + (4t+1)\,\hat{\mathbf{j}}$, evaluate $\vec{\mathbf{r}}(0)$, $\vec{\mathbf{r}}(1)$, and $\vec{\mathbf{r}}(-4)$. Does this function have any domain restrictions?

Hint

Substitute the appropriate values of t into the function.

Answer

$\vec{\mathbf{r}}(0) = \hat{\mathbf{j}}$, $\vec{\mathbf{r}}(1) = -2\hat{\mathbf{i}} + 5\hat{\mathbf{j}}$, $\vec{\mathbf{r}}(-4) = 28\hat{\mathbf{i}} - 15\hat{\mathbf{j}}$

The domain of $\vec{\mathbf{r}}(t) = (t^2 - 3t)\hat{\mathbf{i}} + (4t+1)\hat{\mathbf{j}}$ is all real numbers.

Example 12.1.1 illustrates an important concept. The domain of a vector-valued function consists of real numbers. The domain can be all real numbers or a subset of the real numbers. The range of a vector-valued function consists of vectors. Each real number in the domain of a vector-valued function is mapped to either a two- or a three-dimensional vector.

Graphing Vector-Valued Functions

Recall that a plane vector consists of two quantities: direction and magnitude. Given any point in the plane (the *initial point*), if we move in a specific direction for a specific distance, we arrive at a second point. This represents the *terminal point* of the vector. We calculate the components of the vector by subtracting the coordinates of the initial point from the coordinates of the terminal point.

A vector is considered to be in *standard position* if the initial point is located at the origin. When graphing a vector-valued function, we typically graph the vectors in the domain of the function in standard position, because doing so guarantees the uniqueness of the graph. This convention applies to the graphs of three-dimensional vector-valued functions as well. The graph of a vector-valued function of the form

$$\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}}$$

consists of the set of all points $(f(t),\ g(t))$, and the path it traces is called a **plane curve**. The graph of a vector-valued function of the form

$$\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\,\hat{\mathbf{k}}$$

consists of the set of all points $(f(t),\ g(t),\ h(t))$, and the path it traces is called a **space curve**. Any representation of a plane curve or space curve using a vector-valued function is called a **vector parameterization** of the curve.

Each plane curve and space curve has an **orientation**, indicated by arrows drawn in on the curve, that shows the direction of motion along the curve as the value of the parameter t increases.

Example 12.1.2 : Graphing a Vector-Valued Function

Create a graph of each of the following vector-valued functions:

a. The plane curve represented by $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} + 3\sin t\,\hat{\mathbf{j}}$, $0 \le t \le 2\pi$
b. The plane curve represented by $\vec{\mathbf{r}}(t) = 4\cos(t^3)\,\hat{\mathbf{i}} + 3\sin(t^3)\,\hat{\mathbf{j}}$, $0 \le t \le \sqrt[3]{2\pi}$
c. The space curve represented by $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} + 4\sin t\,\hat{\mathbf{j}} + t\,\hat{\mathbf{k}}$, $0 \le t \le 4\pi$

Solution

a. As with any graph, we start with a table of values. We then graph each of the vectors in the second column of the table in standard position and connect the terminal points of each vector to form a curve (Figure 12.1.1). This curve turns out to be an ellipse centered at the origin.

Table 12.1.1: Table of Values for $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} + 3\sin t\,\hat{\mathbf{j}}$, $0 \le t \le 2\pi$

t	$\vec{\mathbf{r}}(t)$	t	$\vec{\mathbf{r}}(t)$
0	$4\hat{\mathbf{i}}$	π	$-4\hat{\mathbf{i}}$
$\frac{\pi}{4}$	$2\sqrt{2}\hat{\mathbf{i}} + \frac{3\sqrt{2}}{2}\hat{\mathbf{j}}$	$\frac{5\pi}{4}$	$-2\sqrt{2}\hat{\mathbf{i}} - \frac{3\sqrt{2}}{2}\hat{\mathbf{j}}$
$\frac{\pi}{2}$	$3\hat{\mathbf{j}}$	$\frac{3\pi}{2}$	$-3\hat{\mathbf{j}}$
$\frac{3\pi}{4}$	$-2\sqrt{2}\hat{\mathbf{i}} + \frac{3\sqrt{2}}{2}\hat{\mathbf{j}}$	$\frac{7\pi}{4}$	$2\sqrt{2}\hat{\mathbf{i}} - \frac{3\sqrt{2}}{2}\hat{\mathbf{j}}$
2π	$4\hat{\mathbf{i}}$		

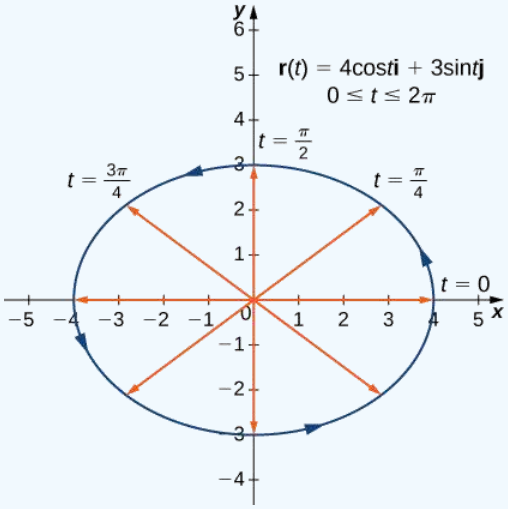


Figure 12.1.1: The graph of the first vector-valued function is an ellipse.

b. The table of values for $\vec{\mathbf{r}}(t) = 4\cos(t^3)\,\hat{\mathbf{i}} + 3\sin(t^3)\,\hat{\mathbf{j}}$, $0 \le t \le \sqrt[3]{2\pi}$ is as follows:

Table of Values for $\vec{\mathbf{r}}(t) = 4\cos(t^3)\,\hat{\mathbf{i}} + 3\sin(t^3)\,\hat{\mathbf{j}}$, $0 \le t \le \sqrt[3]{2\pi}$

t	$\vec{\mathbf{r}}(t)$	t	$\vec{\mathbf{r}}(t)$
0	$4\hat{\mathbf{i}}$	$\sqrt[3]{\pi}$	$-4\hat{\mathbf{i}}$
$\sqrt[3]{\frac{\pi}{4}}$	$2\sqrt{2}\hat{\mathbf{i}} + \frac{3\sqrt{2}}{2}\hat{\mathbf{j}}$	$\sqrt[3]{\frac{5\pi}{4}}$	$-2\sqrt{2}\hat{\mathbf{i}} - \frac{3\sqrt{2}}{2}\hat{\mathbf{j}}$
$\sqrt[3]{\frac{\pi}{2}}$	$3\hat{\mathbf{j}}$	$\sqrt[3]{\frac{3\pi}{2}}$	$-3\hat{\mathbf{j}}$
$\sqrt[3]{\frac{3\pi}{4}}$	$-2\sqrt{2}\hat{\mathbf{i}} + \frac{3\sqrt{2}}{2}\hat{\mathbf{j}}$	$\sqrt[3]{\frac{7\pi}{4}}$	$2\sqrt{2}\hat{\mathbf{i}} - \frac{3\sqrt{2}}{2}\hat{\mathbf{j}}$
$\sqrt[3]{2\pi}$	$4\hat{\mathbf{i}}$		

The graph of this curve is also an ellipse centered at the origin.

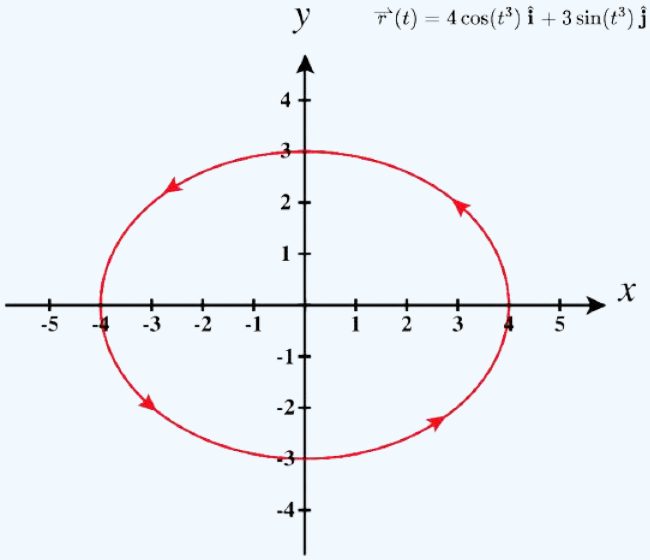


Figure 12.1.2: The graph of the second vector-valued function is also an ellipse.

c. We go through the same procedure for a three-dimensional vector function.

Table of Values for $\mathrm{r}(\mathrm{t}) = 4\cos \mathrm{t}\,\hat{\mathbf{i}} + 4\sin \mathrm{t}\,\hat{\mathbf{j}} + \mathrm{t}\hat{\mathbf{k}}$, $0 \le \mathrm{t} \le 4\pi$

t	$\vec{\mathbf{r}}(t)$	t	$\vec{\mathbf{r}}(t)$
0	$4\hat{\mathbf{i}}$	π	$-4\hat{\mathbf{i}} + \pi\hat{\mathbf{k}}$
$\frac{\pi}{4}$	$2\sqrt{2}\hat{\mathbf{i}} + 2\sqrt{2}\hat{\mathbf{j}} + \frac{\pi}{4}\hat{\mathbf{k}}$	$\frac{5\pi}{4}$	$-2\sqrt{2}\hat{\mathbf{i}} - 2\sqrt{2}\hat{\mathbf{j}} + \frac{5\pi}{4}\hat{\mathbf{k}}$
$\frac{\pi}{2}$	$4\hat{\mathbf{j}} + \frac{\pi}{2}\hat{\mathbf{k}}$	$\frac{3\pi}{2}$	$-4\hat{\mathbf{j}} + \frac{3\pi}{2}\hat{\mathbf{k}}$
$\frac{3\pi}{4}$	$-2\sqrt{2}\hat{\mathbf{i}} + 2\sqrt{2}\hat{\mathbf{j}} + \frac{3\pi}{4}\hat{\mathbf{k}}$	$\frac{7\pi}{4}$	$2\sqrt{2}\hat{\mathbf{i}} - 2\sqrt{2}\hat{\mathbf{j}} + \frac{7\pi}{4}\hat{\mathbf{k}}$
2π	$4\hat{\mathbf{j}} + 2\pi\hat{\mathbf{k}}$		

The values then repeat themselves, except for the fact that the coefficient of $\hat{\mathbf{k}}$ is always increasing (12.1.3). This curve is called a helix. Notice that if the $\hat{\mathbf{k}}$ component is eliminated, then the function becomes $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} + 4\sin t\,\hat{\mathbf{j}}$, which is a circle of radius 4 centered at the origin.

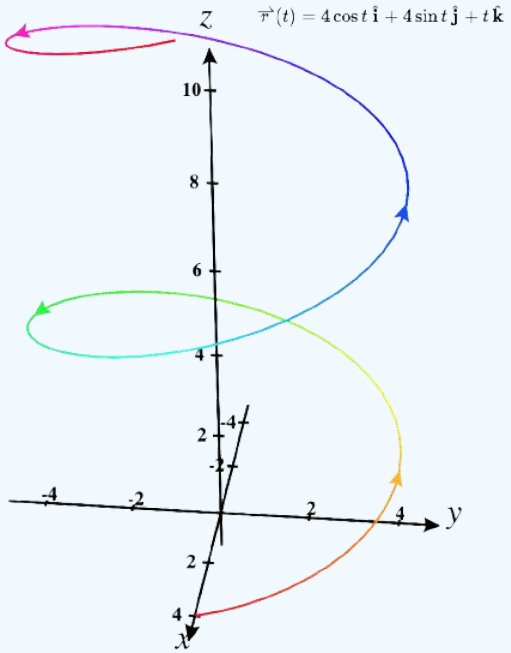


Figure 12.1.3: The graph of the third vector-valued function is a helix.

You may notice that the graphs in parts a. and b. are identical. This happens because the function describing the curve in part b. is a so-called reparameterization of the function describing the curve in part a. In fact, any curve has an infinite number of reparameterizations; for example, we can replace t with $2t$ in any of the three previous curves without changing the shape of the curve. The interval over which t is defined may change, but that is all. We return to this idea later in this chapter when we study arc-length parameterization. As mentioned, the name of the shape of the curve of the graph in 12.1.3 is a *helix*. The curve resembles a spring, with a circular cross-section looking down along the z-axis. It is possible for a helix to be elliptical in cross-section as well. For example, the vector-valued function $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} + 3\sin t\,\hat{\mathbf{j}} + t\,\hat{\mathbf{k}}$ describes an elliptical helix. The projection of this helix into the xy-plane is an ellipse. Last, the arrows in the graph of this helix indicate the orientation of the curve as t progresses from 0 to 4π.

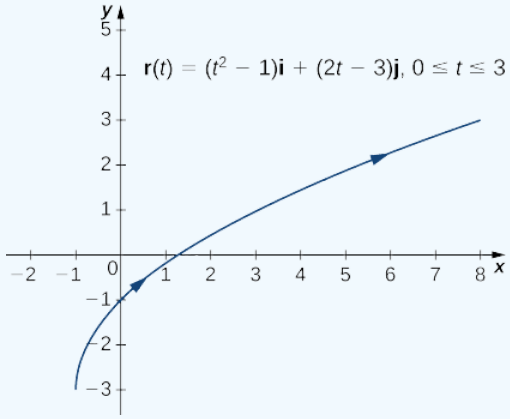
? Exercise 12.1.2

Create a graph of the vector-valued function $\vec{\mathbf{r}}(t) = (t^2 - 1)\hat{\mathbf{i}} + (2t - 3)\hat{\mathbf{j}}$, $0 \le t \le 3$.

Hint

Start by making a table of values, then graph the vectors for each value of t.

Answer



At this point, you may notice a similarity between vector-valued functions and parameterized curves. Indeed, given a vector-valued function $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}}$ we can define $x = f(t)$ and $y = g(t)$. If a restriction exists on the values of t (for example, t is restricted to the interval $[a, b]$ for some constants $a < b$, then this restriction is enforced on the parameter. The graph of the parameterized function would then agree with the graph of the vector-valued function, except that the vector-valued graph would represent vectors rather than points. Since we can parameterize a curve defined by a function $y = f(x)$, it is also possible to represent an arbitrary plane curve by a vector-valued function.

Limits and Continuity of a Vector-Valued Function

We now take a look at the limit of a vector-valued function. This is important to understand to study the calculus of vector-valued functions.

Definition: limit of a vector-valued function

A vector-valued function $\vec{\mathbf{r}}$ approaches the limit $\vec{\mathbf{L}}$ as t approaches a, written

$$\lim_{t \to a} \vec{\mathbf{r}}(t) = \vec{\mathbf{L}},$$

provided

$$\lim_{t \to a} \left\| \vec{\mathbf{r}}(t) - \vec{\mathbf{L}} \right\| = 0.$$

This is a rigorous definition of the limit of a vector-valued function. In practice, we use the following theorem:

Theorem: Limit of a vector-valued function

Let f, g, and h be functions of t. Then the limit of the vector-valued function $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}}$ as t approaches a is given by

$$\lim_{t \to a} \vec{\mathbf{r}}(t) = \left[\lim_{t \to a} f(t)\right] \hat{\mathbf{i}} + \left[\lim_{t \to a} g(t)\right] \hat{\mathbf{j}}, \tag{12.1.1}$$

provided the limits $\lim_{t \to a} f(t)$ and $\lim_{t \to a} g(t)$ exist.

Similarly, the limit of the vector-valued function $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\hat{\mathbf{k}}$ as t approaches a is given by

$$\lim_{t \to a} \vec{\mathbf{r}}(t) = \left[\lim_{t \to a} f(t)\right] \hat{\mathbf{i}} + \left[\lim_{t \to a} g(t)\right] \hat{\mathbf{j}} + \left[\lim_{t \to a} h(t)\right] \hat{\mathbf{k}}, \tag{12.1.2}$$

provided the limits $\lim_{t \to a} f(t)$, $\lim_{t \to a} g(t)$ and $\lim_{t \to a} h(t)$ exist.

In the following example, we show how to calculate the limit of a vector-valued function.

Example 12.1.3: Evaluating the Limit of a Vector-Valued Function

For each of the following vector-valued functions, calculate $\lim\limits_{t\to 3}\vec{\mathbf{r}}(t)$ for

a. $\vec{\mathbf{r}}(t)=(t^2-3t+4)\hat{\mathbf{i}}+(4t+3)\hat{\mathbf{j}}$
b. $\vec{\mathbf{r}}(t)=\frac{2t-4}{t+1}\hat{\mathbf{i}}+\frac{t}{t^2+1}\hat{\mathbf{j}}+(4t-3)\hat{\mathbf{k}}$

Solution

a. Use Equation 12.1.1 and substitute the value $t=3$ into the two component expressions:

$$\begin{aligned}\lim_{t\to 3}\vec{\mathbf{r}}(t) &= \lim_{t\to 3}\left[(t^2-3t+4)\hat{\mathbf{i}}+(4t+3)\hat{\mathbf{j}}\right]\\ &= \left[\lim_{t\to 3}(t^2-3t+4)\right]\hat{\mathbf{i}}+\left[\lim_{t\to 3}(4t+3)\right]\hat{\mathbf{j}}\\ &= 4\hat{\mathbf{i}}+15\hat{\mathbf{j}}\end{aligned}$$

b. Use Equation 12.1.2 and substitute the value $t=3$ into the three component expressions:

$$\begin{aligned}\lim_{t\to 3}\vec{\mathbf{r}}(t) &= \lim_{t\to 3}\left(\frac{2t-4}{t+1}\hat{\mathbf{i}}+\frac{t}{t^2+1}\hat{\mathbf{j}}+(4t-3)\hat{\mathbf{k}}\right)\\ &= \left[\lim_{t\to 3}\left(\frac{2t-4}{t+1}\right)\right]\hat{\mathbf{i}}+\left[\lim_{t\to 3}\left(\frac{t}{t^2+1}\right)\right]\hat{\mathbf{j}}+\left[\lim_{t\to 3}(4t-3)\right]\hat{\mathbf{k}}\\ &= \tfrac{1}{2}\hat{\mathbf{i}}+\tfrac{3}{10}\hat{\mathbf{j}}+9\hat{\mathbf{k}}\end{aligned}$$

Exercise 12.1.3

Calculate $\lim\limits_{t\to 2}\vec{\mathbf{r}}(t)$ for the function $\vec{\mathbf{r}}(t)=\sqrt{t^2+3t-1}\ \hat{\mathbf{i}}-(4t-3)\hat{\mathbf{j}}-\sin\frac{(t+1)\pi}{2}\hat{\mathbf{k}}$

Hint

Use Equation 12.1.2 from the preceding theorem.

Answer

$$\lim_{t\to 2}\vec{\mathbf{r}}(t)=3\hat{\mathbf{i}}-5\hat{\mathbf{j}}+\hat{\mathbf{k}}$$

Now that we know how to calculate the limit of a vector-valued function, we can define **continuity at a point** for such a function.

Definitions

Let f, g, and h be functions of t. Then, the vector-valued function $\vec{\mathbf{r}}(t)=f(t)\hat{\mathbf{i}}+g(t)\hat{\mathbf{j}}$ is **continuous at point $t=a$** if the following three conditions hold:

1. $\vec{\mathbf{r}}(a)$ exists
2. $\lim\limits_{t\to a}\vec{\mathbf{r}}(t)$ exists
3. $\lim\limits_{t\to a}\vec{\mathbf{r}}(t)=\vec{\mathbf{r}}(a)$

Similarly, the vector-valued function $\vec{\mathbf{r}}(t)=f(t)\hat{\mathbf{i}}+g(t)\hat{\mathbf{j}}+h(t)\hat{\mathbf{k}}$ is **continuous at point $t=a$** if the following three conditions hold:

1. $\vec{\mathbf{r}}(a)$ exists
2. $\lim\limits_{t\to a}\vec{\mathbf{r}}(t)$ exists
3. $\lim\limits_{t\to a}\vec{\mathbf{r}}(t)=\vec{\mathbf{r}}(a)$

Summary

- A vector-valued function is a function of the form $\vec{\mathbf{r}}(t) = f(t)\hat{\mathbf{i}} + g(t)\hat{\mathbf{j}}$ or $\vec{\mathbf{r}}(t) = f(t)\hat{\mathbf{i}} + g(t)\hat{\mathbf{j}} + h(t)\hat{\mathbf{k}}$, where the component functions f, g, and h are real-valued functions of the parameter t.
- The graph of a vector-valued function of the form $\vec{\mathbf{r}}(t) = f(t)\hat{\mathbf{i}} + g(t)\hat{\mathbf{j}}$ is called a *plane curve*. The graph of a vector-valued function of the form $\vec{\mathbf{r}}(t) = f(t)\hat{\mathbf{i}} + g(t)\hat{\mathbf{j}} + h(t)\hat{\mathbf{k}}$ is called a *space curve*.
- It is possible to represent an arbitrary plane curve by a vector-valued function.
- To calculate the limit of a vector-valued function, calculate the limits of the component functions separately.

Key Equations

- **Vector-valued function**
  $\vec{\mathbf{r}}(t) = f(t)\hat{\mathbf{i}} + g(t)\hat{\mathbf{j}}$ or $\vec{\mathbf{r}}(t) = f(t)\hat{\mathbf{i}} + g(t)\hat{\mathbf{j}} + h(t)\hat{\mathbf{k}}$,or $\vec{\mathbf{r}}(t) = \langle f(t), g(t)\rangle$ or $\vec{\mathbf{r}}(t) = \langle f(t), g(t), h(t)\rangle$
- **Limit of a vector-valued function**
  $\lim\limits_{t\to a}\vec{\mathbf{r}}(t) = [\lim\limits_{t\to a} f(t)]\,\hat{\mathbf{i}} + [\lim\limits_{t\to a} g(t)]\,\hat{\mathbf{j}}$ or $\lim\limits_{t\to a}\vec{\mathbf{r}}(t) = [\lim\limits_{t\to a} f(t)]\,\hat{\mathbf{i}} + [\lim\limits_{t\to a} g(t)]\,\hat{\mathbf{j}} + [\lim\limits_{t\to a} h(t)]\hat{\mathbf{k}}$

Glossary

component functions

the component functions of the vector-valued function $\vec{\mathbf{r}}(t) = f(t)\hat{\mathbf{i}} + g(t)\hat{\mathbf{j}}$ are $f(t)$ and $g(t)$, and the component functions of the vector-valued function $\vec{\mathbf{r}}(t) = f(t)\hat{\mathbf{i}} + g(t)\hat{\mathbf{j}} + h(t)\hat{\mathbf{k}}$ are $f(t)$, $g(t)$ and $h(t)$

helix

a three-dimensional curve in the shape of a spiral

limit of a vector-valued function

a vector-valued function $\vec{\mathbf{r}}(t)$ has a limit $\vec{\mathbf{L}}$ as t approaches a if $\lim t \to a \left|\vec{\mathbf{r}}(t) - \vec{\mathbf{L}}\right| = 0$

plane curve

the set of ordered pairs $(f(t), g(t))$ together with their defining parametric equations $x = f(t)$ and $y = g(t)$

reparameterization

an alternative parameterization of a given vector-valued function

space curve

the set of ordered triples $(f(t), g(t), h(t))$ together with their defining parametric equations $x = f(t)$, $y = g(t)$ and $z = h(t)$

vector parameterization

any representation of a plane or space curve using a vector-valued function

vector-valued function

a function of the form $\vec{\mathbf{r}}(t) = f(t)\hat{\mathbf{i}} + g(t)\hat{\mathbf{j}}$ or $\vec{\mathbf{r}}(t) = f(t)\hat{\mathbf{i}} + g(t)\hat{\mathbf{j}} + h(t)\hat{\mathbf{k}}$,where the component functions f, g, and h are real-valued functions of the parameter t.

- Gilbert Strang (MIT) and Edwin “Jed” Herman (Harvey Mudd) with many contributing authors. This content by OpenStax is licensed with a CC-BY-SA-NC 4.0 license. Download for free at http://cnx.org.
- Edited by Paul Seeburger (Monroe Community College)

12.1: Curves in Space and Their Tangents is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **13.1: Vector-Valued Functions and Space Curves** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

12.2: Integrals of Vector Functions; Projectile Motion

Learning Objectives

- Write an expression for the derivative of a vector-valued function.
- Find the tangent vector at a point for a given position vector.
- Find the unit tangent vector at a point for a given position vector and explain its significance.
- Calculate the definite integral of a vector-valued function.

To study the calculus of vector-valued functions, we follow a similar path to the one we took in studying real-valued functions. First, we define the derivative, then we examine applications of the derivative, then we move on to defining integrals. However, we will find some interesting new ideas along the way as a result of the vector nature of these functions and the properties of space curves.

Derivatives of Vector-Valued Functions

Now that we have seen what a vector-valued function is and how to take its limit, the next step is to learn how to differentiate a vector-valued function. The definition of the derivative of a vector-valued function is nearly identical to the definition of a real-valued function of one variable. However, because the range of a vector-valued function consists of vectors, the same is true for the range of the derivative of a vector-valued function.

Definition: Derivative of Vector-Valued Functions

The derivative of a vector-valued function $\vec{\mathbf{r}}(t)$ is

$$\vec{\mathbf{r}}\,'(t) = \lim_{\Delta t \to 0} \frac{\vec{\mathbf{r}}(t+\Delta t) - \vec{\mathbf{r}}(t)}{\Delta t} \tag{12.2.1}$$

provided the limit exists. If $\vec{\mathbf{r}}\,'(t)$ exists, then $\vec{\mathbf{r}}(t)$ is differentiable at t. If $\vec{\mathbf{r}}\,'(t)$ exists for all t in an open interval (a, b) then $\vec{\mathbf{r}}(t)$ is differentiable over the interval (a, b). For the function to be differentiable over the closed interval $[a, b]$, the following two limits must exist as well:

$$\vec{\mathbf{r}}\,'(a) = \lim_{\Delta t \to 0^+} \frac{\vec{\mathbf{r}}(a+\Delta t) - \vec{\mathbf{r}}(a)}{\Delta t}$$

and

$$\vec{\mathbf{r}}\,'(b) = \lim_{\Delta t \to 0^-} \frac{\vec{\mathbf{r}}(b+\Delta t) - \vec{\mathbf{r}}(b)}{\Delta t}$$

Many of the rules for calculating derivatives of real-valued functions can be applied to calculating the derivatives of vector-valued functions as well. Recall that the derivative of a real-valued function can be interpreted as the slope of a tangent line or the instantaneous rate of change of the function. The derivative of a vector-valued function can be understood to be an instantaneous rate of change as well; for example, when the function represents the position of an object at a given point in time, the derivative represents its velocity at that same point in time.

We now demonstrate taking the derivative of a vector-valued function.

Example 12.2.1: Finding the Derivative of a Vector-Valued Function

Use the definition to calculate the derivative of the function

$$\vec{\mathbf{r}}(t) = (3t+4)\,\hat{\mathbf{i}} + (t^2 - 4t + 3)\,\hat{\mathbf{j}}.$$

Solution

Let's use Equation 12.2.1:

$$\begin{aligned}
\vec{\mathbf{r}}\,'(t) &= \lim_{\Delta t \to 0} \frac{\vec{\mathbf{r}}(t+\Delta t) - \vec{\mathbf{r}}(t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{[(3(t+\Delta t)+4)\,\hat{\mathbf{i}} + ((t+\Delta t)^2 - 4(t+\Delta t)+3)\,\hat{\mathbf{j}}] - [(3t+4)\,\hat{\mathbf{i}} + (t^2-4t+3)\,\hat{\mathbf{j}}]}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{(3t+3\Delta t+4)\,\hat{\mathbf{i}} - (3t+4)\,\hat{\mathbf{i}} + (t^2+2t\Delta t+(\Delta t)^2-4t-4\Delta t+3)\,\hat{\mathbf{j}} - (t^2-4t+3)\,\hat{\mathbf{j}}}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{(3\Delta t)\,\hat{\mathbf{i}} + (2t\Delta t+(\Delta t)^2-4\Delta t)\,\hat{\mathbf{j}}}{\Delta t} \\
&= \lim_{\Delta t \to 0} (3\,\hat{\mathbf{i}} + (2t+\Delta t-4)\,\hat{\mathbf{j}}) \\
&= 3\,\hat{\mathbf{i}} + (2t-4)\,\hat{\mathbf{j}}
\end{aligned}$$

? Exercise 12.2.1

Use the definition to calculate the derivative of the function $\vec{\mathbf{r}}(t) = (2t^2+3)\,\hat{\mathbf{i}} + (5t-6)\,\hat{\mathbf{j}}$.

Hint

Use Equation 12.2.1.

Answer

$$\vec{\mathbf{r}}\,'(t) = 4t\,\hat{\mathbf{i}} + 5\,\hat{\mathbf{j}}$$

Notice that in the calculations in Example 12.2.1, we could also obtain the answer by first calculating the derivative of each component function, then putting these derivatives back into the vector-valued function. This is always true for calculating the derivative of a vector-valued function, whether it is in two or three dimensions. We state this in the following theorem. The proof of this theorem follows directly from the definitions of the limit of a vector-valued function and the derivative of a vector-valued function.

Theorem 12.2.1: Differentiation of Vector-Valued Functions

Let f, g, and h be differentiable functions of t.

1. If $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}}$ then
$$\vec{\mathbf{r}}\,'(t) = f\,'(t)\,\hat{\mathbf{i}} + g'(t)\,\hat{\mathbf{j}}.$$
2. If $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\,\hat{\mathbf{k}}$ then
$$\vec{\mathbf{r}}\,'(t) = f\,'(t)\,\hat{\mathbf{i}} + g'(t)\,\hat{\mathbf{j}} + h\,'(t)\,\hat{\mathbf{k}}.$$

✓ Example 12.2.2: Calculating the Derivative of Vector-Valued Functions

Use Theorem 12.2.1 to calculate the derivative of each of the following functions.

a. $\vec{\mathbf{r}}(t) = (6t+8)\,\hat{\mathbf{i}} + (4t^2+2t-3)\,\hat{\mathbf{j}}$
b. $\vec{\mathbf{r}}(t) = 3\cos t\,\hat{\mathbf{i}} + 4\sin t\,\hat{\mathbf{j}}$
c. $\vec{\mathbf{r}}(t) = e^t \sin t\,\hat{\mathbf{i}} + e^t \cos t\,\hat{\mathbf{j}} - e^{2t}\,\hat{\mathbf{k}}$

Solution

We use Theorem 12.2.1 and what we know about differentiating functions of one variable.

a. The first component of
$$\vec{\mathbf{r}}(t) = (6t+8)\,\hat{\mathbf{i}} + (4t^2+2t-3)\,\hat{\mathbf{j}}$$

is $f(t)=6t+8$. The second component is $g(t)=4t^2+2t-3$. We have $f'(t)=6$ and $g'(t)=8t+2$, so the Theorem 12.2.1 gives $\vec{\mathbf{r}}\,'(t)=6\,\hat{\mathbf{i}}+(8t+2)\,\hat{\mathbf{j}}$.

b. The first component is $f(t)=3\cos t$ and the second component is $g(t)=4\sin t$. We have $f'(t)=-3\sin t$ and $g'(t)=4\cos t$, so we obtain $\vec{\mathbf{r}}\,'(t)=-3\sin t\,\hat{\mathbf{i}}+4\cos t\,\hat{\mathbf{j}}$.

c. The first component of $\vec{\mathbf{r}}(t)=e^t\sin t\,\hat{\mathbf{i}}+e^t\cos t\,\hat{\mathbf{j}}-e^{2t}\,\hat{\mathbf{k}}$ is $f(t)=e^t\sin t$, the second component is $g(t)=e^t\cos t$, and the third component is $h(t)=-e^{2t}$. We have $f'(t)=e^t(\sin t+\cos t)$, $g'(t)=e^t(\cos t-\sin t)$, and $h'(t)=-2e^{2t}$, so the theorem gives $\vec{\mathbf{r}}\,'(t)=e^t(\sin t+\cos t)\,\hat{\mathbf{i}}+e^t(\cos t-\sin t)\,\hat{\mathbf{j}}-2e^{2t}\,\hat{\mathbf{k}}$.

? Exercise 12.2.2

Calculate the derivative of the function

$$\vec{\mathbf{r}}(t)=(t\ln t)\,\hat{\mathbf{i}}+(5e^t)\,\hat{\mathbf{j}}+(\cos t-\sin t)\,\hat{\mathbf{k}}.$$

Hint

Identify the component functions and use Theorem 12.2.1

Answer

$$\vec{\mathbf{r}}\,'(t)=(1+\ln t)\,\hat{\mathbf{i}}+5e^t\,\hat{\mathbf{j}}-(\sin t+\cos t)\,\hat{\mathbf{k}}$$

We can extend to vector-valued functions the properties of the derivative that we presented previously. In particular, the constant multiple rule, the sum and difference rules, the product rule, and the chain rule all extend to vector-valued functions. However, in the case of the product rule, there are actually three extensions:

1. for a real-valued function multiplied by a vector-valued function,
2. for the dot product of two vector-valued functions, and
3. for the cross product of two vector-valued functions.

Theorem: Properties of the Derivative of Vector-Valued Functions

Let $\vec{\mathbf{r}}$ and $\vec{\mathbf{u}}$ be differentiable vector-valued functions of t, let f be a differentiable real-valued function of t, and let c be a scalar.

i.	$\frac{d}{dt}[c\vec{\mathbf{r}}(t)]$	$=$	$c\vec{\mathbf{r}}\,'(t)$	Scalar multiple
ii.	$\frac{d}{dt}[\vec{\mathbf{r}}(t)\pm\vec{\mathbf{u}}(t)]$	$=$	$\vec{\mathbf{r}}\,'(t)\pm\vec{\mathbf{u}}\,'(t)$	Sum and difference
iii.	$\frac{d}{dt}[f(t)\vec{\mathbf{u}}(t)]$	$=$	$f'(t)\vec{\mathbf{u}}(t)+f(t)\vec{\mathbf{u}}\,'(t)$	Scalar product
iv.	$\frac{d}{dt}[\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{u}}(t)]$	$=$	$\vec{\mathbf{r}}\,'(t)\cdot\vec{\mathbf{u}}(t)+\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{u}}\,'(t)$	Dot product
v.	$\frac{d}{dt}[\vec{\mathbf{r}}(t)\times\vec{\mathbf{u}}(t)]$	$=$	$\vec{\mathbf{r}}\,'(t)\times\vec{\mathbf{u}}(t)+\vec{\mathbf{r}}(t)\times\vec{\mathbf{u}}\,'(t)$	Cross product
vi.	$\frac{d}{dt}[\vec{\mathbf{r}}(f(t))]$	$=$	$\vec{\mathbf{r}}\,'(f(t))\cdot f'(t)$	Chain rule
vii.	If $\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{r}}(t)$	$=$	c, then $\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{r}}\,'(t)=0$.	

Proof

The proofs of the first two properties follow directly from the definition of the derivative of a vector-valued function. The third property can be derived from the first two properties, along with the product rule. Let $\vec{\mathbf{u}}(t)=g(t)\,\hat{\mathbf{i}}+h(t)\,\hat{\mathbf{j}}$. Then

$$\begin{aligned}\frac{d}{dt}[f(t)\vec{\mathbf{u}}(t)] &= \frac{d}{dt}[f(t)(g(t)\,\hat{\mathbf{i}} + h(t)\,\hat{\mathbf{j}})] \\ &= \frac{d}{dt}[f(t)g(t)\,\hat{\mathbf{i}} + f(t)h(t)\,\hat{\mathbf{j}}] \\ &= \frac{d}{dt}[f(t)g(t)]\,\hat{\mathbf{i}} + \frac{d}{dt}[f(t)h(t)]\,\hat{\mathbf{j}} \\ &= (f'(t)g(t) + f(t)g'(t))\,\hat{\mathbf{i}} + (f'(t)h(t) + f(t)h'(t))\,\hat{\mathbf{j}} \\ &= f'(t)\vec{\mathbf{u}}(t) + f(t)\vec{\mathbf{u}}'(t).\end{aligned}$$

To prove property iv. let $\vec{\mathbf{r}}(t) = f_1(t)\,\hat{\mathbf{i}} + g_1(t)\,\hat{\mathbf{j}}$ and $\vec{\mathbf{u}}(t) = f_2(t)\,\hat{\mathbf{i}} + g_2(t)\,\hat{\mathbf{j}}$. Then

$$\begin{aligned}\frac{d}{dt}[\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{u}}(t)] &= \frac{d}{dt}[f_1(t)f_2(t) + g_1(t)g_2(t)] \\ &= f_1'(t)f_2(t) + f_1(t)f_2'(t) + g_1'(t)g_2(t) + g_1(t)g_2'(t) = f_1'(t)f_2(t) + g_1'(t)g_2(t) + f_1(t)f_2'(t) + g_1(t)g_2'(t) \\ &= (f_1'\,\hat{\mathbf{i}} + g_1'\,\hat{\mathbf{j}})\cdot(f_2\,\hat{\mathbf{i}} + g_2\,\hat{\mathbf{j}}) + (f_1\,\hat{\mathbf{i}} + g_1\,\hat{\mathbf{j}})\cdot(f_2'\,\hat{\mathbf{i}} + g_2'\,\hat{\mathbf{j}}) \\ &= \vec{\mathbf{r}}'(t)\cdot\vec{\mathbf{u}}(t) + \vec{\mathbf{r}}(t)\cdot\vec{\mathbf{u}}'(t).\end{aligned}$$

The proof of property v. is similar to that of property iv. Property vi. can be proved using the chain rule. Last, property vii. follows from property iv:

$$\begin{aligned}\frac{d}{dt}[\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{r}}(t)] &= \frac{d}{dt}[c] \\ \vec{\mathbf{r}}'(t)\cdot\vec{\mathbf{r}}(t) + \vec{\mathbf{r}}(t)\cdot\vec{\mathbf{r}}'(t) &= 0 \\ 2\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{r}}'(t) &= 0 \\ \vec{\mathbf{r}}(t)\cdot\vec{\mathbf{r}}'(t) &= 0\end{aligned}$$

Now for some examples using these properties.

Example 12.2.3: Using the Properties of Derivatives of Vector-Valued Functions

Given the vector-valued functions

$$\vec{\mathbf{r}}(t) = (6t+8)\,\hat{\mathbf{i}} + (4t^2 + 2t - 3)\,\hat{\mathbf{j}} + 5t\,\hat{\mathbf{k}}$$

and

$$\vec{\mathbf{u}}(t) = (t^2 - 3)\,\hat{\mathbf{i}} + (2t+4)\,\hat{\mathbf{j}} + (t^3 - 3t)\,\hat{\mathbf{k}},$$

calculate each of the following derivatives using the properties of the derivative of vector-valued functions.

a. $\dfrac{d}{dt}[\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{u}}(t)]$

b. $\dfrac{d}{dt}[\vec{\mathbf{u}}(t)\times\vec{\mathbf{u}}'(t)]$

Solution

We have $\vec{\mathbf{r}}'(t) = 6\,\hat{\mathbf{i}} + (8t+2)\,\hat{\mathbf{j}} + 5\,\hat{\mathbf{k}}$ and $\vec{\mathbf{u}}'(t) = 2t\,\hat{\mathbf{i}} + 2\,\hat{\mathbf{j}} + (3t^2 - 3)\,\hat{\mathbf{k}}$. Therefore, according to **property iv**:

1.

$$\begin{aligned}\frac{d}{dt}[\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{u}}(t)] &= \vec{\mathbf{r}}\,'(t)\cdot\vec{\mathbf{u}}(t)+\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{u}}\,'(t)\\ &= (6\,\hat{\mathbf{i}}+(8t+2)\,\hat{\mathbf{j}}+5\,\hat{\mathbf{k}})\cdot((t^2-3)\,\hat{\mathbf{i}}+(2t+4)\,\hat{\mathbf{j}}+(t^3-3t)\,\hat{\mathbf{k}})\\ &\quad+((6t+8)\,\hat{\mathbf{i}}+(4t^2+2t-3)\,\hat{\mathbf{j}}+5t\,\hat{\mathbf{k}})\cdot(2t\,\hat{\mathbf{i}}+2\,\hat{\mathbf{j}}+(3t^2-3)\,\hat{\mathbf{k}})\\ &= 6(t^2-3)+(8t+2)(2t+4)+5(t^3-3t)\\ &\quad+2t(6t+8)+2(4t^2+2t-3)+5t(3t^2-3)\\ &= 20t^3+42t^2+26t-16.\end{aligned}$$

2. First, we need to adapt **property v** for this problem:

$$\frac{d}{dt}[\vec{\mathbf{u}}(t)\times\vec{\mathbf{u}}\,'(t)] = \vec{\mathbf{u}}\,'(t)\times\vec{\mathbf{u}}\,'(t)+\vec{\mathbf{u}}(t)\times\vec{\mathbf{u}}\,''(t).$$

Recall that the cross product of any vector with itself is zero. Furthermore, $\vec{\mathbf{u}}\,''(t)$ represents the second derivative of $\vec{\mathbf{u}}(t)$:

$$\vec{\mathbf{u}}\,''(t) = \frac{d}{dt}[\vec{\mathbf{u}}\,'(t)] = \frac{d}{dt}[2t\,\hat{\mathbf{i}}+2\,\hat{\mathbf{j}}+(3t^2-3)\,\hat{\mathbf{k}}] = 2\,\hat{\mathbf{i}}+6t\,\hat{\mathbf{k}}.$$

Therefore,

$$\begin{aligned}\frac{d}{dt}[\vec{\mathbf{u}}(t)\times\vec{\mathbf{u}}\,'(t)] &= 0+((t^2-3)\,\hat{\mathbf{i}}+(2t+4)\,\hat{\mathbf{j}}+(t^3-3t)\,\hat{\mathbf{k}})\times(2\,\hat{\mathbf{i}}+6t\,\hat{\mathbf{k}})\\ &= \begin{vmatrix}\hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}}\\ t^2-3 & 2t+4 & t^3-3t\\ 2 & 0 & 6t\end{vmatrix}\\ &= 6t(2t+4)\,\hat{\mathbf{i}}-(6t(t^2-3)-2(t^3-3t))\,\hat{\mathbf{j}}-2(2t+4)\,\hat{\mathbf{k}}\\ &= (12t^2+24t)\,\hat{\mathbf{i}}+(12t-4t^3)\,\hat{\mathbf{j}}-(4t+8)\,\hat{\mathbf{k}}.\end{aligned}$$

? Exercise 12.2.3

Calculate $\frac{d}{dt}[\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{r}}\,'(t)]$ and $\frac{d}{dt}[\vec{\mathbf{u}}(t)\times\vec{\mathbf{r}}(t)]$ for the vector-valued functions:

- $\vec{\mathbf{r}}(t) = \cos t\,\hat{\mathbf{i}}+\sin t\,\hat{\mathbf{j}}-e^{2t}\,\hat{\mathbf{k}}$
- $\vec{\mathbf{u}}(t) = t\,\hat{\mathbf{i}}+\sin t\,\hat{\mathbf{j}}+\cos t\,\hat{\mathbf{k}}$,

Hint

Follow the same steps as in Example 12.2.3

Answer

$$\frac{d}{dt}[\vec{\mathbf{r}}(t)\cdot\vec{\mathbf{r}}\,'(t)] = 8e^{4t}$$

$$\frac{d}{dt}[\vec{\mathbf{u}}(t)\times\vec{\mathbf{r}}(t)] = -(e^{2t}(\cos t+2\sin t)+\cos 2t)\,\hat{\mathbf{i}}+(e^{2t}(2t+1)-\sin 2t)\,\hat{\mathbf{j}}+(t\cos t+\sin t-\cos 2t)\,\hat{\mathbf{k}}$$

Tangent Vectors and Unit Tangent Vectors

Recall that the derivative at a point can be interpreted as the slope of the tangent line to the graph at that point. In the case of a vector-valued function, the derivative provides a tangent vector to the curve represented by the function. Consider the vector-valued function

$$\vec{\mathbf{r}}(t) = \cos t\,\hat{\mathbf{i}}+\sin t\,\hat{\mathbf{j}} \tag{12.2.2}$$

The derivative of this function is

$$\vec{\mathbf{r}}\,'(t) = -\sin t\,\hat{\mathbf{i}}+\cos t\,\hat{\mathbf{j}}$$

If we substitute the value $t = \pi/6$ into both functions we get

$$\vec{\mathbf{r}}\left(\frac{\pi}{6}\right) = \frac{\sqrt{3}}{2}\,\hat{\mathbf{i}} + \frac{1}{2}\,\hat{\mathbf{j}}$$

and

$$\vec{\mathbf{r}}\,'\left(\frac{\pi}{6}\right) = -\frac{1}{2}\,\hat{\mathbf{i}} + \frac{\sqrt{3}}{2}\,\hat{\mathbf{j}}.$$

The graph of this function appears in Figure 12.2.1, along with the vectors $\vec{\mathbf{r}}\left(\frac{\pi}{6}\right)$ and $\vec{\mathbf{r}}'\left(\frac{\pi}{6}\right)$.

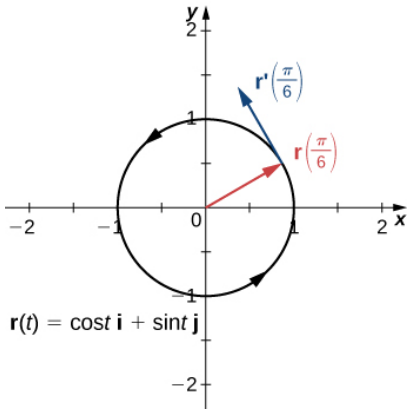


Figure 12.2.1: The tangent line at a point is calculated from the derivative of the vector-valued function $\vec{\mathbf{r}}(t)$.

Notice that the vector $\vec{\mathbf{r}}\,'\left(\frac{\pi}{6}\right)$ is tangent to the circle at the point corresponding to $t = \frac{\pi}{6}$. This is an example of a tangent vector to the plane curve defined by Equation 12.2.2.

Definition: principal unit tangent vector

Let C be a curve defined by a vector-valued function $\vec{\mathbf{r}}$, and assume that $\vec{\mathbf{r}}\,'(t)$ exists when $\mathrm{t} = \mathrm{t}_0$ A tangent vector $\vec{\mathbf{r}}$ at $t = t_0$ is any vector such that, when the tail of the vector is placed at point $\vec{\mathbf{r}}(t_0)$ on the graph, vector $\vec{\mathbf{r}}$ is tangent to curve C. Vector $\vec{\mathbf{r}}\,'(t_0)$ is an example of a tangent vector at point $t = t_0$. Furthermore, assume that $\vec{\mathbf{r}}\,'(t) \neq 0$. The principal unit tangent vector at t is defined to be

$$\vec{\mathbf{T}}(t) = \frac{\vec{\mathbf{r}}\,'(t)}{\|\vec{\mathbf{r}}\,'(t)\|},$$

provided $\|\vec{\mathbf{r}}\,'(t)\| \neq 0$.

The unit tangent vector is exactly what it sounds like: a unit vector that is tangent to the curve. To calculate a unit tangent vector, first find the derivative $\vec{\mathbf{r}}\,'(t)$. Second, calculate the magnitude of the derivative. The third step is to divide the derivative by its magnitude.

Example 12.2.4: Finding a Unit Tangent Vector

Find the unit tangent vector for each of the following vector-valued functions:

a. $\vec{\mathbf{r}}(t) = \cos t\,\hat{\mathbf{i}} + \sin t\,\hat{\mathbf{j}}$
b. $\vec{\mathbf{u}}(t) = (3t^2 + 2t)\,\hat{\mathbf{i}} + (2 - 4t^3)\,\hat{\mathbf{j}} + (6t + 5)\,\hat{\mathbf{k}}$

Solution

a.

$$\begin{aligned}
\text{First step:} \quad \vec{\mathbf{r}}\,'(t) &= -\sin t\,\hat{\mathbf{i}} + \cos t\,\hat{\mathbf{j}} \\
\text{Second step:} \quad \|\vec{\mathbf{r}}\,'(t)\| &= \sqrt{(-\sin t)^2 + (\cos t)^2} = 1 \\
\text{Third step:} \quad \vec{\mathbf{T}}(t) &= \frac{\vec{\mathbf{r}}\,'(t)}{\|\vec{\mathbf{r}}\,'(t)\|} = \frac{-\sin t\,\hat{\mathbf{i}} + \cos t\,\hat{\mathbf{j}}}{1} = -\sin t\,\hat{\mathbf{i}} + \cos t\,\hat{\mathbf{j}}
\end{aligned}$$

b.

$$\begin{aligned}
\text{First step:} \quad \vec{\mathbf{r}}\,'(t) &= (6t+2)\,\hat{\mathbf{i}} - 12t^2\,\hat{\mathbf{j}} + 6\,\hat{\mathbf{k}} \\
\text{Second step:} \quad \|\vec{\mathbf{r}}\,'(t)\| &= \sqrt{(6t+2)^2 + (-12t^2)^2 + 6^2} \\
&= \sqrt{144t^4 + 36t^2 + 24t + 40} \\
&= 2\sqrt{36t^4 + 9t^2 + 6t + 10} \\
\text{Third step:} \quad \vec{\mathbf{T}}(t) &= \frac{\vec{\mathbf{r}}\,'(t)}{\|\vec{\mathbf{r}}\,'(t)\|} = \frac{(6t+2)\,\hat{\mathbf{i}} - 12t^2\,\hat{\mathbf{j}} + 6\,\hat{\mathbf{k}}}{2\sqrt{36t^4 + 9t^2 + 6t + 10}} \\
&= \frac{3t+1}{\sqrt{36t^4 + 9t^2 + 6t + 10}}\,\hat{\mathbf{i}} - \frac{6t^2}{\sqrt{36t^4 + 9t^2 + 6t + 10}}\,\hat{\mathbf{j}} + \frac{3}{\sqrt{36t^4 + 9t^2 + 6t + 10}}\,\hat{\mathbf{k}}
\end{aligned}$$

Exercise 12.2.4

Find the unit tangent vector for the vector-valued function

$$\vec{\mathbf{r}}(t) = (t^2 - 3)\,\hat{\mathbf{i}} + (2t+1)\,\hat{\mathbf{j}} + (t-2)\,\hat{\mathbf{k}}.$$

Hint

Follow the same steps as in Example 12.2.4

Answer

$$\vec{\mathbf{T}}(t) = \frac{2t}{\sqrt{4t^2+5}}\,\hat{\mathbf{i}} + \frac{2}{\sqrt{4t^2+5}}\,\hat{\mathbf{j}} + \frac{1}{\sqrt{4t^2+5}}\,\hat{\mathbf{k}}$$

Integrals of Vector-Valued Functions

We introduced antiderivatives of real-valued functions in Antiderivatives and definite integrals of real-valued functions in The Definite Integral. Each of these concepts can be extended to vector-valued functions. Also, just as we can calculate the derivative of a vector-valued function by differentiating the component functions separately, we can calculate the antiderivative in the same manner. Furthermore, the Fundamental Theorem of Calculus applies to vector-valued functions as well.

The antiderivative of a vector-valued function appears in applications. For example, if a vector-valued function represents the velocity of an object at time t, then its antiderivative represents position. Or, if the function represents the acceleration of the object at a given time, then the antiderivative represents its velocity.

Definition: Definite and Indefinite Integrals of Vector-Valued Functions

Let f, g, and h be integrable real-valued functions over the closed interval $[a, b]$.

1. The indefinite integral of a vector-valued function $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}}$ is

$$\int [f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}}]\,dt = \left[\int f(t)\,dt\right]\,\hat{\mathbf{i}} + \left[\int g(t)\,dt\right]\,\hat{\mathbf{j}}.$$

The definite integral of a vector-valued function is

$$\int_a^b [f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}}]\,dt = \left[\int_a^b f(t)\,dt\right]\,\hat{\mathbf{i}} + \left[\int_a^b g(t)\,dt\right]\,\hat{\mathbf{j}}.$$

2. The indefinite integral of a vector-valued function $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\,\hat{\mathbf{k}}$ is

$$\int [f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\,\hat{\mathbf{k}}]\,dt = \left[\int f(t)\,dt\right]\,\hat{\mathbf{i}} + \left[\int g(t)\,dt\right]\,\hat{\mathbf{j}} + \left[\int h(t)\,dt\right]\,\hat{\mathbf{k}}.$$

The definite integral of the vector-valued function is

$$\int_a^b [f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\,\hat{\mathbf{k}}]\,dt = \left[\int_a^b f(t)\,dt\right]\,\hat{\mathbf{i}} + \left[\int_a^b g(t)\,dt\right]\,\hat{\mathbf{j}} + \left[\int_a^b h(t)\,dt\right]\,\hat{\mathbf{k}}.$$

Since the indefinite integral of a vector-valued function involves indefinite integrals of the component functions, each of these component integrals contains an integration constant. They can all be different. For example, in the two-dimensional case, we can have

$$\int f(t)\,dt = F(t) + C_1 \text{ and } \int g(t)\,dt = G(t) + C_2,$$

where F and G are antiderivatives of f and g, respectively. Then

$$\begin{aligned}\int [f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}}]\,dt &= \left[\int f(t)\,dt\right]\hat{\mathbf{i}} + \left[\int g(t)\,dt\right]\hat{\mathbf{j}} \\ &= (F(t)+C_1)\,\hat{\mathbf{i}} + (G(t)+C_2)\,\hat{\mathbf{j}} \\ &= F(t)\,\hat{\mathbf{i}} + G(t)\,\hat{\mathbf{j}} + C_1\,\hat{\mathbf{i}} + C_2\,\hat{\mathbf{j}} \\ &= F(t)\,\hat{\mathbf{i}} + G(t)\,\hat{\mathbf{j}} + \vec{\mathbf{C}}\end{aligned}$$

where $\vec{\mathbf{C}} = C_1\,\hat{\mathbf{i}} + C_2\,\hat{\mathbf{j}}$. Therefore, the *integration constants* becomes a *constant vector*.

Example 12.2.5: Integrating Vector-Valued Functions

Calculate each of the following integrals:

a. $\displaystyle\int [(3t^2+2t)\,\hat{\mathbf{i}} + (3t-6)\,\hat{\mathbf{j}} + (6t^3+5t^2-4)\,\hat{\mathbf{k}}]\,dt$

b. $\displaystyle\int [\langle t, t^2, t^3\rangle \times \langle t^3, t^2, t\rangle]\,dt$

c. $\displaystyle\int_0^{\frac{\pi}{3}} [\sin 2t\,\hat{\mathbf{i}} + \tan t\,\hat{\mathbf{j}} + e^{-2t}\,\hat{\mathbf{k}}]\,dt$

Solution

a. We use the first part of the definition of the integral of a space curve:

b. $$\begin{aligned}\int [(3t^2+2t)\,\hat{\mathbf{i}} + (3t-6)\,\hat{\mathbf{j}} + (6t^3+5t^2-4)\,\hat{\mathbf{k}}]\,dt &= \left[\int 3t^2+2t\,dt\right]\hat{\mathbf{i}} + \left[\int 3t-6\,dt\right]\hat{\mathbf{j}} + \left[\int 6t^3+5t^2-4\,dt\right]\hat{\mathbf{k}} \\ &= (t^3+t^2)\,\hat{\mathbf{i}} + \left(\frac{3}{2}t^2 - 6t\right)\hat{\mathbf{j}} + \left(\frac{3}{2}t^4 + \frac{5}{3}t^3 - 4t\right)\hat{\mathbf{k}} + \vec{\mathbf{C}}.\end{aligned}$$

c. First calculate $\langle t, t^2, t^3\rangle \times \langle t^3, t^2, t\rangle$:

$$\begin{aligned}\langle t, t^2, t^3\rangle \times \langle t^3, t^2, t\rangle &= \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ t & t^2 & t^3 \\ t^3 & t^2 & t \end{vmatrix} \\ &= (t^2(t) - t^3(t^2))\,\hat{\mathbf{i}} - (t^2 - t^3(t^3))\,\hat{\mathbf{j}} + (t(t^2) - t^2(t^3))\,\hat{\mathbf{k}} \\ &= (t^3 - t^5)\,\hat{\mathbf{i}} + (t^6 - t^2)\,\hat{\mathbf{j}} + (t^3 - t^5)\,\hat{\mathbf{k}}.\end{aligned}$$

Next, substitute this back into the integral and integrate:

$$\begin{aligned}\int [\langle t, t^2, t^3\rangle \times \langle t^3, t^2, t\rangle]\,dt &= \int (t^3 - t^5)\,\hat{\mathbf{i}} + (t^6 - t^2)\,\hat{\mathbf{j}} + (t^3 - t^5)\,\hat{\mathbf{k}}\,dt \\ &= \left(\frac{t^4}{4} - \frac{t^6}{6}\right)\hat{\mathbf{i}} + \left(\frac{t^7}{7} - \frac{t^3}{3}\right)\hat{\mathbf{j}} + \left(\frac{t^4}{4} - \frac{t^6}{6}\right)\hat{\mathbf{k}} + \vec{\mathbf{C}}.\end{aligned}$$

d. Use the second part of the definition of the integral of a space curve:

$$\begin{aligned}\int_0^{\frac{\pi}{3}} [\sin 2t\,\hat{\mathbf{i}} + \tan t\,\hat{\mathbf{j}} + e^{-2t}\,\hat{\mathbf{k}}]\,dt &= \left[\int_0^{\frac{\pi}{3}} \sin 2t\,dt\right]\hat{\mathbf{i}} + \left[\int_0^{\frac{\pi}{3}} \tan t\,dt\right]\hat{\mathbf{j}} + \left[\int_0^{\frac{\pi}{3}} e^{-2t}\,dt\right]\hat{\mathbf{k}} \\ &= \left(-\tfrac{1}{2}\cos 2t\right)\Big|_0^{\pi/3}\hat{\mathbf{i}} - \left(\ln|\cos t|\right)\Big|_0^{\pi/3}\hat{\mathbf{j}} - \left(\tfrac{1}{2}e^{-2t}\right)\Big|_0^{\pi/3}\hat{\mathbf{k}} \\ &= \left(-\tfrac{1}{2}\cos\tfrac{2\pi}{3} + \tfrac{1}{2}\cos 0\right)\hat{\mathbf{i}} - \left(\ln\left(\cos\tfrac{\pi}{3}\right) - \ln(\cos 0)\right)\hat{\mathbf{j}} - \left(\tfrac{1}{2}e^{-2\pi/3} - \tfrac{1}{2}e^{-2(0)}\right)\hat{\mathbf{k}} \\ &= \left(\tfrac{1}{4} + \tfrac{1}{2}\right)\hat{\mathbf{i}} - (-\ln 2)\,\hat{\mathbf{j}} - \left(\tfrac{1}{2}e^{-2\pi/3} - \tfrac{1}{2}\right)\hat{\mathbf{k}} \\ &= \tfrac{3}{4}\hat{\mathbf{i}} + (\ln 2)\,\hat{\mathbf{j}} + \left(\tfrac{1}{2} - \tfrac{1}{2}e^{-2\pi/3}\right)\hat{\mathbf{k}}.\end{aligned}$$

? Exercise 12.2.5

Calculate the following integral:

$$\int_1^3 [(2t+4)\,\hat{\mathbf{i}} + (3t^2 - 4t)\,\hat{\mathbf{j}}]\,dt$$

Hint

Use the definition of the definite integral of a plane curve.

Answer

$$\int_1^3 [(2t+4)\,\hat{\mathbf{i}} + (3t^2 - 4t)\,\hat{\mathbf{j}}]\,dt = 16\,\hat{\mathbf{i}} + 10\,\hat{\mathbf{j}}$$

Summary

- To calculate the derivative of a vector-valued function, calculate the derivatives of the component functions, then put them back into a new vector-valued function.
- Many of the properties of differentiation of scalar functions also apply to vector-valued functions.
- The derivative of a vector-valued function $\vec{\mathbf{r}}(t)$ is also a tangent vector to the curve. The unit tangent vector $\vec{\mathbf{T}}(t)$ is calculated by dividing the derivative of a vector-valued function by its magnitude.
- The antiderivative of a vector-valued function is found by finding the antiderivatives of the component functions, then putting them back together in a vector-valued function.
- The definite integral of a vector-valued function is found by finding the definite integrals of the component functions, then putting them back together in a vector-valued function.

Key Equations

- **Derivative of a vector-valued function**

$$\vec{\mathbf{r}}\,'(t) = \lim_{\Delta t \to 0} \frac{\vec{\mathbf{r}}(t+\Delta t) - \vec{\mathbf{r}}(t)}{\Delta t}$$

- **Principal unit tangent vector**

$$\vec{\mathbf{T}}(t) = \frac{\vec{\mathbf{r}}\,'(t)}{\|\vec{\mathbf{r}}\,'(t)\|}$$

- **Indefinite integral of a vector-valued function**

$$\int [f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\,\hat{\mathbf{k}}]\,dt = \left[\int f(t)\,dt\right]\hat{\mathbf{i}} + \left[\int g(t)\,dt\right]\hat{\mathbf{j}} + \left[\int h(t)\,dt\right]\hat{\mathbf{k}}$$

- **Definite integral of a vector-valued function**

$$\int_a^b [f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\,\hat{\mathbf{k}}]\,dt = \left[\int_a^b f(t)\,dt\right]\hat{\mathbf{i}} + \left[\int_a^b g(t)\,dt\right]\hat{\mathbf{j}} + \left[\int_a^b h(t)\,dt\right]\hat{\mathbf{k}}$$

Glossary

definite integral of a vector-valued function

the vector obtained by calculating the definite integral of each of the component functions of a given vector-valued function, then using the results as the components of the resulting function

derivative of a vector-valued function

the derivative of a vector-valued function $\vec{\mathbf{r}}(t)$ is $\vec{\mathbf{r}}\,'(t) = \lim_{\Delta t \to 0} \frac{\vec{\mathbf{r}}(t+\Delta t) - \vec{\mathbf{r}}(t)}{\Delta t}$, provided the limit exists

indefinite integral of a vector-valued function

a vector-valued function with a derivative that is equal to a given vector-valued function

principal unit tangent vector

a unit vector tangent to a curve C

tangent vector

to $\vec{\mathbf{r}}(t)$ at $t = t_0$ any vector $\vec{\mathbf{v}}$ such that, when the tail of the vector is placed at point $\vec{\mathbf{r}}(t_0)$ on the graph, vector $\vec{\mathbf{v}}$ is tangent to curve C

12.2: Integrals of Vector Functions; Projectile Motion is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **13.2: Calculus of Vector-Valued Functions** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

12.3: Arc Length in Space

Learning Objectives

- Determine the length of a particle's path in space by using the arc-length function.
- Explain the meaning of the curvature of a curve in space and state its formula.
- Describe the meaning of the normal and binormal vectors of a curve in space.

In this section, we study formulas related to curves in both two and three dimensions, and see how they are related to various properties of the same curve. For example, suppose a vector-valued function describes the motion of a particle in space. We would like to determine how far the particle has traveled over a given time interval, which can be described by the arc length of the path it follows. Or, suppose that the vector-valued function describes a road we are building and we want to determine how sharply the road curves at a given point. This is described by the curvature of the function at that point. We explore each of these concepts in this section.

Arc Length for Vector Functions

We have seen how a vector-valued function describes a curve in either two or three dimensions. Recall that the formula for the arc length of a curve defined by the parametric functions $x = x(t), y = y(t), t_1 \leq t \leq t_2$ is given by

$$s = \int_{t_1}^{t_2} \sqrt{(x'(t))^2 + (y'(t))^2}\, dt.$$

In a similar fashion, if we define a smooth curve using a vector-valued function $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}}$, where $a \leq t \leq b$, the arc length is given by the formula

$$s = \int_a^b \sqrt{(f'(t))^2 + (g'(t))^2}\, dt.$$

In three dimensions, if the vector-valued function is described by $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\,\hat{\mathbf{k}}$ over the same interval $a \leq t \leq b$, the arc length is given by

$$s = \int_a^b \sqrt{(f'(t))^2 + (g'(t))^2 + (h'(t))^2}\, dt.$$

Theorem: Arc-Length Formulas for Plane and Space curves

Plane curve: Given a smooth curve C defined by the function $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}}$, where t lies within the interval $[a, b]$, the arc length of C over the interval is

$$s = \int_a^b \sqrt{[f'(t)]^2 + [g'(t)]^2}\, dt \tag{12.3.1}$$

$$= \int_a^b \|\vec{\mathbf{r}}\,'(t)\|\, dt. \tag{12.3.2}$$

Space curve: Given a smooth curve C defined by the function $\vec{\mathbf{r}}(t) = f(t)\,\hat{\mathbf{i}} + g(t)\,\hat{\mathbf{j}} + h(t)\,\hat{\mathbf{k}}$, where t lies within the interval $[a, b]$, the arc length of C over the interval is

$$s = \int_a^b \sqrt{[f'(t)]^2 + [g'(t)]^2 + [h'(t)]^2}\, dt \tag{12.3.3}$$

$$= \int_a^b \|\vec{\mathbf{r}}\,'(t)\|\, dt. \tag{12.3.4}$$

The two formulas are very similar; they differ only in the fact that a space curve has three component functions instead of two. Note that the formulas are defined for smooth curves: curves where the vector-valued function $\vec{\mathbf{r}}(t)$ is differentiable with a non-

zero derivative. The smoothness condition guarantees that the curve has no cusps (or corners) that could make the formula problematic.

Example 12.3.1: Finding the Arc Length

Calculate the arc length for each of the following vector-valued functions:

a. $\vec{\mathbf{r}}(t) = (3t-2)\,\hat{\mathbf{i}} + (4t+5)\,\hat{\mathbf{j}}, \quad 1 \le t \le 5$
b. $\vec{\mathbf{r}}(t) = \langle t\cos t, t\sin t, 2t\rangle, 0 \le t \le 2\pi$

Solution

a. Using Equation 12.3.2, $\vec{\mathbf{r}}\,'(t) = 3\,\hat{\mathbf{i}} + 4\,\hat{\mathbf{j}}$, so

$$\begin{aligned} s &= \int_a^b \|\vec{\mathbf{r}}\,'(t)\|dt \\ &= \int_1^5 \sqrt{3^2+4^2}\,dt \\ &= \int_1^5 5dt = 5t\big|_1^5 = 20. \end{aligned}$$

b. Using Equation 12.3.4, $\vec{\mathbf{r}}\,'(t) = \langle \cos t - t\sin t, \sin t + t\cos t, 2\rangle$, so

$$\begin{aligned} s &= \int_a^b \| \vec{\mathbf{r}}\,'(t) \| \, dt \\ &= \int_0^{2\pi} \sqrt{(\cos t - t\sin t)^2 + (\sin t + t\cos t)^2 + 2^2}\; dt \\ &= \int_0^{2\pi} \sqrt{(\cos^2 t - 2t\sin t\cos t + t^2\sin^2 t) + (\sin^2 t + 2t\sin t\cos t + t^2\cos^2 t) + 4}\; dt \\ &= \int_0^{2\pi} \sqrt{\cos^2 t + \sin^2 t + t^2(\cos^2 t + \sin^2 t) + 4}\; dt \\ &= \int_0^{2\pi} \sqrt{t^2+5}\,dt \end{aligned}$$

Here we can use a table integration formula

$$\int \sqrt{u^2+a^2}\,du = \frac{u}{2}\sqrt{u^2+a^2} + \frac{a^2}{2}\ln\left|u+\sqrt{u^2+a^2}\right| + C,$$

so we obtain

$$\begin{aligned} \int_0^{2\pi} \sqrt{t^2+5}\,dt &= \frac{1}{2}\left(t\sqrt{t^2+5} + 5\ln\left|t+\sqrt{t^2+5}\right|\right)_0^{2\pi} \\ &= \frac{1}{2}\left(2\pi\sqrt{4\pi^2+5} + 5\ln\left(2\pi+\sqrt{4\pi^2+5}\right)\right) - \frac{5}{2}\ln\sqrt{5} \\ &\approx 25.343 \text{ units.} \end{aligned}$$

Exercise 12.3.1

Calculate the arc length of the parameterized curve

$$\vec{\mathbf{r}}(t) = \langle 2t^2+1, 2t^2-1, t^3\rangle, \quad 0 \le t \le 3.$$

Hint

Use Equation 12.3.4

Answer

$\vec{\mathbf{r}}\,'(t) = \langle 4t, 4t, 3t^2 \rangle$, so $s = \frac{1}{27}(113^{3/2} - 32^{3/2}) \approx 37.785$ units

We now return to the helix introduced earlier in this chapter. A vector-valued function that describes a helix can be written in the form

$$\vec{\mathbf{r}}(t) = R\cos\left(\frac{2\pi Nt}{h}\right)\hat{\mathbf{i}} + R\sin\left(\frac{2\pi Nt}{h}\right)\hat{\mathbf{j}} + t\,\hat{\mathbf{k}}, 0 \le t \le h,$$

where R represents the radius of the helix, h represents the height (distance between two consecutive turns), and the helix completes N turns. Let's derive a formula for the arc length of this helix using Equation 12.3.4. First of all,

$$\vec{\mathbf{r}}\,'(t) = -\frac{2\pi NR}{h}\sin\left(\frac{2\pi Nt}{h}\right)\hat{\mathbf{i}} + \frac{2\pi NR}{h}\cos\left(\frac{2\pi Nt}{h}\right)\hat{\mathbf{j}} + \hat{\mathbf{k}}.$$

Therefore,

$$\begin{aligned}
s &= \int_a^b \|\vec{\mathbf{r}}\,'(t)\|dt \\
&= \int_0^h \sqrt{\left(-\frac{2\pi NR}{h}\sin\left(\frac{2\pi Nt}{h}\right)\right)^2 + \left(\frac{2\pi NR}{h}\cos\left(\frac{2\pi Nt}{h}\right)\right)^2 + 1^2}\,dt \\
&= \int_0^h \sqrt{\frac{4\pi^2N^2R^2}{h^2}\left(\sin^2\left(\frac{2\pi Nt}{h}\right) + \cos^2\left(\frac{2\pi Nt}{h}\right)\right) + 1}\,dt \\
&= \int_0^h \sqrt{\frac{4\pi^2N^2R^2}{h^2} + 1}\,dt \\
&= \left[t\sqrt{\frac{4\pi^2N^2R^2}{h^2} + 1}\right]_0^h \\
&= h\sqrt{\frac{4\pi^2N^2R^2 + h^2}{h^2}} \\
&= \sqrt{4\pi^2N^2R^2 + h^2}.
\end{aligned}$$

This gives a formula for the length of a wire needed to form a helix with N turns that has radius R and height h.

Arc-Length Parameterization

We now have a formula for the arc length of a curve defined by a vector-valued function. Let's take this one step further and examine what an **arc-length function** is.

If a vector-valued function represents the position of a particle in space as a function of time, then the arc-length function measures how far that particle travels as a function of time. The formula for the arc-length function follows directly from the formula for arc length:

$$s = \int_a^t \sqrt{(f'(u))^2 + (g'(u))^2 + (h'(u))^2}\,du. \tag{12.3.5}$$

If the curve is in two dimensions, then only two terms appear under the square root inside the integral. The reason for using the independent variable u is to distinguish between time and the variable of integration. Since $s(t)$ measures distance traveled as a function of time, $s'(t)$ measures the speed of the particle at any given time. Since we have a formula for $s(t)$ in Equation 12.3.5, we can differentiate both sides of the equation:

$$
\begin{aligned}
s'(t) &= \frac{d}{dt}\left[\int_a^t \sqrt{(f'(u))^2+(g'(u))^2+(h'(u))^2}\,du\right] \\
&= \frac{d}{dt}\left[\int_a^t \|\vec{\mathbf{r}}\,'(u)\|du\right] \\
&= \|\vec{\mathbf{r}}\,'(t)\|.
\end{aligned}
$$

If we assume that $\vec{\mathbf{r}}(t)$ defines a smooth curve, then the arc length is always increasing, so $s'(t) > 0$ for $t > a$. Last, if $\vec{\mathbf{r}}(t)$ is a curve on which $\|\vec{\mathbf{r}}\,'(t)\| = 1$ for all t, then

$$
s(t) = \int_a^t \|\vec{\mathbf{r}}\,'(u)\|\,du = \int_a^t 1\,du = t - a,
$$

which means that t represents the arc length as long as $a = 0$.

Theorem: Arc-Length Function

Let $\vec{\mathbf{r}}(t)$ describe a smooth curve for $t \geq a$. Then the arc-length function is given by

$$
s(t) = \int_a^t \|\vec{\mathbf{r}}\,'(u)\|\,du
$$

Furthermore,

$$
\frac{ds}{dt} = \|\vec{\mathbf{r}}\,'(t)\| > 0.
$$

If $\|\vec{\mathbf{r}}\,'(t)\| = 1$ for all $t \geq a$, then the parameter t represents the arc length from the starting point at $t = a$.

A useful application of this theorem is to find an alternative parameterization of a given curve, called an **arc-length parameterization**. Recall that any vector-valued function can be reparameterized via a change of variables. For example, if we have a function $\vec{\mathbf{r}}(t) = \langle 3\cos t, 3\sin t\rangle, 0 \leq t \leq 2\pi$ that parameterizes a circle of radius 3, we can change the parameter from t to $4t$, obtaining a new parameterization $\vec{\mathbf{r}}(t) = \langle 3\cos 4t, 3\sin 4t\rangle$. The new parameterization still defines a circle of radius 3, but now we need only use the values $0 \leq t \leq \pi/2$ to traverse the circle once.

Suppose that we find the arc-length function $s(t)$ and are able to solve this function for t as a function of s. We can then reparameterize the original function $\vec{\mathbf{r}}(t)$ by substituting the expression for t back into $\vec{\mathbf{r}}(t)$. The vector-valued function is now written in terms of the parameter s. Since the variable s represents the arc length, we call this an *arc-length parameterization* of the original function $\vec{\mathbf{r}}(t)$. One advantage of finding the arc-length parameterization is that the distance traveled along the curve starting from $s = 0$ is now equal to the parameter s. The arc-length parameterization also appears in the context of curvature (which we examine later in this section) and line integrals.

Example 12.3.2: Finding an Arc-Length Parameterization

Find the arc-length parameterization for each of the following curves:

a. $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} + 4\sin t\,\hat{\mathbf{j}}, \quad t \geq 0$
b. $\vec{\mathbf{r}}(t) = \langle t+3, 2t-4, 2t\rangle, \quad t \geq 3$

Solution

a. First we find the arc-length function using Equation 12.3.5:

$$\begin{aligned} s(t) &= \int_a^t \|\vec{\mathbf{r}}\,'(u)\|\,du \\ &= \int_0^t \|\langle -4\sin u, 4\cos u\rangle\|\,du \\ &= \int_0^t \sqrt{(-4\sin u)^2 + (4\cos u)^2}\,du \\ &= \int_0^t \sqrt{16\sin^2 u + 16\cos^2 u}\,du \\ &= \int_0^t 4\,du = 4t, \end{aligned}$$

which gives the relationship between the arc length s and the parameter t as $s = 4t$; so, $t = s/4$. Next we replace the variable t in the original function $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} + 4\sin t\,\hat{\mathbf{j}}$ with the expression $s/4$ to obtain

$$\vec{\mathbf{r}}(s) = 4\cos\left(\frac{s}{4}\right)\hat{\mathbf{i}} + 4\sin\left(\frac{s}{4}\right)\hat{\mathbf{j}}.$$

This is the arc-length parameterization of $\vec{\mathbf{r}}(t)$. Since the original restriction on t was given by $t \geq 0$, the restriction on s becomes $s/4 \geq 0$, or $s \geq 0$.

b. The arc-length function is given by Equation 12.3.5:

$$\begin{aligned} s(t) &= \int_a^t \|\vec{\mathbf{r}}\,'(u)\|\,du \\ &= \int_3^t \|\langle 1, 2, 2\rangle\|\,du \\ &= \int_3^t \sqrt{1^2 + 2^2 + 2^2}\,du \\ &= \int_3^t 3\,du \\ &= 3t - 9. \end{aligned}$$

Therefore, the relationship between the arc length s and the parameter t is $s = 3t - 9$, so $t = \frac{s}{3} + 3$. Substituting this into the original function $\vec{\mathbf{r}}(t) = \langle t + 3, 2t - 4, 2t\rangle$ yields

$$\vec{\mathbf{r}}(s) = \langle\left(\frac{s}{3} + 3\right) + 3,\ 2\left(\frac{s}{3} + 3\right) - 4,\ 2\left(\frac{s}{3} + 3\right)\rangle = \langle\frac{s}{3} + 6, \frac{2s}{3} + 2, \frac{2s}{3} + 6\rangle.$$

This is an arc-length parameterization of $\vec{\mathbf{r}}(t)$. The original restriction on the parameter t was $t \geq 3$, so the restriction on s is $(s/3) + 3 \geq 3$, or $s \geq 0$.

? Exercise 12.3.2

Find the arc-length function for the helix

$$\vec{\mathbf{r}}(t) = \langle 3\cos t, 3\sin t, 4t\rangle, \quad t \geq 0.$$

Then, use the relationship between the arc length and the parameter t to find an arc-length parameterization of $\vec{\mathbf{r}}(t)$.

Hint

Start by finding the arc-length function.

Answer

$s = 5t$, or $t = s/5$. Substituting this into $\vec{\mathbf{r}}(t) = \langle 3\cos t, 3\sin t, 4t\rangle$ gives

$$\vec{\mathbf{r}}(s) = \langle 3\cos\left(\frac{s}{5}\right), 3\sin\left(\frac{s}{5}\right), \frac{4s}{5}\rangle, \quad s \geq 0$$

Curvature

An important topic related to arc length is curvature. The concept of curvature provides a way to measure how sharply a smooth curve turns. A circle has constant curvature. The smaller the radius of the circle, the greater the curvature.

Think of driving down a road. Suppose the road lies on an arc of a large circle. In this case you would barely have to turn the wheel to stay on the road. Now suppose the radius is smaller. In this case you would need to turn more sharply to stay on the road. In the case of a curve other than a circle, it is often useful first to inscribe a circle to the curve at a given point so that it is tangent to the curve at that point and "hugs" the curve as closely as possible in a neighborhood of the point (Figure 12.3.1). The curvature of the graph at that point is then defined to be the same as the curvature of the inscribed circle.

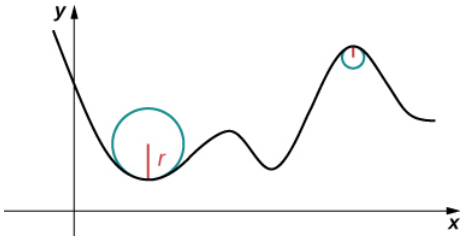


Figure 12.3.1: The graph represents the curvature of a function $y = f(x)$. The sharper the turn in the graph, the greater the curvature, and the smaller the radius of the inscribed circle.

Definition: Curvature

Let C be a smooth curve in the plane or in space given by $\vec{\mathbf{r}}(s)$, where s is the arc-length parameter. The curvature κ at s is

$$\kappa = \left\| \frac{d\vec{\mathbf{T}}}{ds} \right\| = \|\vec{\mathbf{T}}\,'(s)\|.$$

Visit this video for more information about the curvature of a space curve.

The formula in the definition of curvature is not very useful in terms of calculation. In particular, recall that $\vec{\mathbf{T}}(t)$ represents the unit tangent vector to a given vector-valued function $\vec{\mathbf{r}}(t)$, and the formula for $\vec{\mathbf{T}}(t)$ is

$$\vec{\mathbf{T}}(t) = \frac{\vec{\mathbf{r}}\,'(t)}{\|\vec{\mathbf{r}}\,'(t)\|}.$$

To use the formula for curvature, it is first necessary to express $\vec{\mathbf{r}}(t)$ in terms of the arc-length parameter s, then find the unit tangent vector $\vec{\mathbf{T}}(s)$ for the function $\vec{\mathbf{r}}(s)$, then take the derivative of $\vec{\mathbf{T}}(s)$ with respect to s. This is a tedious process. Fortunately, there are equivalent formulas for curvature.

Theorem: Alternate Formulas of Curvature

If C is a smooth curve given by $\vec{\mathbf{r}}(t)$, then the curvature κ of C at t is given by

$$\kappa = \frac{\|\vec{\mathbf{T}}\,'(t)\|}{\|\vec{\mathbf{r}}\,'(t)\|}. \tag{12.3.6}$$

If C is a three-dimensional curve, then the curvature can be given by the formula

$$\kappa = \frac{\|\vec{\mathbf{r}}\,'(t) \times \vec{\mathbf{r}}\,''(t)\|}{\|\vec{\mathbf{r}}\,'(t)\|^3}. \tag{12.3.7}$$

If C is the graph of a function $y = f(x)$ and both y' and y'' exist, then the curvature κ at point (x, y) is given by

$$\kappa = \frac{|y''|}{[1+(y')^2]^{3/2}}. \tag{12.3.8}$$

Proof

The first formula follows directly from the chain rule:

$$\frac{d\vec{\mathbf{T}}}{dt} = \frac{d\vec{\mathbf{T}}}{ds}\frac{ds}{dt},$$

where s is the arc length along the curve C. Dividing both sides by ds/dt, and taking the magnitude of both sides gives

$$\left\| \frac{d\vec{\mathbf{T}}}{ds} \right\| = \left\| \frac{\vec{\mathbf{T}}\,'(t)}{\frac{ds}{dt}} \right\|.$$

Since $ds/dt = \|\vec{\mathbf{r}}\,'(t)\|$, this gives the formula for the curvature κ of a curve C in terms of any parameterization of C:

$$\kappa = \frac{\|\vec{\mathbf{T}}\,'(t)\|}{\|\vec{\mathbf{r}}\,'(t)\|}.$$

In the case of a three-dimensional curve, we start with the formulas $\vec{\mathbf{T}}(t) = (\vec{\mathbf{r}}\,'(t))/\|\vec{\mathbf{r}}\,'(t)\|$ and $ds/dt = \|\vec{\mathbf{r}}\,'(t)\|$. Therefore, $\vec{\mathbf{r}}\,'(t) = (ds/dt)\vec{\mathbf{T}}(t)$. We can take the derivative of this function using the scalar product formula:

$$\vec{\mathbf{r}}\,''(t) = \frac{d^2s}{dt^2}\vec{\mathbf{T}}(t) + \frac{ds}{dt}\vec{\mathbf{T}}\,'(t).$$

Using these last two equations we get

$$\begin{aligned} \vec{\mathbf{r}}\,'(t) \times \vec{\mathbf{r}}\,''(t) &= \frac{ds}{dt}\vec{\mathbf{T}}(t) \times \left(\frac{d^2s}{dt^2}\vec{\mathbf{T}}(t) + \frac{ds}{dt}\vec{\mathbf{T}}\,'(t) \right) \\ &= \frac{ds}{dt}\frac{d^2s}{dt^2}\vec{\mathbf{T}}(t) \times \vec{\mathbf{T}}(t) + (\frac{ds}{dt})^2 \vec{\mathbf{T}}(t) \times \vec{\mathbf{T}}\,'(t). \end{aligned}$$

Since $\vec{\mathbf{T}}(t) \times \vec{\mathbf{T}}(t) = 0$, this reduces to

$$\vec{\mathbf{r}}\,'(t) \times \vec{\mathbf{r}}\,''(t) = \left(\frac{ds}{dt} \right)^2 \vec{\mathbf{T}}(t) \times \vec{\mathbf{T}}\,'(t).$$

Since $\vec{\mathbf{T}}\,'$ is parallel to $\vec{\mathbf{N}}$, and $\vec{\mathbf{T}}$ is orthogonal to $\vec{\mathbf{N}}$, it follows that $\vec{\mathbf{T}}$ and $\vec{\mathbf{T}}\,'$ are orthogonal. This means that $\|\vec{\mathbf{T}} \times \vec{\mathbf{T}}\,'\| = \|\vec{\mathbf{T}}\|\|\vec{\mathbf{T}}\,'\| \sin(\pi/2) = \|\vec{\mathbf{T}}\,'\|$, so

$$\|\vec{\mathbf{r}}\,'(t) \times \vec{\mathbf{r}}\,''(t)\| = \left(\frac{ds}{dt} \right)^2 \|\vec{\mathbf{T}}\,'(t)\|.$$

Now we solve this equation for $\|\vec{\mathbf{T}}\,'(t)\|$ and use the fact that $ds/dt = \|\vec{\mathbf{r}}\,'(t)\|$:

$$\|\vec{\mathbf{T}}\,'(t)\| = \frac{\|\vec{\mathbf{r}}\,'(t) \times \vec{\mathbf{r}}\,''(t)\|}{\|\vec{\mathbf{r}}\,'(t)\|^2}.$$

Then, we divide both sides by $\|\vec{\mathbf{r}}\,'(t)\|$. This gives

$$\kappa = \frac{\|\vec{\mathbf{T}}\,'(t)\|}{\|\vec{\mathbf{r}}\,'(t)\|} = \frac{\|\vec{\mathbf{r}}\,'(t) \times \vec{\mathbf{r}}\,''(t)\|}{\|\vec{\mathbf{r}}\,'(t)\|^3}.$$

This proves 12.3.7. To prove 12.3.8, we start with the assumption that curve C is defined by the function $y = f(x)$. Then, we can define $\vec{\mathbf{r}}(t) = x\,\hat{\mathbf{i}} + f(x)\,\hat{\mathbf{j}} + 0\,\hat{\mathbf{k}}$. Using the previous formula for curvature:

$$\vec{\mathbf{r}}\,'(t) = \hat{\mathbf{i}} + f'(x)\,\hat{\mathbf{j}}$$

$$\vec{\mathbf{r}}\,''(t) = f''(x)\,\hat{\mathbf{j}}$$

$$\vec{\mathbf{r}}\,'(t) \times \vec{\mathbf{r}}\,''(t) = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ 1 & f'(x) & 0 \\ 0 & f''(x) & 0 \end{vmatrix} = f''(x)\,\hat{\mathbf{k}}.$$

Therefore,

$$\kappa = \frac{\|\vec{\mathbf{r}}\,'(t) \times \vec{\mathbf{r}}\,''(t)\|}{\|\vec{\mathbf{r}}\,'(t)\|^3} = \frac{|f''(x)|}{(1+[f'(x)]^2)^{3/2}}$$

Example 12.3.3: Finding Curvature

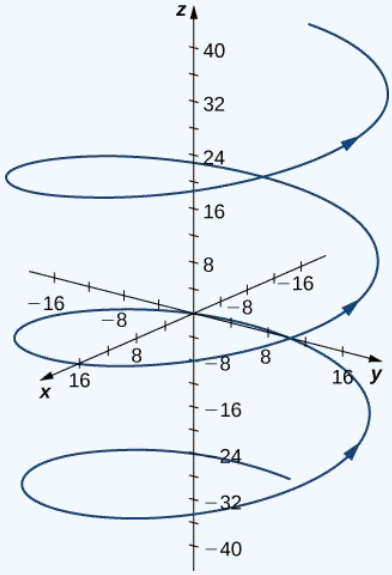
Find the curvature for each of the following curves at the given point:

a. $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} + 4\sin t\,\hat{\mathbf{j}} + 3t\,\hat{\mathbf{k}}, \quad t = \dfrac{4\pi}{3}$

b. $f(x) = \sqrt{4x - x^2}, x = 2$

Solution

a. This function describes a helix.



The curvature of the helix at $t = (4\pi)/3$ can be found by using 12.3.6. First, calculate $\vec{\mathbf{T}}(t)$:

$$\begin{aligned} \vec{\mathbf{T}}(t) &= \frac{\vec{\mathbf{r}}\,'(t)}{\|\vec{\mathbf{r}}\,'(t)\|} \\ &= \frac{\langle -4\sin t, 4\cos t, 3\rangle}{\sqrt{(-4\sin t)^2 + (4\cos t)^2 + 3^2}} \\ &= \langle -\frac{4}{5}\sin t, \frac{4}{5}\cos t, \frac{3}{5}\rangle. \end{aligned}$$

Next, calculate $\vec{\mathbf{T}}\,'(t)$:

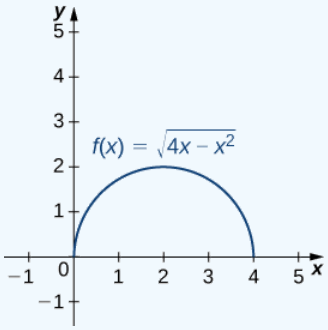
$$\vec{\mathbf{T}}\,'(t) = \langle -\frac{4}{5}\cos t, -\frac{4}{5}\sin t, 0\rangle.$$

Last, apply 12.3.6:

$$\kappa = \frac{\|\vec{\mathbf{T}}'(t)\|}{\|\vec{\mathbf{r}}'(t)\|} = \frac{\|\langle -\frac{4}{5}\cos t, -\frac{4}{5}\sin t, 0\rangle\|}{\|\langle -4\sin t, 4\cos t, 3\rangle\|}$$
$$= \frac{\sqrt{(-\frac{4}{5}\cos t)^2 + (-\frac{4}{5}\sin t)^2 + 0^2}}{\sqrt{(-4\sin t)^2 + (4\cos t)^2 + 3^2}}$$
$$= \frac{4/5}{5} = \frac{4}{25}.$$

The curvature of this helix is constant at all points on the helix.

2. This function describes a semicircle.



To find the curvature of this graph, we must use 12.3.8. First, we calculate y' and y'':

$$y = \sqrt{4x - x^2} = (4x - x^2)^{1/2}$$
$$y' = \frac{1}{2}(4x - x^2)^{-1/2}(4 - 2x) = (2 - x)(4x - x^2)^{-1/2}$$
$$y'' = -(4x - x^2)^{-1/2} + (2 - x)(-\frac{1}{2})(4x - x^2)^{-3/2}(4 - 2x)$$
$$= -\frac{4x - x^2}{(4x - x^2)^{3/2}} - \frac{(2 - x)^2}{(4x - x^2)^{3/2}}$$
$$= \frac{x^2 - 4x - (4 - 4x + x^2)}{(4x - x^2)^{3/2}}$$
$$= -\frac{4}{(4x - x^2)^{3/2}}.$$

Then, we apply 12.3.8:

$$\begin{aligned}\kappa &= \frac{|y''|}{[1+(y')^2]^{3/2}} \\ &= \frac{\left|-\frac{4}{(4x-x^2)^{3/2}}\right|}{\left[1+((2-x)(4x-x^2)^{-1/2})^2\right]^{3/2}} = \frac{\left|\frac{4}{(4x-x^2)^{3/2}}\right|}{\left[1+\frac{(2-x)^2}{4x-x^2}\right]^{3/2}} \\ &= \frac{\left|\frac{4}{(4x-x^2)^{3/2}}\right|}{\left[\frac{4x-x^2+x^2-4x+4}{4x-x^2}\right]^{3/2}} = \left|\frac{4}{(4x-x^2)^{3/2}}\right| \cdot \frac{(4x-x^2)^{3/2}}{8} \\ &= \frac{1}{2}.\end{aligned}$$

The curvature of this circle is equal to the reciprocal of its radius. There is a minor issue with the absolute value in 12.3.8; however, a closer look at the calculation reveals that the denominator is positive for any value of x.

? Exercise 12.3.3

Find the curvature of the curve defined by the function

$$y = 3x^2 - 2x + 4$$

at the point $x = 2$.

Hint

Use 12.3.8

Answer

$\kappa = \frac{6}{101^{3/2}} \approx 0.0059$

The Normal and Binormal Vectors

We have seen that the derivative $\vec{\mathbf{r}}\,'(t)$ of a vector-valued function is a tangent vector to the curve defined by $\vec{\mathbf{r}}(t)$, and the unit tangent vector $\vec{\mathbf{T}}(t)$ can be calculated by dividing $\vec{\mathbf{r}}\,'(t)$ by its magnitude. When studying motion in three dimensions, two other vectors are useful in describing the motion of a particle along a path in space: the principal unit normal vector and the **binormal vector**.

Definition: Binormal Vectors

Let C be a three-dimensional **smooth** curve represented by $\vec{\mathbf{r}}$ over an open interval I. If $\vec{\mathbf{T}}\,'(t) \neq \vec{\mathbf{0}}$, then the principal unit normal vector at t is defined to be

$$\vec{\mathbf{N}}(t) = \frac{\vec{\mathbf{T}}\,'(t)}{\|\vec{\mathbf{T}}\,'(t)\|}. \tag{12.3.9}$$

The binormal vector at t is defined as

$$\vec{\mathbf{B}}(t) = \vec{\mathbf{T}}(t) \times \vec{\mathbf{N}}(t), \tag{12.3.10}$$

where $\vec{\mathbf{T}}(t)$ is the unit tangent vector.

Note that, by definition, the binormal vector is orthogonal to both the unit tangent vector and the normal vector. Furthermore, $\vec{\mathbf{B}}(t)$ is always a unit vector. This can be shown using the formula for the magnitude of a cross product.

$$\|\vec{\mathbf{B}}(t)\| = \|\vec{\mathbf{T}}(t) \times \vec{\mathbf{N}}(t)\| = \|\vec{\mathbf{T}}(t)\|\|\vec{\mathbf{N}}(t)\| \sin\theta,$$

where θ is the angle between $\vec{\mathbf{T}}(t)$ and $\vec{\mathbf{N}}(t)$. Since $\vec{\mathbf{N}}(t)$ is the derivative of a unit vector, property (vii) of the derivative of a vector-valued function tells us that $\vec{\mathbf{T}}(t)$ and $\vec{\mathbf{N}}(t)$ are orthogonal to each other, so $\theta = \pi/2$. Furthermore, they are both unit vectors, so their magnitude is 1. Therefore, $\|\vec{\mathbf{T}}(t)\|\|\vec{\mathbf{N}}(t)\| \sin\theta = (1)(1)\sin(\pi/2) = 1$ and $\vec{\mathbf{B}}(t)$ is a unit vector.

The principal unit normal vector can be challenging to calculate because the unit tangent vector involves a quotient, and this quotient often has a square root in the denominator. In the three-dimensional case, finding the cross product of the unit tangent vector and the unit normal vector can be even more cumbersome. Fortunately, we have alternative formulas for finding these two vectors, and they are presented in Motion in Space.

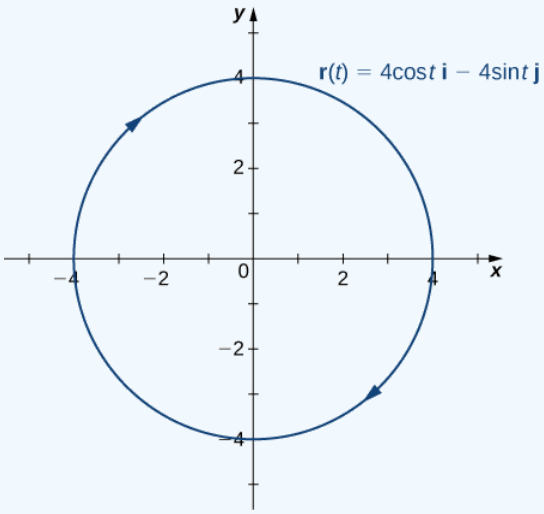
Example 12.3.4: Finding the Principal Unit Normal Vector and Binormal Vector

For each of the following vector-valued functions, find the principal unit normal vector. Then, if possible, find the binormal vector.

1. $\vec{\mathbf{r}}(t) = 4\cos t\,\hat{\mathbf{i}} - 4\sin t\,\hat{\mathbf{j}}$
2. $\vec{\mathbf{r}}(t) = (6t+2)\,\hat{\mathbf{i}} + 5t^2\,\hat{\mathbf{j}} - 8t\,\hat{\mathbf{k}}$

Solution

1. This function describes a circle.



To find the principal unit normal vector, we first must find the unit tangent vector $\vec{\mathbf{T}}(t)$:

$$\begin{aligned}
\vec{\mathbf{T}}(t) &= \frac{\vec{\mathbf{r}}\,'(t)}{\|\vec{\mathbf{r}}\,'(t)\|} \\
&= \frac{-4\sin t\,\hat{\mathbf{i}} - 4\cos t\,\hat{\mathbf{j}}}{\sqrt{(-4\sin t)^2 + (-4\cos t)^2}} \\
&= \frac{-4\sin t\,\hat{\mathbf{i}} - 4\cos t\,\hat{\mathbf{j}}}{\sqrt{16\sin^2 t + 16\cos^2 t}} \\
&= \frac{-4\sin t\,\hat{\mathbf{i}} - 4\cos t\,\hat{\mathbf{j}}}{\sqrt{16(\sin^2 t + \cos^2 t)}} \\
&= \frac{-4\sin t\,\hat{\mathbf{i}} - 4\cos t\,\hat{\mathbf{j}}}{4} \\
&= -\sin t\,\hat{\mathbf{i}} - \cos t\,\hat{\mathbf{j}}.
\end{aligned}$$

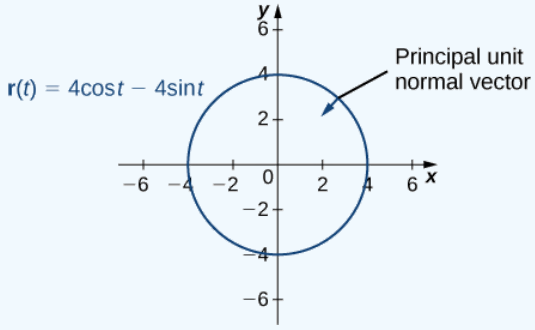
Next, we use 12.3.9:

$$\begin{aligned}\vec{\mathbf{N}}(t) &= \frac{\vec{\mathbf{T}}'(t)}{\|\vec{\mathbf{T}}'(t)\|} \\ &= \frac{-\cos t\,\hat{\mathbf{i}} + \sin t\,\hat{\mathbf{j}}}{\sqrt{(-\cos t)^2 + (\sin t)^2}} \\ &= \frac{-\cos t\,\hat{\mathbf{i}} + \sin t\,\hat{\mathbf{j}}}{\sqrt{\cos^2 t + \sin^2 t}} \\ &= -\cos t\,\hat{\mathbf{i}} + \sin t\,\hat{\mathbf{j}}.\end{aligned}$$

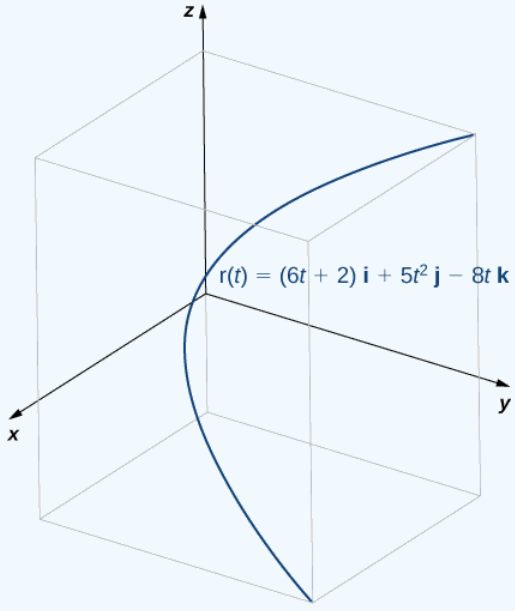
Notice that the unit tangent vector and the principal unit normal vector are orthogonal to each other for all values of t:

$$\begin{aligned}\vec{\mathbf{T}}(t) \cdot \vec{\mathbf{N}}(t) &= \langle -\sin t, -\cos t\rangle \cdot \langle -\cos t, \sin t\rangle \\ &= \sin t \cos t - \cos t \sin t \\ &= 0.\end{aligned}$$

Furthermore, the principal unit normal vector points toward the center of the circle from every point on the circle. Since $\vec{\mathbf{r}}(t)$ defines a curve in two dimensions, we cannot calculate the binormal vector.



2. This function looks like this:



To find the principal unit normal vector, we first find the unit tangent vector $\vec{\mathbf{T}}(t)$:

$$\begin{aligned}\vec{\mathbf{T}}(t) &= \frac{\vec{\mathbf{r}}\,'(t)}{\|\vec{\mathbf{r}}\,'(t)\|} \\ &= \frac{6\,\hat{\mathbf{i}} + 10t\,\hat{\mathbf{j}} - 8\,\hat{\mathbf{k}}}{\sqrt{6^2 + (10t)^2 + (-8)^2}} \\ &= \frac{6\,\hat{\mathbf{i}} + 10t\,\hat{\mathbf{j}} - 8\,\hat{\mathbf{k}}}{\sqrt{36 + 100t^2 + 64}} \\ &= \frac{6\,\hat{\mathbf{i}} + 10t\,\hat{\mathbf{j}} - 8\,\hat{\mathbf{k}}}{\sqrt{100(t^2+1)}} \\ &= \frac{3\,\hat{\mathbf{i}} - 5t\,\hat{\mathbf{j}} - 4\,\hat{\mathbf{k}}}{5\sqrt{t^2+1}} \\ &= \frac{3}{5}(t^2+1)^{-1/2}\,\hat{\mathbf{i}} - t(t^2+1)^{-1/2}\,\hat{\mathbf{j}} - \frac{4}{5}(t^2+1)^{-1/2}\,\hat{\mathbf{k}}.\end{aligned}$$

Next, we calculate $\vec{\mathbf{T}}\,'(t)$ and $\|\vec{\mathbf{T}}\,'(t)\|$:

$$\begin{aligned}\vec{\mathbf{T}}\,'(t) &= \frac{3}{5}(-\frac{1}{2})(t^2+1)^{-3/2}(2t)\,\hat{\mathbf{i}} - ((t^2+1)^{-1/2} - t(\frac{1}{2})(t^2+1)^{-3/2}(2t))\,\hat{\mathbf{j}} - \frac{4}{5}(-\frac{1}{2})(t^2+1)^{-3/2}(2t)\,\hat{\mathbf{k}} \\ &= -\frac{3t}{5(t^2+1)^{3/2}}\,\hat{\mathbf{i}} - \frac{1}{(t^2+1)^{3/2}}\,\hat{\mathbf{j}} + \frac{4t}{5(t^2+1)^{3/2}}\,\hat{\mathbf{k}} \\ \|\vec{\mathbf{T}}\,'(t)\| &= \sqrt{\left(-\frac{3t}{5(t^2+1)^{3/2}}\right)^2 + \left(-\frac{1}{(t^2+1)^{3/2}}\right)^2 + \left(\frac{4t}{5(t^2+1)^{3/2}}\right)^2} \\ &= \sqrt{\frac{9t^2}{25(t^2+1)^3} + \frac{1}{(t^2+1)^3} + \frac{16t^2}{25(t^2+1)^3}} \\ &= \sqrt{\frac{25t^2+25}{25(t^2+1)^3}} \\ &= \sqrt{\frac{1}{(t^2+1)^2}} \\ &= \frac{1}{t^2+1}.\end{aligned}$$

Therefore, according to 12.3.9:

$$\begin{aligned}\vec{\mathbf{N}}(t) &= \frac{\vec{\mathbf{T}}\,'(t)}{\|\vec{\mathbf{T}}\,'(t)\|} \\ &= \left(-\frac{3t}{5(t^2+1)^{3/2}}\,\hat{\mathbf{i}} - \frac{1}{(t^2+1)^{3/2}}\,\hat{\mathbf{j}} + \frac{4t}{5(t^2+1)^{3/2}}\,\hat{\mathbf{k}}\right)(t^2+1) \\ &= -\frac{3t}{5(t^2+1)^{1/2}}\,\hat{\mathbf{i}} - \frac{5}{5(t^2+1)^{1/2}}\,\hat{\mathbf{j}} + \frac{4t}{5(t^2+1)^{1/2}}\,\hat{\mathbf{k}} \\ &= -\frac{3t\,\hat{\mathbf{i}} + 5\,\hat{\mathbf{j}} - 4t\,\hat{\mathbf{k}}}{5\sqrt{t^2+1}}.\end{aligned}$$

Once again, the unit tangent vector and the principal unit normal vector are orthogonal to each other for all values of t:

$$\begin{aligned}\vec{\mathbf{T}}(t)\cdot\vec{\mathbf{N}}(t) &= \left(\frac{3\,\hat{\mathbf{i}}-5t\,\hat{\mathbf{j}}-4\,\hat{\mathbf{k}}}{5\sqrt{t^2+1}}\right)\cdot\left(-\frac{3t\,\hat{\mathbf{i}}+5\,\hat{\mathbf{j}}-4t\,\hat{\mathbf{k}}}{5\sqrt{t^2+1}}\right)\\ &= \frac{3(-3t)-5t(-5)-4(4t)}{25(t^2+1)}\\ &= \frac{-9t+25t-16t}{25(t^2+1)}\\ &= 0.\end{aligned}$$

Last, since $\vec{\mathbf{r}}(t)$ represents a three-dimensional curve, we can calculate the binormal vector using 12.3.10:

$$\begin{aligned}\vec{\mathbf{B}}(t) &= \vec{\mathbf{T}}(t)\times\vec{\mathbf{N}}(t)\\ &= \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \dfrac{3}{5\sqrt{t^2+1}} & -\dfrac{5t}{5\sqrt{t^2+1}} & -\dfrac{4}{5\sqrt{t^2+1}} \\ -\dfrac{3t}{5\sqrt{t^2+1}} & -\dfrac{5}{5\sqrt{t^2+1}} & \dfrac{4t}{5\sqrt{t^2+1}} \end{vmatrix}\\ &= \left(\left(-\frac{5t}{5\sqrt{t^2+1}}\right)\left(\frac{4t}{5\sqrt{t^2+1}}\right)-\left(-\frac{4}{5\sqrt{t^2+1}}\right)\left(-\frac{5}{5\sqrt{t^2+1}}\right)\right)\hat{\mathbf{i}}\\ &\quad-\left(\left(\frac{3}{5\sqrt{t^2+1}}\right)\left(\frac{4t}{5\sqrt{t^2+1}}\right)-\left(-\frac{4}{5\sqrt{t^2+1}}\right)\left(-\frac{3t}{5\sqrt{t^2+1}}\right)\right)\hat{\mathbf{j}}\\ &\quad+\left(\left(\frac{3}{5\sqrt{t^2+1}}\right)\left(-\frac{5}{5\sqrt{t^2+1}}\right)-\left(-\frac{5t}{5\sqrt{t^2+1}}\right)\left(-\frac{3t}{5\sqrt{t^2+1}}\right)\right)\hat{\mathbf{k}}\\ &= \left(\frac{-20t^2-20}{25(t^2+1)}\right)\hat{\mathbf{i}}+\left(\frac{-15-15t^2}{25(t^2+1)}\right)\hat{\mathbf{k}}\\ &= -20\left(\frac{t^2+1}{25(t^2+1)}\right)\hat{\mathbf{i}}-15\left(\frac{t^2+1}{25(t^2+1)}\right)\hat{\mathbf{k}}\\ &= -\frac{4}{5}\,\hat{\mathbf{i}}-\frac{3}{5}\,\hat{\mathbf{k}}.\end{aligned}$$

? Exercise 12.3.4

Find the unit normal vector for the vector-valued function $\vec{\mathbf{r}}(t)=(t^2-3t)\,\hat{\mathbf{i}}+(4t+1)\,\hat{\mathbf{j}}$ and evaluate it at $t=2$.

Hint

First, find $\vec{\mathbf{T}}(t)$, then use 12.3.9

Answer

$\vec{\mathbf{N}}(2)=\dfrac{\sqrt{2}}{2}(\,\hat{\mathbf{i}}-\hat{\mathbf{j}})$

For any smooth curve in three dimensions that is defined by a vector-valued function, we now have formulas for the unit tangent vector $\vec{\mathbf{T}}$, the unit normal vector $\vec{\mathbf{N}}$, and the binormal vector $\vec{\mathbf{B}}$. The unit normal vector and the binormal vector form a plane that is perpendicular to the curve at any point on the curve, called the normal plane. In addition, these three vectors form a frame of reference in three-dimensional space called the *Frenet frame of reference* (also called the **TNB** frame) (Figure 12.3.2). Last, the plane determined by the vectors $\vec{\mathbf{T}}$ and $\vec{\mathbf{N}}$ forms the osculating plane of C at any point P on the curve.

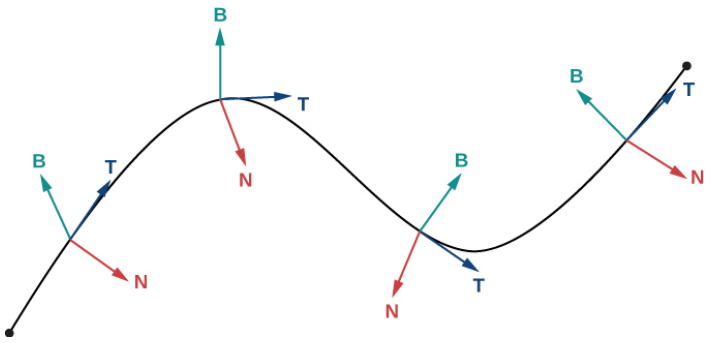


Figure 12.3.2: This figure depicts a Frenet frame of reference. At every point P *on a three-dimensional curve, the unit tangent, unit normal, and binormal vectors form a three-dimensional frame of reference.*

Suppose we form a circle in the osculating plane of C at point P on the curve. Assume that the circle has the same curvature as the curve does at point P and let the circle have radius r. Then, the curvature of the circle is given by $\frac{1}{r}$. We call r the radius of curvature of the curve, and it is equal to the reciprocal of the curvature. If this circle lies on the concave side of the curve and is tangent to the curve at point P, then this circle is called the *osculating circle* of C at P, as shown in Figure 12.3.3

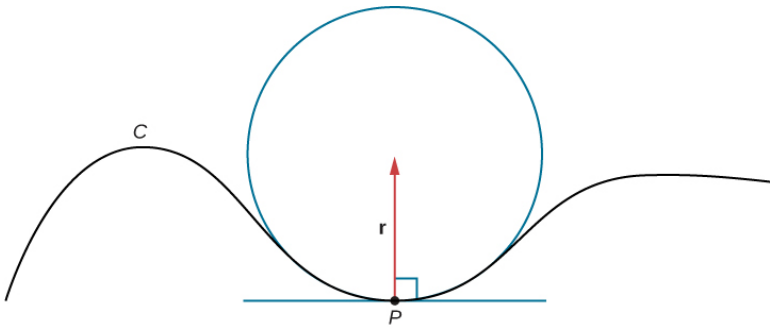


Figure 12.3.3: In this osculating circle, the circle is tangent to curve C at point P *and shares the same curvature.*

For more information on osculating circles, see this demonstration on curvature and torsion, this article on osculating circles, and this discussion of Serret formulas.

To find the equation of an osculating circle in two dimensions, we need find only the center and radius of the circle.

Example 12.3.5: Finding the Equation of an Osculating Circle

Find the equation of the osculating circle of the curve defined by the function $y = x^3 - 3x + 1$ at $x = 1$.

Solution

Figure 12.3.4shows the graph of $y = x^3 - 3x + 1$.

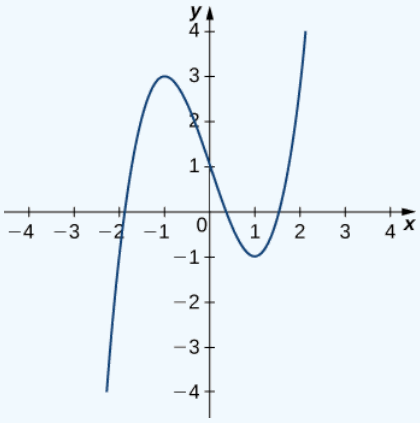


Figure 12.3.4: We want to find the osculating circle of this graph at the point where $x = 1$.

First, let's calculate the curvature at $x = 1$:

$$\kappa = \frac{|f''(x)|}{\left(1+[f'(x)]^2\right)^{3/2}} = \frac{|6x|}{(1+[3x^2-3]^2)^{3/2}}.$$

This gives $\kappa = 6$. Therefore, the radius of the osculating circle is given by $R = \frac{1}{\kappa} = \frac{1}{6}$. Next, we then calculate the coordinates of the center of the circle. When $x = 1$, the slope of the tangent line is zero. Therefore, the center of the osculating circle is directly above the point on the graph with coordinates $(1, -1)$. The center is located at $(1, -\frac{5}{6})$. The formula for a circle with radius r and center (h, k) is given by $(x-h)^2 + (y-k)^2 = r^2$. Therefore, the equation of the osculating circle is $(x-1)^2 + (y+\frac{5}{6})^2 = \frac{1}{36}$. The graph and its osculating circle appears in the following graph.

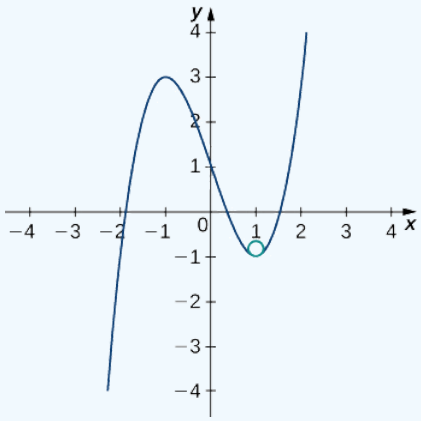


Figure 12.3.5: The osculating circle has radius $R = \frac{1}{6}$.

Exercise 12.3.5

Find the equation of the osculating circle of the curve defined by the vector-valued function $y = 2x^2 - 4x + 5$ at $x = 1$.

Hint

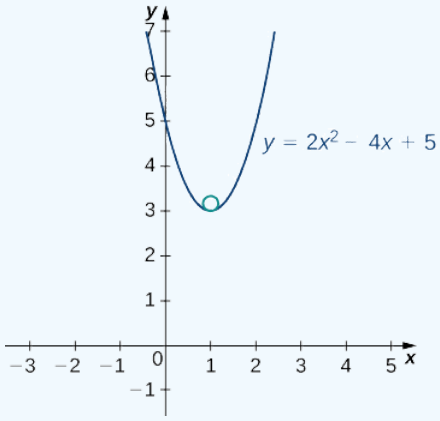
Use 12.3.8to find the curvature of the graph, then draw a graph of the function around $x = 1$ to help visualize the circle in relation to the graph.

Answer

$\kappa = \frac{4}{[1+(4x-4)^2]^{3/2}}$

At the point $x = 1$, the curvature is equal to 4. Therefore, the radius of the osculating circle is $\frac{1}{4}$.

A graph of this function appears next:



The vertex of this parabola is located at the point $(1,3)$. Furthermore, the center of the osculating circle is directly above the vertex. Therefore, the coordinates of the center are $(1, \frac{13}{4})$. The equation of the osculating circle is $(x-1)^2+(y-\frac{13}{4})^2=\frac{1}{16}$.

Key Concepts

- The arc-length function for a vector-valued function is calculated using the integral formula $s(t)=\int_a^b \|\vec{\mathbf{r}}\,'(t)\|\,dt$. This formula is valid in both two and three dimensions.
- The curvature of a curve at a point in either two or three dimensions is defined to be the curvature of the inscribed circle at that point. The arc-length parameterization is used in the definition of curvature.
- There are several different formulas for curvature. The curvature of a circle is equal to the reciprocal of its radius.
- The principal unit normal vector at t is defined to be

$$\vec{\mathbf{N}}(t)=\frac{\vec{\mathbf{T}}\,'(t)}{\|\vec{\mathbf{T}}\,'(t)\|}.$$

- The binormal vector at t is defined as $\vec{\mathbf{B}}(t)=\vec{\mathbf{T}}(t)\times\vec{\mathbf{N}}(t)$, where $\vec{\mathbf{T}}(t)$ is the unit tangent vector.
- The Frenet frame of reference is formed by the unit tangent vector, the principal unit normal vector, and the binormal vector.
- The osculating circle is tangent to a curve at a point and has the same curvature as the tangent curve at that point.

Key Equations

- **Arc length of space curve**
  $s=\int_a^b \sqrt{[f\,'(t)]^2+[g\,'(t)]^2+[h\,'(t)]^2}\,dt=\int_a^b \|\vec{\mathbf{r}}\,'(t)\|\,dt$
- **Arc-length function**
  $s(t)=\int_a^t \sqrt{f\,'(u))^2+(g\,'(u))^2+(h\,'(u))^2}\,du$ *or* $s(t)=\int_a^t \|\vec{\mathbf{r}}\,'(u)\|\,du$
- $\kappa=\frac{\|\vec{\mathbf{T}}\,'(t)\|}{\|\vec{\mathbf{r}}\,'(t)\|}$ *or* $\kappa=\frac{\|\vec{\mathbf{r}}\,'(t)\times\vec{\mathbf{r}}\,''(t)\|}{\|\vec{\mathbf{r}}\,'(t)\|^3}$ *or* $\kappa=\frac{|y\,''|}{[1+(y\,')^2]^{3/2}}$
- **Principal unit normal vector**
  $\vec{\mathbf{N}}(t)=\frac{\vec{\mathbf{T}}\,'(t)}{\|\vec{\mathbf{T}}\,'(t)\|}$
- **Binormal vector**
  $\vec{\mathbf{B}}(t)=\vec{\mathbf{T}}(t)\times\vec{\mathbf{N}}(t)$

Glossary

arc-length function

a function $s(t)$ that describes the arc length of curve C as a function of t

arc-length parameterization

a reparameterization of a vector-valued function in which the parameter is equal to the arc length

binormal vector

a unit vector orthogonal to the unit tangent vector and the unit normal vector

curvature

the derivative of the unit tangent vector with respect to the arc-length parameter

Frenet frame of reference

(TNB frame) a frame of reference in three-dimensional space formed by the unit tangent vector, the unit normal vector, and the binormal vector

normal plane

a plane that is perpendicular to a curve at any point on the curve

osculating circle

a circle that is tangent to a curve C at a point P and that shares the same curvature

osculating plane

the plane determined by the unit tangent and the unit normal vector

principal unit normal vector

a vector orthogonal to the unit tangent vector, given by the formula $\frac{\vec{\mathbf{T}}'(t)}{\|\vec{\mathbf{T}}'(t)\|}$

radius of curvature

the reciprocal of the curvature

smooth

curves where the vector-valued function $\vec{\mathbf{r}}(t)$ is differentiable with a non-zero derivative

---

12.3: Arc Length in Space is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **13.3: Arc Length and Curvature** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

12.4: Curvature and Normal Vectors of a Curve

For a parametrically defined curve we had the definition of arc length. Since vector valued functions are parametrically defined curves in disguise, we have the same definition. We have the added benefit of notation with vector valued functions in that the square root of the sum of the squares of the derivatives is just the magnitude of the velocity vector.

Definition: Arc Length

Let

$$\mathbf{r}(t) = x(t)\,\hat{\mathbf{i}} + y(t)\,\hat{\mathbf{j}} + z(t)\,\hat{\mathbf{k}}$$

be a differentiable vector valued function on [a,b]. Then the arc length s is defined by

$$s = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\, dt = \int_a^b |v(t)|\, dt.$$

Example 12.4.1

Suppose that

$$\mathbf{r}(t) = 3t\,\hat{\mathbf{i}} + 2\,\hat{\mathbf{j}} + t^2\,\hat{\mathbf{k}}$$

Set up the integral that defines the arc length of the curve from 2 to 3. Then use a calculator or computer to approximate the arc length.

Solution

We use the arc length formula

$$s = \int_2^3 \sqrt{9 + 0 + 4t^2}\, dt = \int_2^3 \sqrt{9 + 4t^2}\, dt.$$

Notice that we could do this integral by hand by letting $t = 9/2 \tan q$, however the question only asked us to use a machine to approximate the integral:

$$s = 5.8386.$$

Parameterization by Arc Length

Recall that like parametric equations, vector valued function describe not just the path of the particle, but also how the particle is moving. Among all representations of a curve there is a "simplest" one. If the particle travels at the constant rate of one unit per second, then we say that the curve is *parameterized by arc length*. We have seen this concept before in the definition of radians. On a unit circle one radian is one unit of arc length around the circle. When we say "simplest" we in no way mean that the equations are simple to find, but rather that the dynamics of the particle are simple. To aid us in parameterizing by arc length, we define the arc length function.

Definition: Arc Length Function

If $\mathbf{r}(t)$ is a differentiable vector valued function, then the *arc length function* is defined by

$$s(t) = \int_0^t ||\mathbf{v}(u)||\, du.$$

Remark: By the **second fundamental theorem of calculus**, we have

$$s'(t) = ||v(t)||.$$

If a vector valued function is parameterized by arc length, then

$$s(t) = t.$$

If we have a vector valued function$r(t)$ with arc length s(t), then we can introduce a new variable

$$s = s^{-1}(t).$$

So that the vector valued function $r(s)$ will have arc length equal to

$$s\left(s^{-1}(t)\right) = t.$$

and $r(s)$ will be parameterized by arc length. Unfortunately, this process is usually impossible for two reasons.

1. The integral that defines arc length involves a square root in the integrand; this integral is usually impossible to determine.
2. Even if the integral is possible to evaluate, finding the inverse of a function is often impossible. There are a few special curves that can be parameterized by arc length and one is demonstrated below.

Example 12.4.2: Parameterizing by Arc Length

Find the arc length parameterization of the helix defined by

$$\mathbf{r}(t) = \cos t\,\hat{\mathbf{i}} + \sin t\,\hat{\mathbf{j}} + t\hat{\mathbf{k}}.$$

Solution

First find the arc length function

$$s(t) = \int_0^t \sqrt{\sin^2 u + \cos^2 u + 1}\; dt = \int_0^t \sqrt{2}\, dt = \sqrt{2}\, t.$$

Solving for t gives

$$t = \frac{s}{\sqrt{2}}.$$

Now substitute back into the position equation to get

$$\mathbf{r}(s) = \cos\frac{s}{\sqrt{2}}\,\hat{\mathbf{i}} + \sin\frac{s}{\sqrt{2}}\,\hat{\mathbf{j}} + \frac{s}{\sqrt{2}}\,\hat{\mathbf{k}}.$$

Concepts: Curvature and Normal Vector

Consider a car driving along a curvy road. The tighter the curve, the more difficult the driving is. In math we have a number, the *curvature*, that describes this "tightness". If the curvature is zero then the curve looks like a line near this point. While if the curvature is a large number, then the curve has a sharp bend.

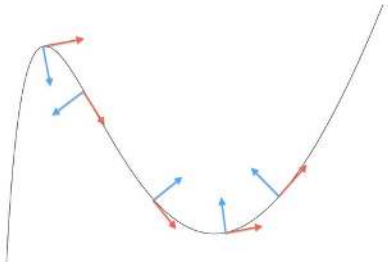


Figure 12.4.1: Below image is a part of a curve $\mathbf{r}(t)$ Red arrows represent unit tangent vectors, $\hat{\mathbf{T}}$, and blue arrows represent unit normal vectors, $\hat{\mathbf{N}}$.

Before learning what curvature of a curve is and how to find the value of that curvature, we must first learn about *unit tangent vector*. As the name suggests, unit tangent vectors are unit vectors (vectors with length of 1) that are tangent to the curve at certain points. Because tangent lines at certain point of a curve are defined as lines that barely touch the curve at the given point, we can deduce that tangent lines or vectors have slopes equivalent to the instantaneous slope of a curve at the given point. In other words,

$$\mathbf{T} = \frac{d\mathbf{r}}{dt},$$

which means

$$\hat{\mathbf{T}} = \frac{\mathbf{T}}{|\mathbf{T}|} = \frac{d\mathbf{r}/dt}{|d\mathbf{r}/dt|}.$$

Based on what we learned previosuly, we know that $\frac{d\mathbf{r}}{dt} = \mathbf{v}$, where $\mathbf{v}$ is the velocity at which a point is moving at a given time. Furthermore, the absolute value of the velocity vector is the speed vector of the curve, meaning $\left|\frac{d\mathbf{r}}{dt}\right| = \frac{ds}{dt}$. So the formula for unit tangent vector can be simplified to:

$$\hat{\mathbf{T}} = \frac{\text{velocity}}{\text{speed}} = \frac{d\mathbf{r}/dt}{ds/dt}.$$

And now, let's think about the unit tangent vector when the curve is explained in terms of arc length, that is, $r(s)$ instead of $r(t)$. This means:

$$\mathbf{T} = \frac{d\mathbf{r}}{ds}$$

$$\text{and } \hat{\mathbf{T}} = \frac{d\mathbf{r}/ds}{ds/ds} = \frac{d\mathbf{r}}{ds}.$$

With this information, we will be learning what curvature really is and how we can calculate the curvature, denoted as κ.

Curvature of a Curve

Curvature is a measure of how much the curve deviates from a straight line. In other words, the curvature of a curve at a point is a measure of how much the change in a curve at a point is changing, meaning the curvature is the magnitude of the second derivative of the curve at given point (let's assume that the curve is defined in terms of the arc length s to make things easier). This means:

$$k = \left|\frac{d^2\mathbf{r}}{ds^2}\right|.$$

Since we know that $\hat{\mathbf{T}} = d\mathbf{r}/ds$, we can formulate an equation for κ in terms of $\hat{\mathbf{T}}$:

$$k = \left|\frac{d\hat{\mathbf{T}}}{ds}\right|.$$

Never the less, we know that most curves are written in parametric equations in terms of some dummy variable, most commonly t. So let's assume that the curve is in terms of t, such that $\mathbf{r}(t)$ is a curve. In such case, we must formulate another equation to find the curvature without taking derivatives in terms of s.

First, we know that

$$k = \left|\frac{d\hat{\mathbf{T}}}{ds}\right|$$

Using Chain Rule, we get

$$k = \left|\frac{d\hat{\mathbf{T}}}{dt} \cdot \frac{dt}{ds}\right|$$

$$= \frac{1}{|ds/dt|}\left|\frac{d\hat{\mathbf{T}}}{dt}\right|$$

therefore

$$k = \frac{1}{|\mathbf{v}|}\left|\frac{d\hat{\mathbf{T}}}{dt}\right|.$$

Definition of Curvature (repeat)

More formally, if $\mathbf{T}(t)$ is the *unit tangent vector* function then the curvature, k, is defined at the rate at which the *unit tangent vector* changes with respect to arc length.

$$k = ||\frac{d}{ds}(\mathbf{T}(t))|| = ||\mathbf{r}''(s)||$$

As stated previously, this is not a practical definition, since parameterizing by arc length is typically impossible. Instead we use the chain rule to get

$$||\frac{d}{ds}(\mathbf{T}(t))|| = ||\mathbf{T}'(t)\frac{dt}{ds}||$$

$$\frac{||\mathbf{T}'(t)||}{||\frac{ds}{dt}||} = \frac{||\mathbf{T}'(t)||}{||\mathbf{r}'(t)||}.$$

This formula is more practical to use, but still cumbersome. $\mathbf{T}'(t)$ is typically a mess. Instead we can borrow from the formula for the normal vector to get the curvature

$$K(t) = \frac{||r'(t) \times r''(t)||}{||r'(t)||^3}.$$

Normal Vector of a Curve

A unit normal vector of a curve, by its definition, is perpendicular to the curve at given point. This means a normal vector of a curve at a given point is perpendicular to the tangent vector at the same point. Furthermore, a normal vector points towards the center of curvature, and the derivative of tangent vector also points towards the center of curvature. In summary, normal vector of a curve is the derivative of tangent vector of a curve.

$$\mathbf{N} = \frac{d\hat{\mathbf{T}}}{ds} \text{or} \frac{d\hat{\mathbf{T}}}{dt}$$

To find the unit normal vector, we simply divide the normal vector by its magnitude:

$$\hat{\mathbf{N}} = \frac{d\hat{\mathbf{T}}/ds}{\left|d\hat{\mathbf{T}}/ds\right|} \text{or} \frac{d\hat{\mathbf{T}}/dt}{\left|d\hat{\mathbf{T}}/dt\right|}.$$

Notice that $\left|d\hat{\mathbf{T}}/ds\right|$ can be replaced with κ, such that:

$$\hat{\mathbf{N}} = \frac{1}{\kappa}\frac{d\hat{\mathbf{T}}}{ds}$$

$$\therefore \hat{\mathbf{N}} = \frac{1}{\kappa}\frac{d\hat{\mathbf{T}}}{ds} \text{or} \frac{d\hat{\mathbf{T}}/dt}{\left|d\hat{\mathbf{T}}/dt\right|}.$$

Example 12.4.3

Find the curvature at $t = \frac{\pi}{2}$ if

$$r(t) = \cos t\,\hat{\mathbf{i}} - \frac{1}{t}\hat{\mathbf{j}} + \sin t\,\hat{\mathbf{k}}.$$

Solution

We take derivatives

$$\mathbf{r}'(t) = -\sin t\,\hat{\mathbf{i}} + \frac{1}{t^2}\,\hat{\mathbf{j}} + \cos t\,\hat{\mathbf{k}}$$

$$\mathbf{r}''(t) = -\cos t\,\hat{\mathbf{i}} - \frac{2}{t^3}\,\hat{\mathbf{j}} - \sin t\,\hat{\mathbf{k}}.$$

Plugging in $t = \frac{\pi}{2}$ gives

$$\begin{aligned}\mathbf{r}'\left(\frac{\pi}{2}\right) &= -\hat{\mathbf{i}} + \frac{4}{\pi^2}\,\hat{\mathbf{j}} \\ &= -\frac{16}{\pi^3}\,\hat{\mathbf{j}} - \hat{\mathbf{k}} \\ &\mathbf{r}''\left(\frac{\pi}{2}\right).\end{aligned}$$

Now take the cross product to get

$$\mathbf{r}'(\pi/2) \times \mathbf{r}''(\pi/2) = -\frac{4}{\pi^2}\,\hat{\mathbf{i}} - \hat{\mathbf{j}} + \frac{16}{\pi^3}\,\hat{\mathbf{k}}$$

Finally, we plug this information into the curvature formula to get

$$\frac{\sqrt{\frac{16}{\pi^4} + 1 + \frac{256}{\pi^6}}}{\left(\sqrt{1 + \frac{16}{\pi^4}}\right)^3} \approx 0.952.$$

Curvature of a Plane Curve

If a curve resides only in the xy-plane and is defined by the function $y = f(t)$ then there is an easier formula for the curvature. We can parameterize the curve by

$$\mathbf{r}(t) = t\,\hat{\mathbf{i}} + f(t)\,\hat{\mathbf{j}}.$$

We have

$$\mathbf{r}'(t) = \hat{\mathbf{i}} + f'(t)\,\hat{\mathbf{j}}$$

$$\mathbf{r}''(t) = f''(t)\,\hat{\mathbf{j}}.$$

Their cross product is just

$$r'(t) \times r''(t) = f''(t)\hat{\mathbf{k}}$$

which has magnitude

$$||\mathbf{r}'(t) \times r''(t)|| = |f''(t)|.$$

The curvature formula gives

Definition: Curvature of Plane Curve

$$K(t) = \frac{|f''(t)|}{\left[1 + (f'(t))^2\right]^{3/2}}.$$

Example 12.4.4

Find the curvature for the curve

$$y = \sin x$$

.

Solution

We have

$$f'(x) = \cos\, x$$

$$f''(x) = -\sin\, x.$$

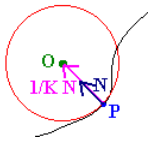
Plugging into the curvature formula gives

$$K(t) = \frac{|-\sin\, t|}{[1+\cos^2 t]^{3/2}}$$

The Osculating Circle

In first year calculus, we saw how to approximate a curve with a line, parabola, etc. Instead we can find the best fitting *circle* at the point on the curve. If P is a point on the curve, then the best fitting circle will have the same curvature as the curve and will pass through the point P. We will see that the curvature of a circle is a constant $1/r$, where r is the radius of the circle. The center of the osculating circle will be on the line containing the normal vector to the circle. In particular the center can be found by adding

$$OP + 1/KN.$$



Exercise 12.4.2

Find the equation of osculating circle to $y = x^2$ at $x = -1$.

Solution

The Normal Component of Acceleration Revisited

How is the normal component of acceleration related to the curvature. If you remember, the normal component the acceleration tells us how fast the particle is changing direction. If a curve has a sharp bend (high curvature) then the directional change will be faster. We now show that there is a definite relationship between the normal component of acceleration and curvature.

$$\mathbf{a}(t) = a_{\mathbf{T}}\mathbf{T}(t) + a_{\mathbf{N}}\mathbf{N}(t)$$

We have

$$\begin{aligned}
\mathbf{a}(t) &= \mathbf{r}''(t) \\
&= \frac{d}{dt}(\mathbf{r}'(t)) \\
&= \frac{d}{dt}(||\mathbf{r}'(t)||\mathbf{T}(t)) \\
&= \frac{d}{dt}(||r'(t)||)\mathbf{T}(t) + ||r'(t)||\mathbf{T}'(t)) \\
&= s''(t)\mathbf{T}(t) + s'\mathbf{T}'(t) \\
&= s''(t)\mathbf{T}(t) + s'||\mathbf{T}'(t)||\mathbf{N}(t) = s''(t)\mathbf{T}(t) + ks'^2\mathbf{N}(t).
\end{aligned}$$

So that the tangential component of the acceleration is $s''(t)$ and the normal component is $k(t)s'^2(t)$.

Exercise 12.4.3

Find the tangential and normal components of $\mathbf{r}(t) = t\,\hat{\mathbf{i}} - 2t\,\hat{\mathbf{j}} + t^2\,\hat{\mathbf{k}}$.

Contributors and Attributions

- Larry Green (Lake Tahoe Community College)
- Joseph Sanghun Lee (UCD)
- Integrated by Justin Marshall.

12.4: Curvature and Normal Vectors of a Curve is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **2.3: Curvature and Normal Vectors of a Curve** is licensed CC BY 4.0.

12.5: Tangential and Normal Components of Acceleration

Learning Objectives

- Describe the velocity and acceleration vectors of a particle moving in space.
- Explain the tangential and normal components of acceleration.
- State Kepler's laws of planetary motion.

We have now seen how to describe curves in the plane and in space, and how to determine their properties, such as arc length and curvature. All of this leads to the main goal of this chapter, which is the description of motion along plane curves and space curves. We now have all the tools we need; in this section, we put these ideas together and look at how to use them.

Motion Vectors in the Plane and in Space

Our starting point is using vector-valued functions to represent the position of an object as a function of time. All of the following material can be applied either to curves in the plane or to space curves. For example, when we look at the orbit of the planets, the curves defining these orbits all lie in a plane because they are elliptical. However, a particle traveling along a helix moves on a curve in three dimensions.

Definition: Speed, Velocity, and Acceleration

Let $\vec{\mathbf{r}}(t)$ be a twice-differentiable vector-valued function of the parameter t that represents the position of an object as a function of time.

The velocity vector $\vec{\mathbf{v}}(t)$ of the object is given by

$$\text{Velocity } = \vec{\mathbf{v}}(t) = \vec{\mathbf{r}}\,'(t). \tag{12.5.1}$$

The acceleration vector $\vec{\mathbf{a}}(t)$ is defined to be

$$\text{Acceleration } = \vec{\mathbf{a}}(t) = \vec{\mathbf{v}}\,'(t) = \vec{\mathbf{r}}\,''(t). \tag{12.5.2}$$

The *speed* is defined to be

$$\text{Speed } = v(t) = \|\vec{\mathbf{v}}(t)\| = \|\vec{\mathbf{r}}\,'(t)\| = \frac{ds}{dt}. \tag{12.5.3}$$

Since $\vec{\mathbf{r}}(t)$ can be in either two or three dimensions, these vector-valued functions can have either two or three components. In two dimensions, we define $\vec{\mathbf{r}}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}}$ and in three dimensions $\vec{\mathbf{r}}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}} + z(t)\hat{\mathbf{k}}$. Then the velocity, acceleration, and speed can be written as shown in the following table.

Table 12.5.1: Formulas for Position, Velocity, Acceleration, and Speed

Quantity	Two Dimensions	Three Dimensions
Position	$\vec{\mathbf{r}}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}}$	$\vec{\mathbf{r}}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}} + z(t)\hat{\mathbf{k}}$
Velocity	$\vec{\mathbf{v}}(t) = x'(t)\hat{\mathbf{i}} + y'(t)\hat{\mathbf{j}}$	$\vec{\mathbf{v}}(t) = x'(t)\hat{\mathbf{i}} + y'(t)\hat{\mathbf{j}} + z'(t)\hat{\mathbf{k}}$
Acceleration	$\vec{\mathbf{a}}(t) = x''(t)\hat{\mathbf{i}} + y''(t)\hat{\mathbf{j}}$	$\vec{\mathbf{a}}(t) = x''(t)\hat{\mathbf{i}} + y''(t)\hat{\mathbf{j}} + z''(t)\hat{\mathbf{k}}$
Speed	$\|\vec{\mathbf{v}}(t)\| = \sqrt{(x'(t))^2 + (y'(t))^2}$	$\|\vec{\mathbf{v}}(t)\| = \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2}$

Example 12.5.1: Studying Motion Along a Parabola

A particle moves in a parabolic path defined by the vector-valued function $\vec{\mathbf{r}}(t) = t^2\hat{\mathbf{i}} + \sqrt{5 - t^2}\,\hat{\mathbf{j}}$, where t measures time in seconds.

1. Find the velocity, acceleration, and speed as functions of time.
2. Sketch the curve along with the velocity vector at time $t = 1$.

Solution

1. We use Equations 12.5.1, 12.5.2, and 12.5.3:

$$\begin{aligned}\vec{\mathbf{v}}(t) &= \vec{\mathbf{r}}\,'(t) = 2t\hat{\mathbf{i}} - \frac{t}{\sqrt{5-t^2}}\hat{\mathbf{j}} \\ \vec{\mathbf{a}}(t) &= \vec{\mathbf{v}}\,'(t) = 2\hat{\mathbf{i}} - 5(5-t^2)^{-\frac{3}{2}}\hat{\mathbf{j}} \\ ||\vec{\mathbf{v}}(t)|| &= ||\vec{\mathbf{r}}\,'(t)|| \\ &= \sqrt{(2t)^2 + \left(-\frac{t}{\sqrt{5-t^2}}\right)^2} \\ &= \sqrt{4t^2 + \frac{t^2}{5-t^2}} \\ &= \sqrt{\frac{21t^2 - 4t^4}{5-t^2}}.\end{aligned}$$

2. The graph of $\vec{\mathbf{r}}(t) = t^2\hat{\mathbf{i}} + \sqrt{5-t^2}\,\hat{\mathbf{j}}$ is a portion of a parabola (Figure 12.5.1).

When $t = 1$, $\vec{\mathbf{r}}(1) = (1)^2\hat{\mathbf{i}} + \sqrt{5-(1)^2}\,\hat{\mathbf{j}} = \hat{\mathbf{i}} + \sqrt{4}\hat{\mathbf{j}} = \hat{\mathbf{i}} + 2\hat{\mathbf{j}}$.

Thus the particle would be located at the point $(1, 2)$ when $t = 1$.

The velocity vector at $t = 1$ is

$$\begin{aligned}\vec{\mathbf{v}}(1) &= \vec{\mathbf{r}}\,'(1) = 2(1)\hat{\mathbf{i}} - \frac{1}{\sqrt{5-1^2}}\hat{\mathbf{j}} \\ &= 2\hat{\mathbf{i}} - \frac{1}{2}\hat{\mathbf{j}}\end{aligned}$$

and the acceleration vector at $t = 1$ is

$$\vec{\mathbf{a}}(1) = \vec{\mathbf{v}}\,'(1) = 2\hat{\mathbf{i}} - 5(5-1^2)^{-3/2}\hat{\mathbf{j}} = 2\hat{\mathbf{i}} - \frac{5}{8}\hat{\mathbf{j}}.$$

Notice that the velocity vector is tangent to the path, as is always the case.

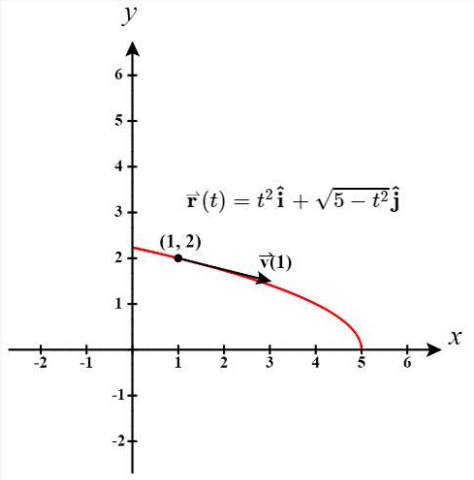


Figure 12.5.1: This graph depicts the velocity vector at time $t = 1$ for a particle moving in a parabolic path.

Exercise 12.5.1

A particle moves in a path defined by the vector-valued function $\vec{\mathbf{r}}(t) = (t^2 - 3t)\,\hat{\mathbf{i}} + (2t - 4)\,\hat{\mathbf{j}} + (t + 2)\,\hat{\mathbf{k}}$, where t measures time in seconds and where distance is measured in feet. Find the velocity, acceleration, and speed as functions of time.

Hint

Use Equations 12.5.1, 12.5.2, and 12.5.3

Answer

$$\vec{\mathbf{v}}(t) = \vec{\mathbf{r}}\,'(t) = (2t - 3)\,\hat{\mathbf{i}} + 2\,\hat{\mathbf{j}} + \hat{\mathbf{k}}$$

$$\vec{\mathbf{a}}(t) = \vec{\mathbf{v}}\,'(t) = 2\,\hat{\mathbf{i}}$$

$$||\vec{\mathbf{r}}\,'(t)|| = \sqrt{(2t - 3)^2 + 2^2 + 1^2} = \sqrt{4t^2 - 12t + 14}$$

The units for velocity and speed are feet per second, and the units for acceleration are feet per second squared.

To gain a better understanding of the velocity and acceleration vectors, imagine you are driving along a curvy road. If you do not turn the steering wheel, you would continue in a straight line and run off the road. The speed at which you are traveling when you run off the road, coupled with the direction, gives a vector representing your velocity, as illustrated in Figure 12.5.2

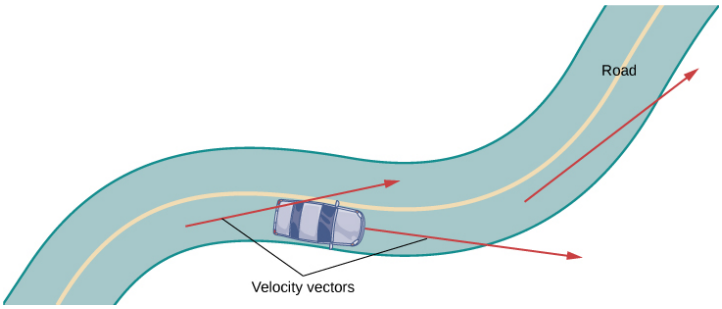


Figure 12.5.2: At each point along a road traveled by a car, the velocity vector of the car is tangent to the path traveled by the car.

However, the fact that you must turn the steering wheel to stay on the road indicates that your velocity is always changing (even if your speed is not) because your *direction* is constantly changing to keep you on the road. As you turn to the right, your acceleration vector also points to the right. As you turn to the left, your acceleration vector points to the left. This indicates that your velocity and acceleration vectors are constantly changing, regardless of whether your actual speed varies (Figure 12.5.3).

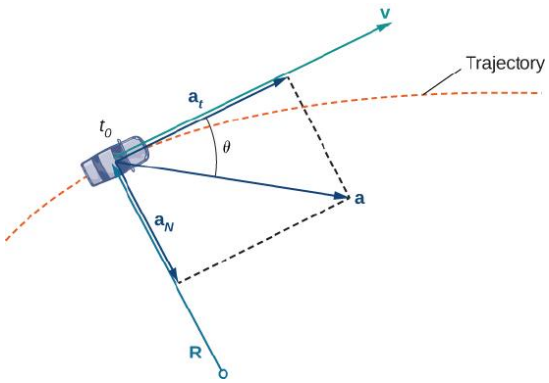


Figure 12.5.3: The dashed line represents the trajectory of an object (a car, for example). The acceleration vector points toward the inside of the turn at all times.

Components of the Acceleration Vector

We can combine some of the concepts discussed in Arc Length and Curvature with the acceleration vector to gain a deeper understanding of how this vector relates to motion in the plane and in space. Recall that the unit tangent vector $\vec{\mathbf{T}}$ and the unit

normal vector $\vec{\mathbf{N}}$ form an osculating plane at any point P on the curve defined by a vector-valued function $\vec{\mathbf{r}}(t)$. The following theorem shows that the acceleration vector $\vec{\mathbf{a}}(t)$ lies in the osculating plane and can be written as a linear combination of the unit tangent and the unit normal vectors.

Theorem 12.5.1: The Plane of the Acceleration Vector

The acceleration vector $\vec{\mathbf{a}}(t)$ of an object moving along a curve traced out by a twice-differentiable function $\vec{\mathbf{r}}(t)$ lies in the plane formed by the unit tangent vector $\vec{\mathbf{T}}(t)$ and the principal unit normal vector $\vec{\mathbf{N}}(t)$ to C. Furthermore,

$$\vec{\mathbf{a}}(t) = v'(t)\vec{\mathbf{T}}(t) + [v(t)]^2\kappa\vec{\mathbf{N}}(t)$$

Here, $v(t) = ||\vec{\mathbf{v}}(t)||$ is the speed of the object and κ is the curvature of C traced out by $\vec{\mathbf{r}}(t)$.

Proof

Because $\vec{\mathbf{v}}(t) = \vec{\mathbf{r}}\,'(t)$ and $\vec{\mathbf{T}}(t) = \dfrac{\vec{\mathbf{r}}\,'(t)}{||\vec{\mathbf{r}}\,'(t)||}$, we have $\vec{\mathbf{v}}(t) = ||\vec{\mathbf{r}}\,'(t)||\vec{\mathbf{T}}(t) = v(t)\vec{\mathbf{T}}(t)$.

Now we differentiate this equation:

$$\vec{\mathbf{a}}(t) = \vec{\mathbf{v}}\,'(t) = \frac{d}{dt}\Big(v(t)\vec{\mathbf{T}}(t)\Big) = v'(t)\vec{\mathbf{T}}(t) + v(t)\vec{\mathbf{T}}\,'(t)$$

Since $\vec{\mathbf{N}}(t) = \dfrac{\vec{\mathbf{T}}\,'(t)}{||\vec{\mathbf{T}}\,'(t)||}$, we know $\vec{\mathbf{T}}\,'(t) = ||\vec{\mathbf{T}}\,'(t)||\vec{\mathbf{N}}(t)$, so

$$\vec{\mathbf{a}}(t) = v'(t)\vec{\mathbf{T}}(t) + v(t)||\vec{\mathbf{T}}\,'(t)||\vec{\mathbf{N}}(t).$$

A formula for curvature is $\kappa = \dfrac{||\vec{\mathbf{T}}\,'(t)||}{||\vec{\mathbf{r}}\,'(t)||}$, so $\vec{\mathbf{T}}\,'(t) = \kappa||\vec{\mathbf{r}}\,'(t)|| = \kappa v(t)$.

This gives $\vec{\mathbf{a}}(t) = v'(t)\vec{\mathbf{T}}(t) + \kappa(v(t))^2\vec{\mathbf{N}}(t)$.

□

The coefficients of $\vec{\mathbf{T}}(t)$ and $\vec{\mathbf{N}}(t)$ are referred to as the **tangential component of acceleration** and the **normal component of acceleration**, respectively. We write $a_{\vec{\mathbf{T}}}$ to denote the tangential component and $a_{\vec{\mathbf{N}}}$ to denote the normal component.

Theorem 12.5.2: Tangential and Normal Components of Acceleration

Let $\vec{\mathbf{r}}(t)$ be a vector-valued function that denotes the position of an object as a function of time. Then $\vec{\mathbf{a}}(t) = \vec{\mathbf{r}}\,''(t)$ is the acceleration vector. The tangential and normal components of acceleration $a_{\vec{\mathbf{T}}}$ and $a_{\vec{\mathbf{N}}}$ are given by the formulas

$$a_{\vec{\mathbf{T}}} = \vec{\mathbf{a}} \cdot \vec{\mathbf{T}} = \frac{\vec{\mathbf{v}} \cdot \vec{\mathbf{a}}}{||\vec{\mathbf{v}}||} \tag{12.5.4}$$

and

$$a_{\vec{\mathbf{N}}} = \vec{\mathbf{a}} \cdot \vec{\mathbf{N}} = \frac{||\vec{\mathbf{v}} \times \vec{\mathbf{a}}||}{||\vec{\mathbf{v}}||} = \sqrt{||\vec{\mathbf{a}}||^2 - \left(a_{\vec{\mathbf{T}}}\right)^2}. \tag{12.5.5}$$

These components are related by the formula

$$\vec{\mathbf{a}}(t) = a_{\vec{\mathbf{T}}}\vec{\mathbf{T}}(t) + a_{\vec{\mathbf{N}}}\vec{\mathbf{N}}(t). \tag{12.5.6}$$

Here $\vec{\mathbf{T}}(t)$ is the unit tangent vector to the curve defined by $\vec{\mathbf{r}}(t)$, and $\vec{\mathbf{N}}(t)$ is the unit normal vector to the curve defined by $\vec{\mathbf{r}}(t)$.

The normal component of acceleration is also called the *centripetal component of acceleration* or sometimes the *radial component of acceleration*. To understand centripetal acceleration, suppose you are traveling in a car on a circular track at a constant speed. Then, as we saw earlier, the acceleration vector points toward the center of the track at all times. As a rider in the car, you feel a pull toward the *outside* of the track because you are constantly turning. This sensation acts in the opposite direction of centripetal acceleration. The same holds true for non-circular paths. The reason is that your body tends to travel in a straight line and resists the force resulting from acceleration that push it toward the side. Note that at point B in Figure $12.5.4$ the acceleration vector is pointing backward. This is because the car is decelerating as it goes into the curve.

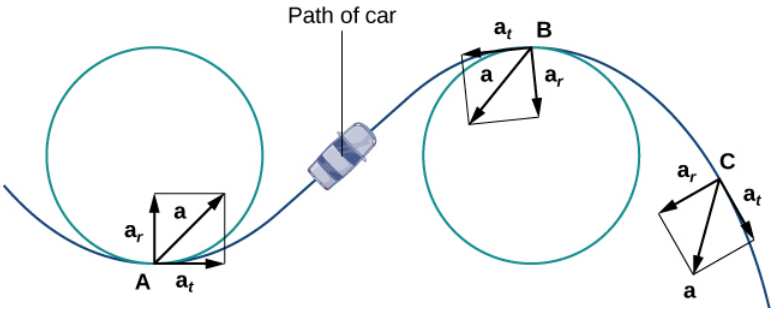


Figure 12.5.4: The tangential and normal components of acceleration can be used to describe the acceleration vector.

The tangential and normal unit vectors at any given point on the curve provide a frame of reference at that point. The tangential and normal components of acceleration are the projections of the acceleration vector onto $\vec{\mathbf{T}}$ and $\vec{\mathbf{N}}$, respectively.

Example 12.5.2: Finding Components of Acceleration

A particle moves in a path defined by the vector-valued function $\vec{\mathbf{r}}(t) = t^2\,\hat{\mathbf{i}} + (2t-3)\,\hat{\mathbf{j}} + (3t^2-3t)\,\hat{\mathbf{k}}$, where t measures time in seconds and distance is measured in feet.

a. Find $a_{\vec{\mathbf{T}}}$ and $a_{\vec{\mathbf{N}}}$ as functions of t.
b. Find $a_{\vec{\mathbf{T}}}$ and $a_{\vec{\mathbf{N}}}$ at time $t = 2$.

Solution

a. Let's start deriving the velocity and acceleration functions:

$$\begin{aligned}\vec{\mathbf{v}}(t) &= \vec{\mathbf{r}}\,'(t)\\ &= 2t\,\hat{\mathbf{i}} + 2\,\hat{\mathbf{j}} + (6t-3)\,\hat{\mathbf{k}}\\ \vec{\mathbf{a}}(t) &= \vec{\mathbf{v}}\,'(t)\\ &= 2\,\hat{\mathbf{i}} + 6\,\hat{\mathbf{k}}\end{aligned}$$

Now we apply Equation 12.5.4:

$$\begin{aligned}a_{\vec{\mathbf{T}}} &= \frac{\vec{\mathbf{v}}\cdot\vec{\mathbf{a}}}{||\vec{\mathbf{v}}||}\\ &= \frac{(2t\,\hat{\mathbf{i}} + 2\,\hat{\mathbf{j}} + (6t-3)\,\hat{\mathbf{k}})\cdot(2\,\hat{\mathbf{i}} + 6\,\hat{\mathbf{k}})}{||2t\,\hat{\mathbf{i}} + 2\,\hat{\mathbf{j}} + (6t-3)\,\hat{\mathbf{k}}||}\\ &= \frac{4t + 6(6t-3)}{\sqrt{(2t)^2 + 2^2 + (6t-3)^2}}\\ &= \frac{40t-18}{40t^2 - 36t + 13}\end{aligned}$$

Now we can apply Equation 12.5.5:

$$
\begin{aligned}
a_{\vec{\mathbf{N}}} &= \sqrt{||\vec{\mathbf{a}}||^2 - \left(a_{\vec{\mathbf{T}}}\right)^2} \\
&= \sqrt{||2\,\hat{\mathbf{i}} + 6\,\hat{\mathbf{k}}||^2 - \left(\frac{40t-18}{\sqrt{40t^2-36t+13}}\right)^2} \\
&= \sqrt{4+36-\frac{(40t-18)^2}{40t^2-36t+13}} \\
&= \sqrt{\frac{40(40t^2-36t+13)-(1600t^2-1440t+324)}{40t^2-36t+13}} \\
&= \sqrt{\frac{196}{40t^2-36t+13}} \\
&= \frac{14}{\sqrt{40t^2-36t+13}}
\end{aligned}
$$

b. We must evaluate each of the answers from part a at $t=2$:

$$
\begin{aligned}
a_{\vec{\mathbf{T}}}(2) &= \frac{40(2)-18}{\sqrt{40(2)^2-36(2)+13}} \\
&= \frac{80-18}{\sqrt{160-72+13}} \\
&= \frac{62}{\sqrt{101}} \\
a_{\vec{\mathbf{N}}}(2) &= \frac{14}{\sqrt{40(2)^2-36(2)+13}} \\
&= \frac{14}{\sqrt{160-72+13}} = \frac{14}{\sqrt{101}}.
\end{aligned}
$$

The units of acceleration are feet per second squared, as are the units of the normal and tangential components of acceleration.

Exercise 12.5.2

An object moves in a path defined by the vector-valued function $\vec{\mathbf{r}}(t) = 4t\,\hat{\mathbf{i}} + t^2\,\hat{\mathbf{j}}$, where t measures time in seconds.

a. Find $a_{\vec{\mathbf{T}}}$ and $a_{\vec{\mathbf{N}}}$ as functions of t.
b. Find $a_{\vec{\mathbf{T}}}$ and $a_{\vec{\mathbf{N}}}$ at time $t=-3$.

Hint

Use Equations 12.5.4and 12.5.5

Answer

a.

$$\begin{aligned} a_{\vec{\mathbf{T}}} &= \frac{\vec{\mathbf{v}}(t)\cdot\vec{\mathbf{a}}(t)}{||\vec{\mathbf{v}}(t)||} = \frac{\vec{\mathbf{r}}\,'(t)\cdot\vec{\mathbf{r}}\,''(t)}{||\vec{\mathbf{r}}\,'(t)||} \\ &= \frac{(4\,\hat{\mathbf{i}}+2t\,\hat{\mathbf{j}})\cdot(2\,\hat{\mathbf{j}})}{||4\,\hat{\mathbf{i}}+2t\,\hat{\mathbf{j}}||} \\ &= \frac{4t}{\sqrt{4^2+(2t)^2}} \\ &= \frac{2t}{\sqrt{2+t^2}} \end{aligned}$$

$$\begin{aligned} a_{\vec{\mathbf{N}}} &= \sqrt{||\vec{\mathbf{a}}||^2 - a_{\vec{\mathbf{T}}}^2} \\ &= \sqrt{||2\,\hat{\mathbf{j}}||^2 - \left(\frac{2t}{\sqrt{2+t^2}}\right)^2} \\ &= \sqrt{4-\frac{4t^2}{2+t^2}} \end{aligned}$$

b.

$$\begin{aligned} a_{\vec{\mathbf{T}}}(-3) &= \frac{2(-3)}{\sqrt{2+(-3)^2}} \\ &= \frac{-6}{\sqrt{11}} \end{aligned}$$

$$\begin{aligned} a_{\vec{\mathbf{N}}}(-3) &= \sqrt{4-\frac{4(-3)^2}{2+(-3)^2}} \\ &= \sqrt{4-\frac{36}{11}} \\ &= \sqrt{\frac{8}{11}} \\ &= \frac{2\sqrt{2}}{\sqrt{11}} \end{aligned}$$

Projectile Motion

Now let's look at an application of vector functions. In particular, let's consider the effect of gravity on the motion of an object as it travels through the air, and how it determines the resulting trajectory of that object. In the following, we ignore the effect of air resistance. This situation, with an object moving with an initial velocity but with no forces acting on it other than gravity, is known as projectile motion. It describes the motion of objects from golf balls to baseballs, and from arrows to cannonballs.

First we need to choose a coordinate system. If we are standing at the origin of this coordinate system, then we choose the positive y-axis to be up, the negative y-axis to be down, and the positive x-axis to be forward (i.e., away from the thrower of the object). The effect of gravity is in a downward direction, so Newton's second law tells us that the force on the object resulting from gravity is equal to the mass of the object times the acceleration resulting from gravity, or $\vec{\mathbf{F}}_g = m\vec{\mathbf{a}}$, where $\vec{\mathbf{F}}_g$ represents the force from gravity and $\vec{\mathbf{a}} = -g\,\hat{\mathbf{j}}$ represents the acceleration resulting from gravity at Earth's surface. The value of g in the English system of measurement is approximately 32 ft/sec^2 and it is approximately 9.8 m/sec^2 in the metric system. This is the only force acting on the object. Since gravity acts in a downward direction, we can write the force resulting from gravity in the form $\vec{\mathbf{F}}_g = -mg\,\hat{\mathbf{j}}$, as shown in Figure 12.5.5.

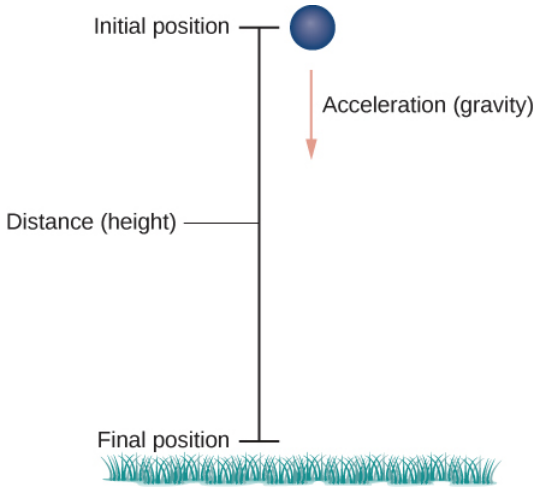


Figure 12.5.5: An object is falling under the influence of gravity.

Newton's second law also tells us that $F = m\vec{\mathbf{a}}$, where $\vec{\mathbf{a}}$ represents the acceleration vector of the object. This force must be equal to the force of gravity at all times, so we therefore know that

$$\begin{aligned} \vec{\mathbf{F}} &= \vec{\mathbf{F}}_g \\ m\vec{\mathbf{a}} &= -mg\,\hat{\mathbf{j}} \\ \vec{\mathbf{a}} &= -g\,\hat{\mathbf{j}}. \end{aligned}$$

Now we use the fact that the acceleration vector is the first derivative of the velocity vector. Therefore, we can rewrite the last equation in the form

$$\vec{\mathbf{v}}\,'(t) = -g\,\hat{\mathbf{j}}$$

By taking the antiderivative of each side of this equation we obtain

$$\vec{\mathbf{v}}(t) = \int -g\,\hat{\mathbf{j}}\; dt = -gt\,\hat{\mathbf{j}} + \vec{\mathbf{C}}_1$$

for some constant vector $\vec{\mathbf{C}}_1$. To determine the value of this vector, we can use the velocity of the object at a fixed time, say at time $t = 0$. We call this velocity the *initial velocity*: $\vec{\mathbf{v}}(0) = \vec{\mathbf{v}}_0$. Therefore, $\vec{\mathbf{v}}(0) = -g(0)\,\hat{\mathbf{j}} + \vec{\mathbf{C}}_1 = \vec{\mathbf{v}}_0$ and $\vec{\mathbf{C}}_1 = \vec{\mathbf{v}}_0$. This gives the velocity vector as $\vec{\mathbf{v}}(t) = -gt\,\hat{\mathbf{j}} + \vec{\mathbf{v}}_0$.

Next we use the fact that velocity $\vec{\mathbf{v}}(t)$ is the derivative of position $\vec{\mathbf{s}}(t)$. This gives the equation

$$\vec{\mathbf{s}}\,'(t) = -gt\,\hat{\mathbf{j}} + \vec{\mathbf{v}}_0.$$

Taking the antiderivative of both sides of this equation leads to

$$\begin{aligned} \vec{\mathbf{s}}(t) &= \int -gt\,\hat{\mathbf{j}} + \vec{\mathbf{v}}_0\; dt \\ &= -\frac{1}{2}gt^2\,\hat{\mathbf{j}} + \vec{\mathbf{v}}_0 t + \vec{\mathbf{C}}_2 \end{aligned}$$

with another unknown constant vector $\vec{\mathbf{C}}_2$. To determine the value of $\vec{\mathbf{C}}_2$, we can use the position of the object at a given time, say at time $t = 0$. We call this position the *initial position*: $\vec{\mathbf{s}}(0) = \vec{\mathbf{s}}_0$. Therefore, $\vec{\mathbf{s}}(0) = -(1/2)g(0)^2\,\hat{\mathbf{j}} + \vec{\mathbf{v}}_0(0) + \vec{\mathbf{C}}_2 = \vec{\mathbf{s}}_0$. This gives the position of the object at any time as

$$\vec{\mathbf{s}}(t) = -\frac{1}{2}gt^2\,\hat{\mathbf{j}} + \vec{\mathbf{v}}_0 t + \vec{\mathbf{s}}_0.$$

Let's take a closer look at the initial velocity and initial position. In particular, suppose the object is thrown upward from the origin at an angle θ to the horizontal, with initial speed $\vec{\mathbf{v}}_0$. How can we modify the previous result to reflect this scenario? First, we can

assume it is thrown from the origin. If not, then we can move the origin to the point from where it is thrown. Therefore, $\vec{\mathbf{s}}_0 = \vec{\mathbf{0}}$, as shown in Figure 12.5.6.

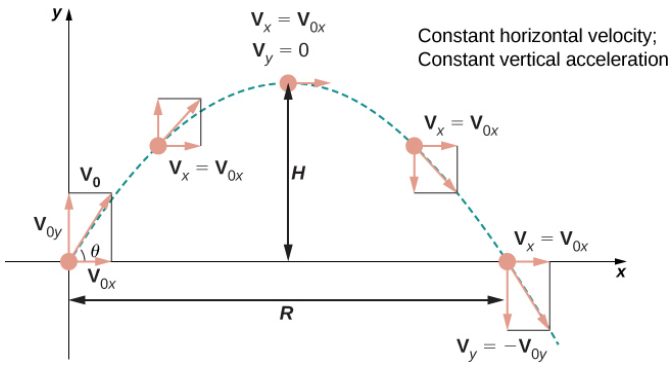


Figure 12.5.6: Projectile motion when the object is thrown upward at an angle θ. The horizontal motion is at constant velocity and the vertical motion is at constant acceleration.

We can rewrite the initial velocity vector in the form $\vec{\mathbf{v}}_0 = v_0 \cos\theta\,\hat{\mathbf{i}} + v_0 \sin\theta\,\hat{\mathbf{j}}$. Then the equation for the position function $\vec{\mathbf{s}}(t)$ becomes

$$\begin{aligned}\vec{\mathbf{s}}(t) &= -\frac{1}{2}gt^2\,\hat{\mathbf{j}} + v_0 t\cos\theta\,\hat{\mathbf{i}} + v_0 t\sin\theta\,\hat{\mathbf{j}} \\ &= v_0 t\cos\theta\,\hat{\mathbf{i}} + v_0 t\sin\theta\,\hat{\mathbf{j}} - \frac{1}{2}gt^2\,\hat{\mathbf{j}} \\ &= v_0 t\cos\theta\,\hat{\mathbf{i}} + \left(v_0 t\sin\theta - \frac{1}{2}gt^2\right)\,\hat{\mathbf{j}}.\end{aligned}$$

The coefficient of $\hat{\mathbf{i}}$ represents the horizontal component of $\vec{\mathbf{s}}(t)$ and is the horizontal distance of the object from the origin at time t. The maximum value of the horizontal distance (measured at the same initial and final altitude) is called the range R. The coefficient of $\hat{\mathbf{j}}$ represents the vertical component of $\vec{\mathbf{s}}(t)$ and is the altitude of the object at time t. The maximum value of the vertical distance is the height H.

Example 12.5.3: Motion of a Cannonball

During an Independence Day celebration, a cannonball is fired from a cannon on a cliff toward the water. The cannon is aimed at an angle of 30° above horizontal and the initial speed of the cannonball is 600 ft/sec. The cliff is 100 ft above the water (Figure 12.5.7).

a. Find the maximum height of the cannonball.
b. How long will it take for the cannonball to splash into the sea?
c. How far out to sea will the cannonball hit the water?

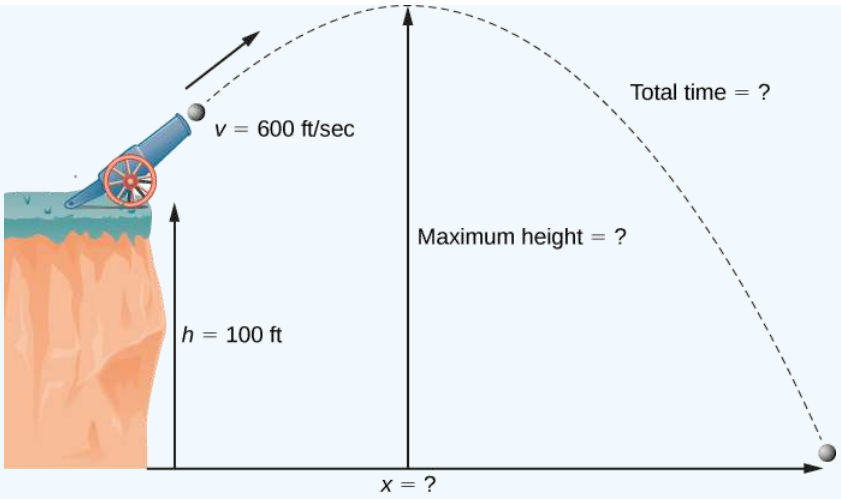


Figure 12.5.7: The flight of a cannonball (ignoring air resistance) is projectile motion.

Solution

We use the equation

$$\vec{\mathbf{s}}(t) = v_0 t\cos\theta\,\hat{\mathbf{i}} + \left(v_0 t\sin\theta - \frac{1}{2}gt^2\right)\,\hat{\mathbf{j}}$$

with $\theta = 30^\circ$, $g = 32\,\dfrac{\text{ft}}{\text{sec}^2}$, and $v_0 = 600\,\dfrac{\text{ft}}{\text{sec}}$. Then the position equation becomes

$$\begin{aligned}\vec{\mathbf{s}}(t) &= 600t(\cos 30^\circ)\,\hat{\mathbf{i}} + \left(600t\sin 30^\circ - \frac{1}{2}(32)t^2\right)\,\hat{\mathbf{j}}\\ &= 300t\sqrt{3}\,\hat{\mathbf{i}} + \left(300t - 16t^2\right)\,\hat{\mathbf{j}}\end{aligned}$$

a. The cannonball reaches its maximum height when the vertical component of its velocity is zero, because the cannonball is neither rising nor falling at that point. The velocity vector is

$$\begin{aligned}\vec{\mathbf{v}}(t) &= \vec{\mathbf{s}}\,'(t)\\ &= 300\sqrt{3}\,\hat{\mathbf{i}} + (300 - 32t)\,\hat{\mathbf{j}}\end{aligned}$$

Therefore, the vertical component of velocity is given by the expression $300 - 32t$. Setting this expression equal to zero and solving for t gives $t = 9.375$ sec. The height of the cannonball at this time is given by the vertical component of the position vector, evaluated at $t = 9.375$.

$$\begin{aligned}\vec{\mathbf{s}}(9.375) &= 300(9.375)\sqrt{3}\,\hat{\mathbf{i}} + (300(9.375) - 16(9.375)^2)\,\hat{\mathbf{j}}\\ &= 4871.39\,\hat{\mathbf{i}} + 1406.25\,\hat{\mathbf{j}}\end{aligned}$$

Therefore, the maximum height of the cannonball is 1406.39 ft above the cannon, or 1506.39 ft above sea level.

b. When the cannonball lands in the water, it is 100 ft below the cannon. Therefore, the vertical component of the position vector is equal to -100. Setting the vertical component of $\vec{\mathbf{s}}(t)$ equal to -100 and solving, we obtain

$$\begin{aligned} 300t - 16t^2 &= -100 \\ 16t^2 - 300t - 100 &= 0 \\ 4t^2 - 75 - 25 &= 0 \end{aligned}$$

$$\begin{aligned} t &= \frac{75 \pm \sqrt{(-75)^2 - 4(4)(-25)}}{2(4)} \\ &= \frac{75 \pm \sqrt{6025}}{8} \\ &= \frac{75 \pm 5\sqrt{241}}{8} \end{aligned}$$

The positive value of t that solves this equation is approximately 19.08. Therefore, the cannonball hits the water after approximately 19.08 sec.

c. To find the distance out to sea, we simply substitute the answer from part (b) into $\vec{\mathbf{s}}(t)$:

$$\begin{aligned} \vec{\mathbf{s}}(19.08) &= 300(19.08)\sqrt{3}\,\hat{\mathbf{i}} + \left(300(19.08) - 16(19.08)^2\right)\,\hat{\mathbf{j}} \\ &= 9914.26\,\hat{\mathbf{i}} - 100.7424\,\hat{\mathbf{j}} \end{aligned}$$

Therefore, the ball hits the water about 9914.26 ft away from the base of the cliff. Notice that the vertical component of the position vector is very close to −100, which tells us that the ball just hit the water. Note that 9914.26 feet is not the true range of the cannon since the cannonball lands in the ocean at a location below the cannon. The range of the cannon would be determined by finding how far out the cannonball is when its height is 100 ft above the water (the same as the altitude of the cannon).

? Exercise 12.5.3

An archer fires an arrow at an angle of 40° above the horizontal with an initial speed of 98 m/sec. The height of the archer is 171.5 cm. Find the horizontal distance the arrow travels before it hits the ground.

Hint

The equation for the position vector needs to account for the height of the archer in meters.

Answer

967.15 m

One final question remains: In general, what is the maximum distance a projectile can travel, given its initial speed? To determine this distance, we assume the projectile is fired from ground level and we wish it to return to ground level. In other words, we want to determine an equation for the range. In this case, the equation of projectile motion is

$$\vec{\mathbf{s}} = v_0 t\cos\theta\,\hat{\mathbf{i}} + \left(v_0 t\sin\theta - \frac{1}{2}gt^2\right)\,\hat{\mathbf{j}}.$$

Setting the second component equal to zero and solving for t yields

$$\begin{aligned} v_0 t\sin\theta - \frac{1}{2}gt^2 &= 0 \\ t\left(v_0\sin\theta - \frac{1}{2}gt\right) &= 0 \end{aligned}$$

Therefore, either $t = 0$ or $t = \dfrac{2v_0\sin\theta}{g}$. We are interested in the second value of t, so we substitute this into $\vec{\mathbf{s}}(t)$, which gives

$$\vec{\mathbf{s}}\left(\frac{2v_0\sin\theta}{g}\right) = v_0\left(\frac{2v_0\sin\theta}{g}\right)\cos\theta\,\hat{\mathbf{i}} + \left(v_0\left(\frac{2v_0\sin\theta}{g}\right)\sin\theta - \frac{1}{2}g\left(\frac{2v_0\sin\theta}{g}\right)^2\right)\hat{\mathbf{j}}$$
$$= \left(\frac{2v_0^2\sin\theta\cos\theta}{g}\right)\hat{\mathbf{i}}$$
$$= \frac{v_0^2\sin 2\theta}{g}\,\hat{\mathbf{i}}.$$

Thus, the expression for the range of a projectile fired at an angle θ is

$$R = \frac{v_0^2\sin 2\theta}{g}\,\hat{\mathbf{i}}.$$

The only variable in this expression is θ. To maximize the distance traveled, take the derivative of the coefficient of $\hat{\mathbf{i}}$ with respect to θ and set it equal to zero:

$$\frac{d}{d\theta}\left(\frac{v_0^2\sin 2\theta}{g}\right) = 0$$
$$\frac{2v_0^2\cos 2\theta}{g} = 0$$
$$\theta = 45^\circ$$

This value of θ) is the smallest positive value that makes the derivative equal to zero. Therefore, in the absence of air resistance, the best angle to fire a projectile (to maximize the range) is at a 45° angle. The distance it travels is given by

$$\vec{\mathbf{s}}\left(\frac{2v_0\sin 45^\circ}{g}\right) = \frac{v_0^2\sin 90^\circ}{g}\,\hat{\mathbf{i}} = \frac{v_0^2}{g}\,\hat{\mathbf{i}}$$

Therefore, the range for an angle of 45° is $\frac{v_0^2}{g}$ units.

Kepler's Laws

During the early 1600s, Johannes Kepler was able to use the amazingly accurate data from his mentor Tycho Brahe to formulate his three laws of planetary motion, now known as Kepler's laws of planetary motion. These laws also apply to other objects in the solar system in orbit around the Sun, such as comets (e.g., Halley's comet) and asteroids. Variations of these laws apply to satellites in orbit around Earth.

Theorem 12.5.2: Kepler's Laws of Planetary Motion

1. The path of any planet about the Sun is elliptical in shape, with the center of the Sun located at one focus of the ellipse (the law of ellipses).
2. A line drawn from the center of the Sun to the center of a planet sweeps out equal areas in equal time intervals (the law of equal areas) (Figure 12.5.8).
3. The ratio of the squares of the periods of any two planets is equal to the ratio of the cubes of the lengths of their semimajor orbital axes (the Law of Harmonies).

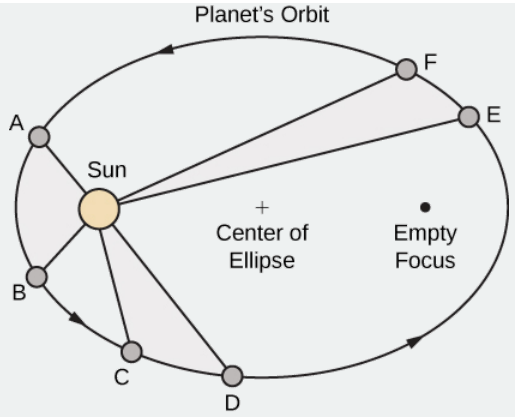


Figure 12.5.8: Kepler's first and second laws are pictured here. The Sun is located at a focus of the elliptical orbit of any planet. Furthermore, the shaded areas are all equal, assuming that the amount of time measured as the planet moves is the same for each region.

Kepler's third law is especially useful when using appropriate units. In particular, *1 astronomical unit* is defined to be the average distance from Earth to the Sun, and is now recognized to be 149,597,870,700 m or, approximately 93,000,000 mi. We therefore write 1 A.U. = 93,000,000 mi. Since the time it takes for Earth to orbit the Sun is 1 year, we use Earth years for units of time. Then, substituting 1 year for the period of Earth and 1 A.U. for the average distance to the Sun, Kepler's third law can be written as

$$T_p^2 = D_p^3$$

for any planet in the solar system, where T_P is the period of that planet measured in Earth years and D_P is the average distance from that planet to the Sun measured in astronomical units. Therefore, if we know the average distance from a planet to the Sun (in astronomical units), we can then calculate the length of its year (in Earth years), and vice versa.

Kepler's laws were formulated based on observations from Brahe; however, they were not proved formally until Sir Isaac Newton was able to apply calculus. Furthermore, Newton was able to generalize Kepler's third law to other orbital systems, such as a moon orbiting around a planet. Kepler's original third law only applies to objects orbiting the Sun.

Proof

Let's now prove Kepler's first law using the calculus of vector-valued functions. First we need a coordinate system. Let's place the Sun at the origin of the coordinate system and let the vector-valued function $\vec{\mathbf{r}}(t)$ represent the location of a planet as a function of time. Newton proved Kepler's law using his second law of motion and his law of universal gravitation. Newton's second law of motion can be written as $\vec{\mathbf{F}} = m\vec{\mathbf{a}}$, where $\vec{\mathbf{F}}$ represents the net force acting on the planet. His law of universal gravitation can be written in the form $\vec{\mathbf{F}} = -\dfrac{GmM}{||\vec{\mathbf{r}}||^2} \cdot \dfrac{\vec{\mathbf{r}}}{||\vec{\mathbf{r}}||}$, which indicates that the force resulting from the gravitational attraction of the Sun points back toward the Sun, and has magnitude $\dfrac{GmM}{||\vec{\mathbf{r}}||^2}$ (Figure 12.5.9).

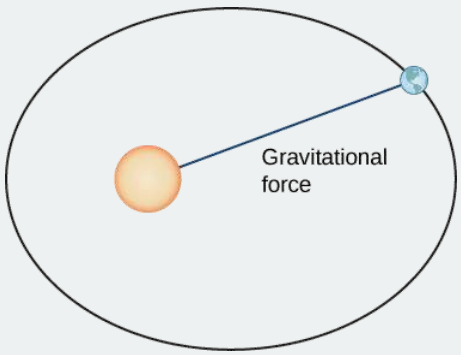


Figure 12.5.9: The gravitational force between Earth and the Sun is equal to the mass of the earth times its acceleration.

Setting these two forces equal to each other, and using the fact that $\vec{\mathbf{a}}(t) = \vec{\mathbf{v}}\,'(t)$, we obtain

$$m\vec{\mathbf{v}}\,'(t) = -\frac{GmM}{\|\vec{\mathbf{r}}\|^2} \cdot \frac{\vec{\mathbf{r}}}{\|\vec{\mathbf{r}}\|},$$

which can be rewritten as

$$\frac{d\vec{\mathbf{v}}}{dt} = -\frac{GM}{||\vec{\mathbf{r}}||^3}\vec{\mathbf{r}}.$$

This equation shows that the vectors $d\vec{\mathbf{v}}/dt$ and $\vec{\mathbf{r}}$ are parallel to each other, so $d\vec{\mathbf{v}}/dt \times \vec{\mathbf{r}} = \vec{\mathbf{0}}$. Next, let's differentiate $\vec{\mathbf{r}} \times \vec{\mathbf{v}}$ with respect to time:

$$\frac{d}{dt}(\vec{\mathbf{r}} \times \vec{\mathbf{v}}) = \frac{d\vec{\mathbf{r}}}{dt} \times \vec{\mathbf{v}} + \vec{\mathbf{r}} \times \frac{d\vec{\mathbf{v}}}{dt} = \vec{\mathbf{v}} \times \vec{\mathbf{v}} + \vec{\mathbf{0}} = \vec{\mathbf{0}}. \tag{12.5.7}$$

This proves that $\vec{\mathbf{r}} \times \vec{\mathbf{v}}$ is a constant vector, which we call $\vec{\mathbf{C}}$. Since $\vec{\mathbf{r}}$ and $\vec{\mathbf{v}}$ are both perpendicular to $\vec{\mathbf{C}}$ for all values of t, they must lie in a plane perpendicular to $\vec{\mathbf{C}}$. Therefore, the motion of the planet lies in a plane.

Next we calculate the expression $d\vec{\mathbf{v}}/dt \times \vec{\mathbf{C}}$:

$$\frac{d\vec{\mathbf{v}}}{dt} \times \vec{\mathbf{C}} = -\frac{GM}{||\vec{\mathbf{r}}||^3}\vec{\mathbf{r}} \times (\vec{\mathbf{r}} \times \vec{\mathbf{v}}) = -\frac{GM}{||\vec{\mathbf{r}}||^3}[(\vec{\mathbf{r}} \cdot \vec{\mathbf{v}})\vec{\mathbf{r}} - (\vec{\mathbf{r}} \cdot \vec{\mathbf{r}})\vec{\mathbf{v}}]. \tag{12.5.8}$$

The last equality in Equation 12.5.8 is from the triple cross product formula (see the cross product section in Introduction to Vectors in Space). We need an expression for $\vec{\mathbf{r}} \cdot \vec{\mathbf{v}}$. To calculate this, we differentiate $\vec{\mathbf{r}} \cdot \vec{\mathbf{r}}$ with respect to time:

$$\frac{d}{dt}(\vec{\mathbf{r}} \cdot \vec{\mathbf{r}}) = \frac{d\vec{\mathbf{r}}}{dt} \cdot \vec{\mathbf{r}} + \vec{\mathbf{r}} \cdot \frac{d\vec{\mathbf{r}}}{dt} = 2\vec{\mathbf{r}} \cdot \frac{d\vec{\mathbf{r}}}{dt} = 2\vec{\mathbf{r}} \cdot \vec{\mathbf{v}}. \tag{12.5.9}$$

Since $\vec{\mathbf{r}} \cdot \vec{\mathbf{r}} = ||\vec{\mathbf{r}}||^2$, we also have

$$\frac{d}{dt}(\vec{\mathbf{r}} \cdot \vec{\mathbf{r}}) = \frac{d}{dt}||\vec{\mathbf{r}}||^2 = 2||\vec{\mathbf{r}}||\frac{d}{dt}||\vec{\mathbf{r}}||. \tag{12.5.10}$$

Combining Equation 12.5.9 and Equation 12.5.10, we get

$$2\vec{\mathbf{r}} \cdot \vec{\mathbf{v}} = 2||\vec{\mathbf{r}}||\frac{d}{dt}||\vec{\mathbf{r}}||$$
$$\vec{\mathbf{r}} \cdot \vec{\mathbf{v}} = ||\vec{\mathbf{r}}||\frac{d}{dt}||\vec{\mathbf{r}}||.$$

Substituting this into Equation 12.5.8 gives us

$$\begin{aligned}\frac{d\vec{\mathbf{v}}}{dt} \times \vec{\mathbf{C}} &= -\frac{GM}{||\vec{\mathbf{r}}||^3}[(\vec{\mathbf{r}} \cdot \vec{\mathbf{v}})\vec{\mathbf{r}} - (\vec{\mathbf{r}} \cdot \vec{\mathbf{r}})\vec{\mathbf{v}}] \\ &= -\frac{GM}{||\vec{\mathbf{r}}||^3}\left[||\vec{\mathbf{r}}||\left(\frac{d}{dt}||\vec{\mathbf{r}}||\right)\vec{\mathbf{r}} - ||\vec{\mathbf{r}}||^2\vec{\mathbf{v}}\right] \\ &= -GM\left[\frac{1}{||\vec{\mathbf{r}}||^2}\left(\frac{d}{dt}||\vec{\mathbf{r}}||\right)\vec{\mathbf{r}} - \frac{1}{||\vec{\mathbf{r}}||}\vec{\mathbf{v}}\right] \\ &= GM\left[\frac{\vec{\mathbf{v}}}{||\vec{\mathbf{r}}||} - \frac{\vec{\mathbf{r}}}{||\vec{\mathbf{r}}||^2}\left(\frac{d}{dt}||\vec{\mathbf{r}}||\right)\right].\end{aligned} \tag{12.5.11}$$

However,

$$\begin{aligned}\frac{d}{dt}\frac{\vec{\mathbf{r}}}{||\vec{\mathbf{r}}||} &= \frac{\frac{d}{dt}(\vec{\mathbf{r}})||\vec{\mathbf{r}}|| - \vec{\mathbf{r}}\frac{d}{dt}||\vec{\mathbf{r}}||}{||\vec{\mathbf{r}}||^2} \\ &= \frac{\frac{d\vec{\mathbf{r}}}{dt}}{||\vec{\mathbf{r}}||} - \frac{\vec{\mathbf{r}}}{||\vec{\mathbf{r}}||^2}\frac{d}{dt}||\vec{\mathbf{r}}|| \\ &= \frac{\vec{\mathbf{v}}}{||\vec{\mathbf{r}}||} - \frac{\vec{\mathbf{r}}}{||\vec{\mathbf{r}}||^2}\frac{d}{dt}||\vec{\mathbf{r}}||.\end{aligned}$$

Therefore, Equation 12.5.11becomes

$$\frac{d\vec{\mathbf{v}}}{dt} \times \vec{\mathbf{C}} = GM\left(\frac{d}{dt}\frac{\vec{\mathbf{r}}}{||\vec{\mathbf{r}}||}\right).$$

Since $\vec{\mathbf{C}}$ is a constant vector, we can integrate both sides and obtain

$$\vec{\mathbf{v}} \times \vec{\mathbf{C}} = GM\frac{\vec{\mathbf{r}}}{||\vec{\mathbf{r}}||} + \vec{\mathbf{D}},$$

where $\vec{\mathbf{D}}$ is a constant vector. Our goal is to solve for $||\vec{\mathbf{r}}||$. Let's start by calculating $\vec{\mathbf{r}} \cdot \left(\vec{\mathbf{v}} \times \vec{\mathbf{C}}\right)$:

$$\vec{\mathbf{r}} \cdot \left(\vec{\mathbf{v}} \times \vec{\mathbf{C}}\right) = GM\frac{||\vec{\mathbf{r}}||^2}{||\vec{\mathbf{r}}||} + \vec{\mathbf{r}} \cdot \vec{\mathbf{D}} = GM||\vec{\mathbf{r}}|| + \vec{\mathbf{r}} \cdot \vec{\mathbf{D}}.$$

However, $\vec{\mathbf{r}} \cdot (\vec{\mathbf{v}} \times \vec{\mathbf{C}}) = (\vec{\mathbf{r}} \times \vec{\mathbf{v}}) \cdot \vec{\mathbf{C}}$, so

$$(\vec{\mathbf{r}} \times \vec{\mathbf{v}}) \cdot \vec{\mathbf{C}} = GM||\vec{\mathbf{r}}|| + \vec{\mathbf{r}} \cdot \vec{\mathbf{D}}.$$

Since $\vec{\mathbf{r}} \times \vec{\mathbf{v}} = \vec{\mathbf{C}}$, we have

$$||\vec{\mathbf{C}}||^2 = GM||\vec{\mathbf{r}}|| + \vec{\mathbf{r}} \cdot \vec{\mathbf{D}}.$$

Note that $\vec{\mathbf{r}} \cdot \vec{\mathbf{D}} = ||\vec{\mathbf{r}}||||\vec{\mathbf{D}}|| \cos\theta$, where θ is the angle between $\vec{\mathbf{r}}$ and $\vec{\mathbf{D}}$. Therefore,

$$||\vec{\mathbf{C}}||^2 = GM||\vec{\mathbf{r}}|| + ||\vec{\mathbf{r}}||||\vec{\mathbf{D}}|| \cos\theta$$

Solving for $||\vec{\mathbf{r}}||$,

$$||\vec{\mathbf{r}}|| = \frac{||\vec{\mathbf{C}}||^2}{GM + ||\vec{\mathbf{D}}|| \cos\theta} = \frac{||\vec{\mathbf{C}}||^2}{GM}\left(\frac{1}{1 + e\cos\theta}\right).$$

where $e = ||\vec{\mathbf{D}}||/GM$. This is the polar equation of a conic with a focus at the origin, which we set up to be the Sun. It is a hyperbola if $e > 1$, a parabola if $e = 1$, or an ellipse if $e < 1$. Since planets have closed orbits, the only possibility is an ellipse. However, at this point it should be mentioned that hyperbolic comets do exist. These are objects that are merely passing through the solar system at speeds too great to be trapped into orbit around the Sun. As they pass close enough to the Sun, the gravitational field of the Sun deflects the trajectory enough so the path becomes hyperbolic.

□

Kepler's third law of planetary motion can be modified to the case of one object in orbit around an object other than the Sun, such as the Moon around the Earth. In this case, Kepler's third law becomes

$$P^2 = \frac{4\pi^2 a^3}{G(m+M)}, \tag{12.5.12}$$

where *m* is the mass of the Moon and *M* is the mass of Earth, *a* represents the length of the major axis of the elliptical orbit, and *P* represents the period.

Example 12.5.4: Using Kepler's Third Law for Nonheliocentric Orbits

Given that the mass of the Moon is 7.35×10^{22} kg, the mass of Earth is 5.97×10^{24} kg, $G = 6.67 \times 10^{-11} \text{m/kg} \cdot \text{sec}^2$, and the period of the moon is 27.3 days, let's find the length of the major axis of the orbit of the Moon around Earth.

Solution

It is important to be consistent with units. Since the universal gravitational constant contains seconds in the units, we need to use seconds for the period of the Moon as well:

$$27.3 \text{ days} \times \frac{24 \text{ hr}}{1 \text{ day}} \times \frac{3600 \text{ esc}}{1 \text{ hour}} = 2,358,720 \text{ sec}$$

Substitute all the data into Equation 12.5.12and solve for a:

$$\begin{aligned} (2,358,720 sec)^2 &= \frac{4\pi^2 a^3}{\left(6.67 \times 10^{-11} \frac{m}{\text{kg} \times \text{sec}^2}\right)(7.35 \times 10^{22}\text{kg} + 5.97 \times 10^{24}\text{kg})} \\ 5.563 \times 10^{12} &= \frac{4\pi^2 a^3}{(6.67 \times 10^{-11}\text{m}^3)(6.04 \times 10^{24})} \\ (5.563 \times 10^{12})(6.67 \times 10^{-11}\text{m}^3)(6.04 \times 10^{24}) &= 4\pi^2 a^3 \\ a^3 &= \frac{2.241 \times 10^{27}}{4\pi^2}\text{m}^3 \\ a &= 3.84 \times 10^8 \text{m} \\ &\approx 384,000 \text{ km}. \end{aligned}$$

Analysis

According to solarsystem.nasa.gov, the actual average distance from the Moon to Earth is 384,400 km. This is calculated using reflectors left on the Moon by Apollo astronauts back in the 1960s.

Exercise 12.5.4

Titan is the largest moon of Saturn. The mass of Titan is approximately $1.35 \times 10^{23} kg$. The mass of Saturn is approximately 5.68×10^{26} kg. Titan takes approximately 16 days to orbit Saturn. Use this information, along with the universal gravitation constant $G = 6.67 \times 10^{-11} \text{m/kg} \cdot \text{sec}^2$ to estimate the distance from Titan to Saturn.

Hint

Make sure your units agree, then use Equation 12.5.12

Answer

$$a \approx 1.224 \times 10^9 \text{m} = 1,224,000 \text{km}$$

Example 12.5.5: Halley's Comet

We now return to the chapter opener, which discusses the motion of Halley's comet around the Sun. Kepler's first law states that Halley's comet follows an elliptical path around the Sun, with the Sun as one focus of the ellipse. The period of Halley's comet is approximately 76.1 years, depending on how closely it passes by Jupiter and Saturn as it passes through the outer solar system. Let's use $T = 76.1$ years. What is the average distance of Halley's comet from the Sun?



Solution

Using the equation $T^2 = D^3$ with $T = 76.1$, we obtain $D^3 = 5791.21$, so $D \approx 17.96$ A.U. This comes out to approximately 1.67×10^9 mi.

A natural question to ask is: What are the maximum (aphelion) and minimum (perihelion) distances from Halley's Comet to the Sun? The eccentricity of the orbit of Halley's Comet is 0.967 (Source: http://nssdc.gsfc.nasa.gov/planetary...cometfact.html). Recall that the formula for the eccentricity of an ellipse is $e = c/a$, where a is the length of the semimajor axis and c is the distance from the center to either focus. Therefore, $0.967 = c/17.96$ and $c \approx 17.37$ A.U. Subtracting this from a gives the perihelion distance $p = a - c = 17.96 - 17.37 = 0.59$ A.U. According to the National Space Science Data Center (Source: http://nssdc.gsfc.nasa.gov/planetary...cometfact.html), the perihelion distance for Halley's comet is 0.587 A.U. To calculate the aphelion distance, we add

$$P = a + c = 17.96 + 17.37 = 35.33 \text{ A.U.}$$

This is approximately 3.3×10^9 mi. The average distance from Pluto to the Sun is 39.5 A.U. (Source: http://www.oarval.org/furthest.htm), so it would appear that Halley's Comet stays just within the orbit of Pluto.

NAVIGATING A BANKED TURN

How fast can a racecar travel through a circular turn without skidding and hitting the wall? The answer could depend on several factors:

- The weight of the car;
- The friction between the tires and the road;
- The radius of the circle;
- The "steepness" of the turn.

In this project we investigate this question for NASCAR racecars at the Bristol Motor Speedway in Tennessee. Before considering this track in particular, we use vector functions to develop the mathematics and physics necessary for answering questions such as this.

A car of mass m moves with constant angular speed ω around a circular curve of radius R (Figure 12.5.9). The curve is banked at an angle θ. If the height of the car off the ground is h, then the position of the car at time t is given by the function $\vec{\mathbf{r}}(t) = < R\cos(\omega t), R\sin(\omega t), h >$.

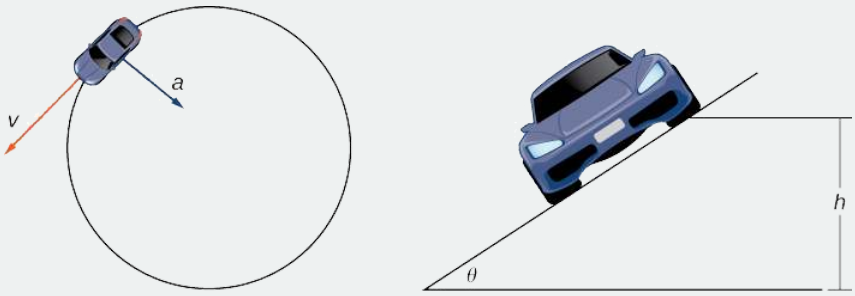


Figure 12.5.9: Views of a race car moving around a track.

1. Find the velocity function $\vec{\mathbf{v}}(t)$ of the car. Show that $\vec{\mathbf{v}}$ is tangent to the circular curve. This means that, without a force to keep the car on the curve, the car will shoot off of it.
2. Show that the speed of the car is ωR. Use this to show that $(2\pi 4)/\|\vec{\mathbf{v}}\| = (2\pi)/\omega$
3. Find the acceleration $\vec{\mathbf{a}}$. Show that this vector points toward the center of the circle and that $\|\vec{\mathbf{a}}\| = R\omega^2$.
4. The force required to produce this circular motion is called the *centripetal force*, and it is denoted $\vec{\mathbf{F}}_{cent}$. This force points toward the center of the circle (not toward the ground). Show that $\|\vec{\mathbf{F}}_{cent}\| = \left(m|\vec{\mathbf{v}}|^2\right)/R$.

As the car moves around the curve, three forces act on it: gravity, the force exerted by the road (this force is perpendicular to the ground), and the friction force (Figure 12.5.10). Because describing the frictional force generated by the tires and the road is complex, we use a standard approximation for the frictional force. Assume that $\vec{\mathbf{f}} = \mu\vec{\mathbf{N}}$ for some positive constant μ. The constant μ is called the *coefficient of friction*.

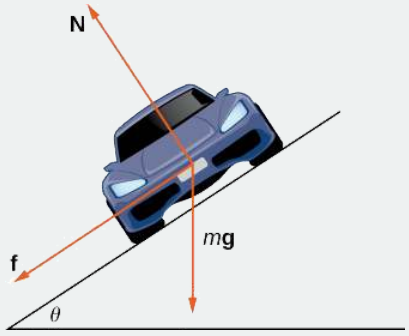


Figure 12.5.10: The car has three forces acting on it: gravity (denoted by $m\vec{\mathbf{g}}$), *the friction force* $\vec{\mathbf{f}}$, *and the force exerted by the road* $\vec{\mathbf{N}}$.

Let v_{max} denote the maximum speed the car can attain through the curve without skidding. In other words, v_{max} is the fastest speed at which the car can navigate the turn. When the car is traveling at this speed, the magnitude of the centripetal force is

$$\|\vec{\mathbf{F}}_{cent}\| = \frac{m(v_{max})^2}{R}.$$

The next three questions deal with developing a formula that relates the speed v_{max} to the banking angle θ.

5. Show that $\vec{\mathbf{N}}\cos\theta = m\vec{\mathbf{g}} + \vec{\mathbf{f}}\sin\theta$. Conclude that $\vec{\mathbf{N}} = (m\vec{\mathbf{g}})/(\cos\theta - \mu\sin\theta)$.
6. The centripetal force is the sum of the forces in the horizontal direction, since the centripetal force points toward the center of the circular curve. Show that

   $$\vec{\mathbf{F}}_{cent} = \vec{\mathbf{N}}\sin\theta + \vec{\mathbf{f}}\cos\theta.$$

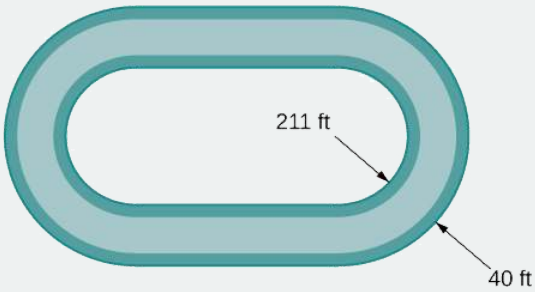
   Conclude that

   $$\vec{\mathbf{F}}_{cent} = \frac{\sin\theta + \mu\cos\theta}{cos\theta - \mu\sin\theta} m\vec{\mathbf{g}}.$$

7. Show that $(v_{\max})^2 = ((\sin\theta + \mu\, cos\theta)/(\cos\theta - \mu\sin\theta))gR$. Conclude that the maximum speed does not actually depend on the mass of the car.
   Now that we have a formula relating the maximum speed of the car and the banking angle, we are in a position to answer the questions like the one posed at the beginning of the project.
   The Bristol Motor Speedway is a NASCAR short track in Bristol, Tennessee. The track has the approximate shape shown in Figure 12.5.11. Each end of the track is approximately semicircular, so when cars make turns they are traveling along an approximately circular curve. If a car takes the inside track and speeds along the bottom of turn 1, the car travels along a semicircle of radius approximately 211 ft with a banking angle of 24°. If the car decides to take the outside track and speeds along the top of turn 1, then the car travels along a semicircle with a banking angle of 28°. (The track has variable angle banking.)



(a)



(b)

Figure 12.5.11: At the Bristol Motor Speedway, Bristol, Tennessee (a), the turns have an inner radius of about 211 ft and a width of 40 ft (b). (credit: part (a) photo by Raniel Diaz, Flickr)

The coefficient of friction for a normal tire in dry conditions is approximately 0.7. Therefore, we assume the coefficient for a NASCAR tire in dry conditions is approximately 0.98.

Before answering the following questions, note that it is easier to do computations in terms of feet and seconds, and then convert the answers to miles per hour as a final step.

8. In dry conditions, how fast can the car travel through the bottom of the turn without skidding?
9. In dry conditions, how fast can the car travel through the top of the turn without skidding?
10. In wet conditions, the coefficient of friction can become as low as 0.1. If this is the case, how fast can the car travel through the bottom of the turn without skidding?
11. Suppose the measured speed of a car going along the outside edge of the turn is 105 mph. Estimate the coefficient of friction for the car's tires.

Key Concepts

- If $\vec{\mathbf{r}}(t)$ represents the position of an object at time t, then $\vec{\mathbf{r}}\,'(t)$ represents the velocity and $\vec{\mathbf{r}}\,''(t)$ represents the acceleration of the object at time t. The magnitude of the velocity vector is speed.
- The acceleration vector always points toward the concave side of the curve defined by $\vec{\mathbf{r}}(t)$. The tangential and normal components of acceleration $a_{\vec{\mathbf{T}}}$ and $a_{\vec{\mathbf{N}}}$ are the projections of the acceleration vector onto the unit tangent and unit normal vectors to the curve.
- Kepler's three laws of planetary motion describe the motion of objects in orbit around the Sun. His third law can be modified to describe motion of objects in orbit around other celestial objects as well.
- Newton was able to use his law of universal gravitation in conjunction with his second law of motion and calculus to prove Kepler's three laws.

Key Equations

- **Velocity**

$$\vec{\mathbf{v}}(t) = \vec{\mathbf{r}}\,'(t)$$

- **Acceleration**

$$\vec{\mathbf{a}}(t) = \vec{\mathbf{v}}\,'(t) = \vec{\mathbf{r}}\,''(t)$$

- **Speed**

$$v(t) = ||\vec{\mathbf{v}}(t)|| = ||\vec{\mathbf{r}}\,'(t)|| = \frac{ds}{dt}$$

- **Tangential component of acceleration**

$$a_{\vec{\mathbf{T}}} = \vec{\mathbf{a}} \cdot \vec{\mathbf{T}} = \frac{\vec{\mathbf{v}} \cdot \vec{\mathbf{a}}}{||\vec{\mathbf{v}}||}$$

- **Normal component of acceleration**

$$a_{\vec{\mathbf{N}}} = \vec{\mathbf{a}} \cdot \vec{\mathbf{N}} = \frac{||\vec{\mathbf{v}} \times \vec{\mathbf{a}}||}{||\vec{\mathbf{v}}||} = \sqrt{||\vec{\mathbf{a}}||^2 - \left(a_{\vec{\mathbf{T}}}\right)^2}$$

Glossary

acceleration vector
the second derivative of the position vector

Kepler's laws of planetary motion
three laws governing the motion of planets, asteroids, and comets in orbit around the Sun

normal component of acceleration
the coefficient of the unit normal vector $\vec{\mathbf{N}}$ when the acceleration vector is written as a linear combination of $\vec{\mathbf{T}}$ and $\vec{\mathbf{N}}$

projectile motion
motion of an object with an initial velocity but no force acting on it other than gravity

tangential component of acceleration
the coefficient of the unit tangent vector $\vec{\mathbf{T}}$ when the acceleration vector is written as a linear combination of $\vec{\mathbf{T}}$ and $\vec{\mathbf{N}}$

velocity vector
the derivative of the position vector

- Gilbert Strang (MIT) and Edwin "Jed" Herman (Harvey Mudd) with many contributing authors. This content by OpenStax is licensed with a CC-BY-SA-NC 4.0 license. Download for free at http://cnx.org.
- Edited by Paul Seeburger
  Paul Seeburger added finding point $(1, 2)$ when $t = 1$ in Example $12.5.1$.
  He also created Figure $12.5.1$.

12.5: Tangential and Normal Components of Acceleration is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **13.4: Motion in Space** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

12.6: Velocity and Acceleration in Polar Coordinates

http://faculty.etsu.edu/gardnerr/211...-12e/c13s6.pdf

https://ecourses.ou.edu/cgi-bin/eboo....6&page=theory

http://pioneer.netserv.chula.ac.th/~...1/24rtheta.pdf

http://math.stackexchange.com/questi...ar-coordinates

12.6: Velocity and Acceleration in Polar Coordinates is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

13: Partial Derivatives

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

13.1: Functions of Several Variables
13.2: Limits and Continuity in Higher Dimensions
13.3: Partial Derivatives
13.4: The Chain Rule
13.5: Directional Derivatives and Gradient Vectors
13.6: Tangent Planes and Differentials
13.7: Extreme Values and Saddle Points
13.8: Lagrange Multipliers

13: Partial Derivatives is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

13.1: Functions of Several Variables

Learning Objectives

- Recognize a function of two variables and identify its domain and range.
- Sketch a graph of a function of two variables.
- Sketch several traces or level curves of a function of two variables.
- Recognize a function of three or more variables and identify its level surfaces.

Our first step is to explain what a function of more than one variable is, starting with functions of two independent variables. This step includes identifying the domain and range of such functions and learning how to graph them. We also examine ways to relate the graphs of functions in three dimensions to graphs of more familiar planar functions.

Functions of Two Variables

The definition of a function of two variables is very similar to the definition for a function of one variable. The main difference is that, instead of mapping values of one variable to values of another variable, we map ordered pairs of variables to another variable.

Definition: Function of Two Variables

A *function of two variables* $z = f(x, y)$ maps each ordered pair (x, y) in a subset D of the real plane R^2 to a unique real number z. The set D is called the domain of the function. The range of f is the set of all real numbers z that has at least one ordered pair $(x, y) \in D$ such that $f(x, y) = z$ as shown in Figure 13.1.1.

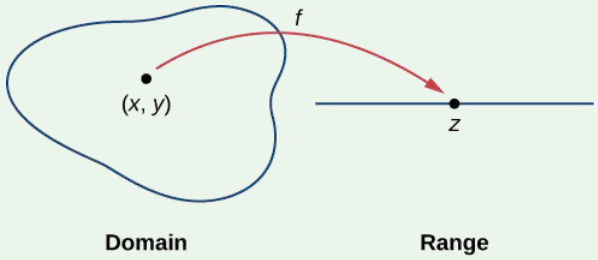


Figure 13.1.1: The domain of a function of two variables consists of ordered pairs (x, y).

Determining the domain of a function of two variables involves taking into account any domain restrictions that may exist. Let's take a look.

Example 13.1.1: Domains and Ranges for Functions of Two Variables

Find the domain and range of each of the following functions:

a. $f(x, y) = 3x + 5y + 2$
b. $g(x, y) = \sqrt{9 - x^2 - y^2}$

Solution

a. This is an example of a linear function in two variables. There are no values or combinations of x and y that cause $f(x, y)$ to be undefined, so the domain of f is R^2. To determine the range, first pick a value for z. We need to find a solution to the equation $f(x, y) = z$, or $3x - 5y + 2 = z$. One such solution can be obtained by first setting $y = 0$, which yields the equation $3x + 2 = z$. The solution to this equation is $x = \dfrac{z-2}{3}$, which gives the ordered pair $\left(\dfrac{z-2}{3}, 0\right)$ as a solution to the equation $f(x, y) = z$ for any value of z. Therefore, the range of the function is all real numbers, or R.

b. For the function $g(x, y)$ to have a real value, the quantity under the square root must be nonnegative:

$$9 - x^2 - y^2 \geq 0.$$

This inequality can be written in the form

$$x^2+y^2 \le 9.$$

Therefore, the domain of $g(x,y)$ is $\{(x,y) \in R^2 \mid x^2+y^2 \le 9\}$. The graph of this set of points can be described as a disk of radius 3 centered at the origin. The domain includes the boundary circle as shown in the following graph.

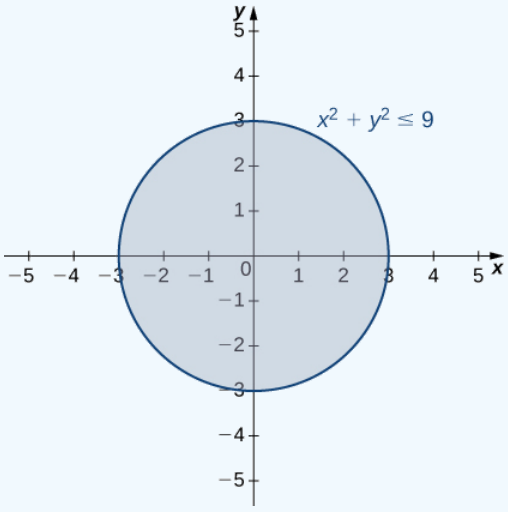


Figure 13.1.2: The domain of the function $g(x,y)=\sqrt{9-x^2-y^2}$ is a closed disk of radius 3.

To determine the range of $g(x,y)=\sqrt{9-x^2-y^2}$ we start with a point (x_0,y_0) on the boundary of the domain, which is defined by the relation $x^2+y^2=9$. It follows that $x_0^2+y_0^2=9$ and

$$\begin{aligned} g(x_0,y_0) &= \sqrt{9-x_0^2-y_0^2} \\ &= \sqrt{9-(x_0^2+y_0^2)} \\ &= \sqrt{9-9} \\ &= 0. \end{aligned}$$

If $x_0^2+y_0^2=0$ (in other words, $x_0=y_0=0$) , then

$$\begin{aligned} g(x_0,y_0) &= \sqrt{9-x_0^2-y_0^2} \\ &= \sqrt{9-(x_0^2+y_0^2)} \\ &= \sqrt{9-0}=3. \end{aligned}$$

This is the maximum value of the function. Given any value c between 0 and 3, we can find an entire set of points inside the domain of g such that $g(x,y)=c$:

$$\begin{aligned} \sqrt{9-x^2-y^2} &= c \\ 9-x^2-y^2 &= c^2 \\ x^2+y^2 &= 9-c^2. \end{aligned}$$

Since $9-c^2>0$, this describes a circle of radius $\sqrt{9-c^2}$ centered at the origin. Any point on this circle satisfies the equation $g(x,y)=c$. Therefore, the range of this function can be written in interval notation as $[0,3]$.

? Exercise 13.1.1

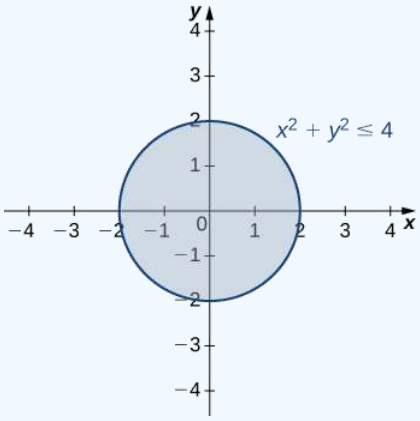
Find the domain and range of the function $f(x,y)=\sqrt{36-9x^2-9y^2}$.

Hint

Determine the set of ordered pairs that do not make the radicand negative.

Solution

The domain is $\{(x,y)|x^2+y^2\leq 4\}$ the shaded circle defined by the inequality $x^2+y^2\leq 4$, which has a circle of radius 2 as its boundary. The range is $[0,6]$.



Graphing Functions of Two Variables

Suppose we wish to graph the function $z=f(x,y)$. This function has two independent variables (x and y) and one dependent variable (z). When graphing a function $y=f(x)$ of one variable, we use the Cartesian plane. We are able to graph any ordered pair (x,y) in the plane, and every point in the plane has an ordered pair (x,y) associated with it. With a function of two variables, each ordered pair (x,y) in the domain of the function is mapped to a real number z. Therefore, the graph of the function f consists of ordered triples (x,y,z). The graph of a function $z=f(x,y)$ of two variables is called a surface.

To understand more completely the concept of plotting a set of ordered triples to obtain a surface in three-dimensional space, imagine the (x,y) coordinate system laying flat. Then, every point in the domain of the function f has a unique z-value associated with it. If z is positive, then the graphed point is located above the xy-plane, if z is negative, then the graphed point is located below the xy-plane. The set of all the graphed points becomes the two-dimensional surface that is the graph of the function f.

Example 13.1.2: Graphing Functions of Two Variables

Create a graph of each of the following functions:

a. $g(x,y)=\sqrt{9-x^2-y^2}$
b. $f(x,y)=x^2+y^2$

Solution

a. In Example $13.1.2$ we determined that the domain of $g(x,y)=\sqrt{9-x^2-y^2}$ is $\{(x,y)\in R^2\mid x^2+y^2\leq 9\}$ and the range is $\{z\in R^2\mid 0\leq z\leq 3\}$. When $x^2+y^2=9$ we have $g(x,y)=0$. Therefore any point on the circle of radius 3 centered at the origin in the xy-plane maps to $z=0$ in R^3. If $x^2+y^2=8$, then $g(x,y)=1$, so any point on the circle of radius $2\sqrt{2}$ centered at the origin in the xy-plane maps to $z=1$ in R^3. As x^2+y^2 gets closer to zero, the value of z approaches 3. When $x^2+y^2=0$, then $g(x,y)=3$. This is the origin in the xy-plane If x^2+y^2 is equal to any other value between 0 and 9, then $g(x,y)$ equals some other constant between 0 and 3. The surface described by this function is a hemisphere centered at the origin with radius 3 as shown in the following graph.

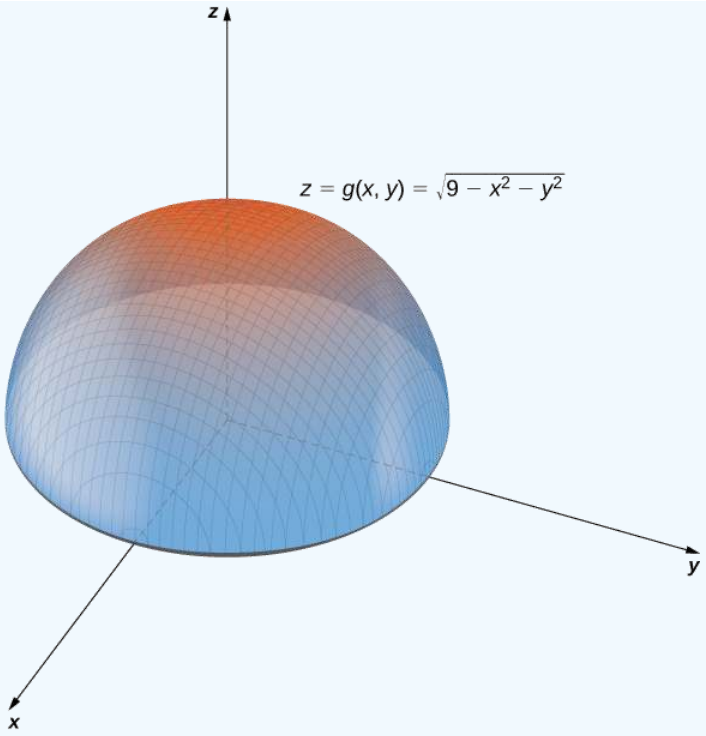


Figure 13.1.3: Graph of the hemisphere represented by the given function of two variables.

b. This function also contains the expression x^2+y^2 . Setting this expression equal to various values starting at zero, we obtain circles of increasing radius. The minimum value of $f(x,y)=x^2+y^2$ is zero (attained when $x=y=0$.. When $x=0$, the function becomes $z=y^2$, and when $y=0$, then the function becomes $z=x^2$. These are cross-sections of the graph, and are parabolas. Recall from Introduction to Vectors in Space that the name of the graph of $f(x,y)=x^2+y^2$ is a **paraboloid**. The graph of f appears in the following graph.

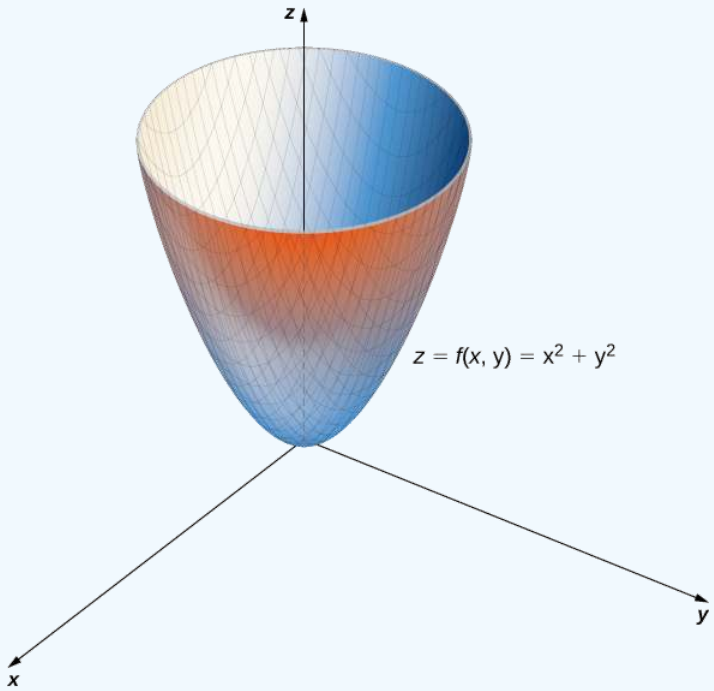


Figure 13.1.4: A paraboloid is the graph of the given function of two variables.

Example 13.1.3: Nuts and Bolts

A profit function for a hardware manufacturer is given by

$$f(x,y) = 16 - (x-3)^2 - (y-2)^2,$$

where x is the number of nuts sold per month (measured in thousands) and y represents the number of bolts sold per month (measured in thousands). Profit is measured in thousands of dollars. Sketch a graph of this function.

Solution

This function is a polynomial function in two variables. The domain of f consists of (x,y) coordinate pairs that yield a nonnegative profit:

$$16 - (x-3)^2 - (y-2)^2 \geq 0$$
$$(x-3)^2 + (y-2)^2 \leq 16.$$

This is a disk of radius 4 centered at $(3,2)$. A further restriction is that both x and y must be nonnegative. When $x = 3$ and $y = 2$, $f(x,y) = 16$. Note that it is possible for either value to be a noninteger; for example, it is possible to sell 2.5 thousand nuts in a month. The domain, therefore, contains thousands of points, so we can consider all points within the disk. For any $z < 16$, we can solve the equation $f(x,y) = 16$:

$$16 - (x-3)^2 - (y-2)^2 = z$$
$$(x-3)^2 + (y-2)^2 = 16 - z.$$

Since $z < 16,$ we know that $16 - z > 0,$ so the previous equation describes a circle with radius $\sqrt{16-z}$ centered at the point $(3,2)$. Therefore. the range of $f(x,y)$ is $\{z \in \mathbb{R} | z \leq 16\}$. The graph of $f(x,y)$ is also a paraboloid, and this paraboloid points downward as shown.

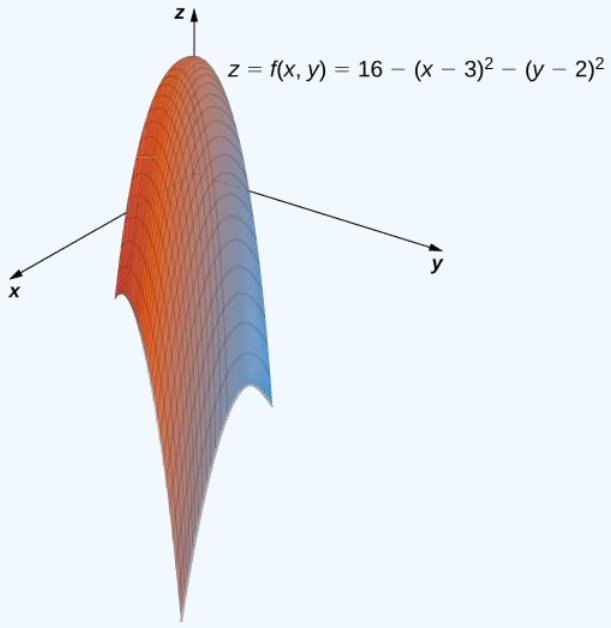
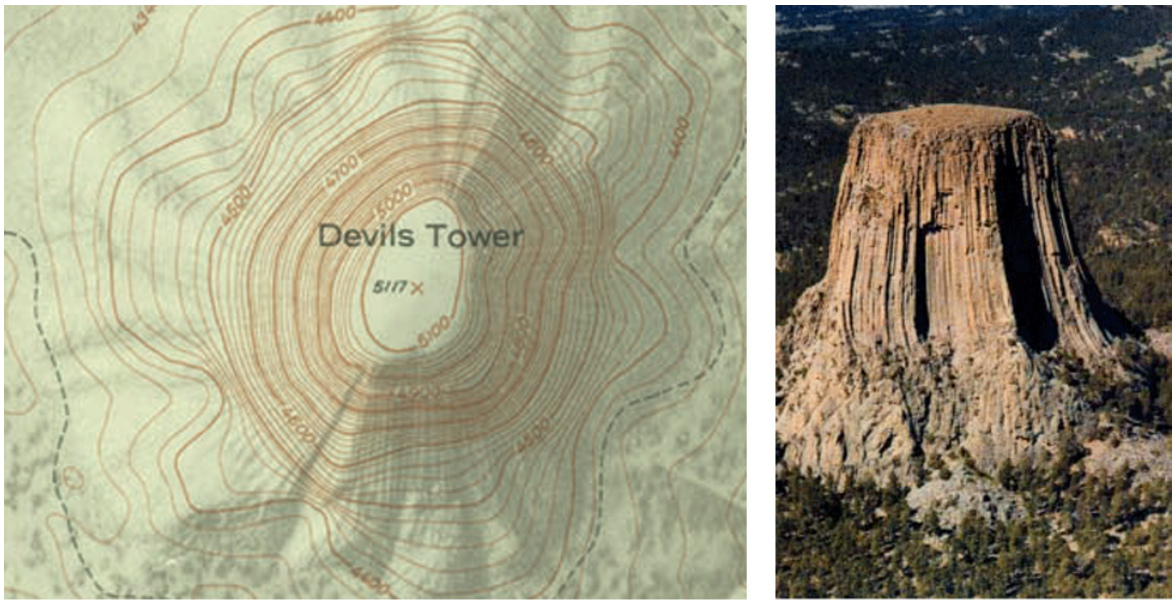


Figure 13.1.5: The graph of the given function of two variables is also a paraboloid.

Level Curves

If hikers walk along rugged trails, they might use a topographical map that shows how steeply the trails change. A *topographical map* contains curved lines called contour lines. Each contour line corresponds to the points on the map that have equal elevation (Figure 13.1.6). A level curve of a function of two variables $f(x,y)$ is completely analogous to a contour line on a topographical map.



(a) (b)

Figure 13.1.6: (a) A topographical map of Devil's Tower, Wyoming. Lines that are close together indicate very steep terrain. (b) A perspective photo of Devil's Tower shows just how steep its sides are. Notice the top of the tower has the same shape as the center of the topographical map.

Definition: Level Curves

Given a function $f(x,y)$ and a number c in the range of f, a level curve of a function of two variables for the value c is defined to be the set of points satisfying the equation $f(x,y)=c$.

Returning to the function $g(x,y)=\sqrt{9-x^2-y^2}$, we can determine the level curves of this function. The range of g is the closed interval $[0,3]$. First, we choose any number in this closed interval—say, $c=2$. The level curve corresponding to $c=2$ is described by the equation

$$\sqrt{9-x^2-y^2}=2.$$

To simplify, square both sides of this equation:

$$9-x^2-y^2=4.$$

Now, multiply both sides of the equation by -1 and add 9 to each side:

$$x^2+y^2=5.$$

This equation describes a circle centered at the origin with radius $\sqrt{5}$. Using values of c between 0 and 3 yields other circles also centered at the origin. If $c=3$, then the circle has radius 0, so it consists solely of the origin. Figure $13.1.7$ is a graph of the level curves of this function corresponding to $c=0,1,2$, and 3. Note that in the previous derivation it may be possible that we introduced extra solutions by squaring both sides. This is not the case here because the range of the square root function is nonnegative.

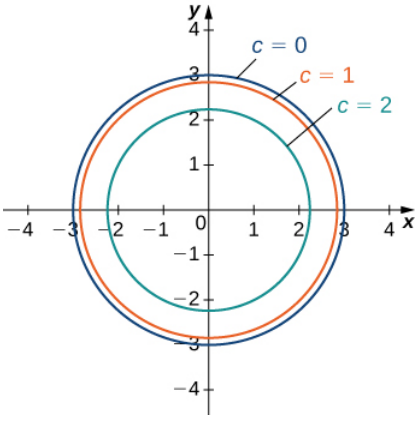


Figure 13.1.7: Level curves of the function $g(x,y)=\sqrt{9-x^2-y^2}$, using $c=0,1,2,$ and $3(c=3$ corresponds to the origin).

A graph of the various level curves of a function is called a *contour map*.

Example 13.1.4: Making a Contour Map

Given the function $f(x,y)=\sqrt{8+8x-4y-4x^2-y^2}$, find the level curve corresponding to $c=0$. Then create a contour map for this function. What are the domain and range of f?

Solution

To find the level curve for $c=0$, we set $f(x,y)=0$ and solve. This gives

$$0=\sqrt{8+8x-4y-4x^2-y^2}\ .$$

We then square both sides and multiply both sides of the equation by -1:

$$4x^2+y^2-8x+4y-8=0.$$

Now, we rearrange the terms, putting the x terms together and the y terms together, and add 8 to each side:

$$4x^2-8x+y^2+4y=8.$$

Next, we group the pairs of terms containing the same variable in parentheses, and factor 4 from the first pair:

$$4(x^2-2x)+(y^2+4y)=8.$$

Then we complete the square in each pair of parentheses and add the correct value to the right-hand side:

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

Next, we factor the left-hand side and simplify the right-hand side:

$$4(x-1)^2+(y+2)^2=16.$$

Last, we divide both sides by 16 :

$$\frac{(x-1)^2}{4}+\frac{(y+2)^2}{16}=1. \tag{13.1.1}$$

This equation describes an ellipse centered at $(1,-2)$. The graph of this ellipse appears in the following graph.

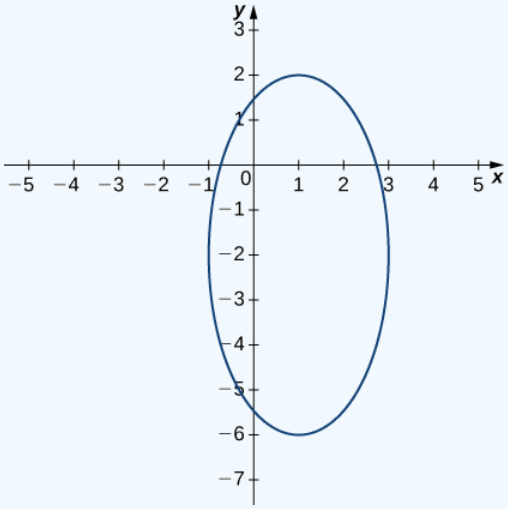


Figure 13.1.8: Level curve of the function $f(x,y)=\sqrt{8+8x-4y-4x^2-y^2}$ corresponding to $c=0$

We can repeat the same derivation for values of c less than 4. Then, Equation 13.1.1 becomes

$$\frac{4(x-1)^2}{16-c^2}+\frac{(y+2)^2}{16-c^2}=1$$

for an arbitrary value of c. Figure 13.1.9 shows a contour map for $f(x,y)$ using the values $c=0,1,2$, and 3. When $c=4$, the level curve is the point $(-1,2)$.

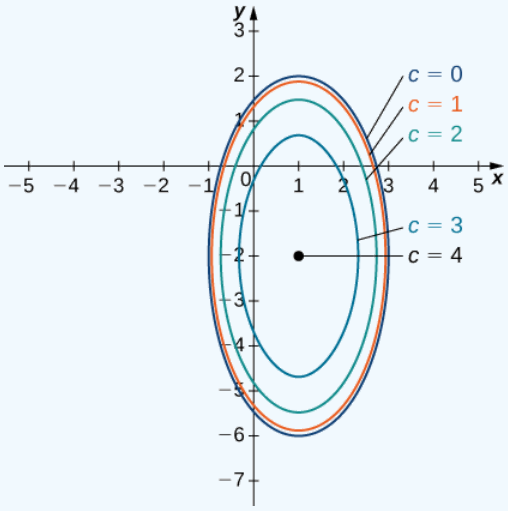


Figure 13.1.9: Contour map for the function $f(x,y)=\sqrt{8+8x-4y-4x^2-y^2}$ using the values $c=0,1,2,3,$ and 4.

Finding the Domain & Range

Since this is a square root function, the radicand must not be negative. So we have

$$8+8x-4y-4x^2-y^2\geq 0$$

Recognizing that the boundary of the domain is an ellipse, we repeat the steps we showed above to obtain

$$\frac{(x-1)^2}{4}+\frac{(y+2)^2}{16}\leq 1$$

So the domain of f can be written: $\left\{(x,y)\,|\,\frac{(x-1)^2}{4}+\frac{(y+2)^2}{16}\leq 1\right\}$.

To find the range of f, we need to consider the possible outputs of this square root function. We know the output cannot be negative, so we need to next check if its output is ever 0. From the work we completed above to find the level curve for $c=0$,

we know the value of f is 0 for any point on that level curve (on the ellipse, $\frac{(x-1)^2}{4}+\frac{(y+2)^2}{16}=1$). So we know the lower bound of the range of this function is 0.

To determine the upper bound for the range of the function in this problem, it's easier if we first complete the square under the radical.

$$\begin{aligned} f(x,y) &= \sqrt{8+8x-4y-4x^2-y^2} \\ &= \sqrt{8-4(x^2-2x\quad)-(y^2+4y\quad)} \\ &= \sqrt{8-4(x^2-2x+1-1)-(y^2+4y+4-4)} \\ &= \sqrt{8-4(x^2-2x+1)+4-(y^2+4y+4)+4} \\ &= \sqrt{16-4(x-1)^2-(y+2)^2} \end{aligned}$$

Now that we have f in this form, we can see how large the radicand can be. Since we are subtracting two perfect squares from 16, we know that the value of the radicand cannot be greater than 16. At the point $(1,-2)$, we can see the radicand will be 16 (since we will be subtracting 0 from 16 at this point. This gives us the maximum value of f, that is $f(1,-2)=\sqrt{16}=4$.

So the range of this function is $[0,4]$.

Exercise 13.1.2

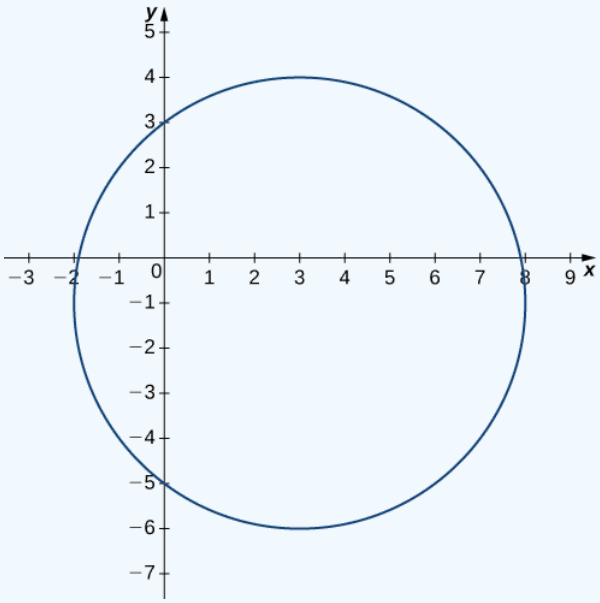
Find and graph the level curve of the function $g(x,y)=x^2+y^2-6x+2y$ corresponding to $c=15$.

Hint

First, set $g(x,y)=15$ and then complete the square.

Solution

The equation of the level curve can be written as $(x-3)^2+(y+1)^2=25,$ which is a circle with radius 5 centered at $(3,-1)$.



Another useful tool for understanding the **graph of a function of two variables** is called a vertical trace. Level curves are always graphed in the xy-plane, but as their name implies, vertical traces are graphed in the xz- or yz-planes.

Definition: Vertical Traces

Consider a function $z = f(x, y)$ with domain $D \subseteq \mathbb{R}^2$. A *vertical trace* of the function can be either the set of points that solves the equation $f(a, y) = z$ for a given constant $x = a$ or $f(x, b) = z$ for a given constant $y = b$.

Example 13.1.5: Finding Vertical Traces

Find vertical traces for the function $f(x, y) = \sin x \cos y$ corresponding to $x = -\frac{\pi}{4}, 0$, and $\frac{\pi}{4}$, and $y = -\frac{\pi}{4}, 0$, and $\frac{\pi}{4}$.

Solution

First set $x = -\frac{\pi}{4}$ in the equation $z = \sin x \cos y$:

$$z = \sin(-\frac{\pi}{4}) \cos y = -\frac{\sqrt{2} \cos y}{2} \approx -0.7071 \cos y.$$

This describes a cosine graph in the plane $x = -\frac{\pi}{4}$. The other values of z appear in the following table.

Vertical Traces Parallel to the $xz - Plane$ for the Function $f(x, y) = \sin x \cos y$

c	Vertical Trace for $x = c$
$-\frac{\pi}{4}$	$z = -\frac{\sqrt{2} \cos y}{2}$
0	$z = 0$
$\frac{\pi}{4}$	$z = \frac{\sqrt{2} \cos y}{2}$

In a similar fashion, we can substitute the $y - values$ in the equation $f(x, y)$ to obtain the traces in the $yz - plane$, as listed in the following table.

Vertical Traces Parallel to the $yz - Plane$ for the Function $f(x, y) = \sin x \cos y$

d	Vertical Trace for $y = d$
$\frac{\pi}{4}$	$z = \frac{\sqrt{2} \sin x}{2}$
0	$z = \sin x$
$-\frac{\pi}{4}$	$z = \frac{\sqrt{2} \sin x}{2}$

The three traces in the $xz - plane$ are cosine functions; the three traces in the $yz - plane$ are sine functions. These curves appear in the intersections of the surface with the planes $x = -\frac{\pi}{4}, x = 0, x = \frac{\pi}{4}$ and $y = -\frac{\pi}{4}, y = 0, y = \frac{\pi}{4}$ as shown in the following figure.

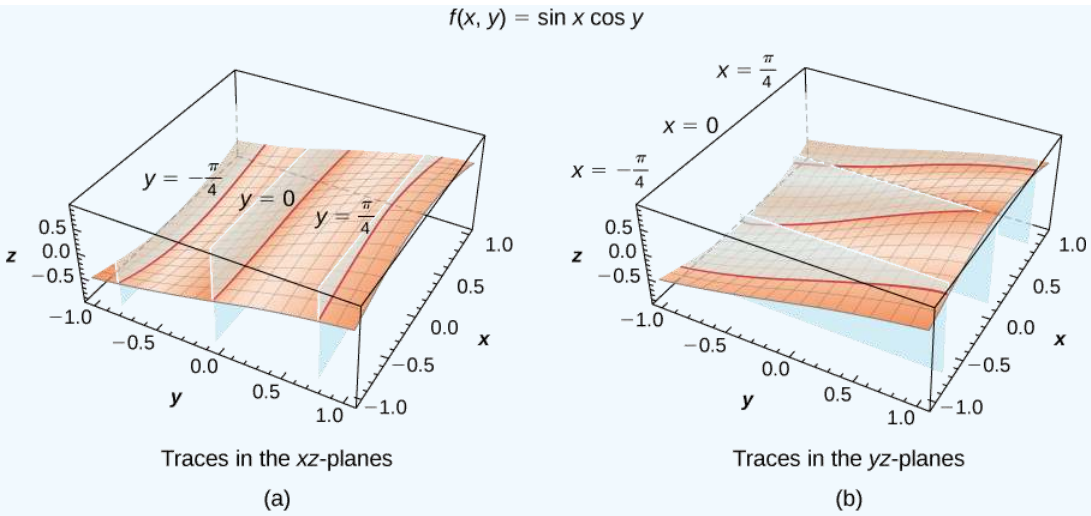


Figure 13.1.10: Vertical traces of the function $f(x,y)$ are cosine curves in the $xz-planes$ (a) and sine curves in the $yz-planes$ (b).

? Exercise 13.1.3

Determine the equation of the vertical trace of the function $g(x,y)=-x^2-y^2+2x+4y-1$ corresponding to $y=3$, and describe its graph.

Hint

Set $y=3$ in the equation $z=-x^2-y^2+2x+4y-1$ and complete the square.

Solution

$z=3-(x-1)^2$. This function describes a parabola opening downward in the plane $y=3$.

Functions of two variables can produce some striking-looking surfaces. Figure 13.1.11shows two examples.

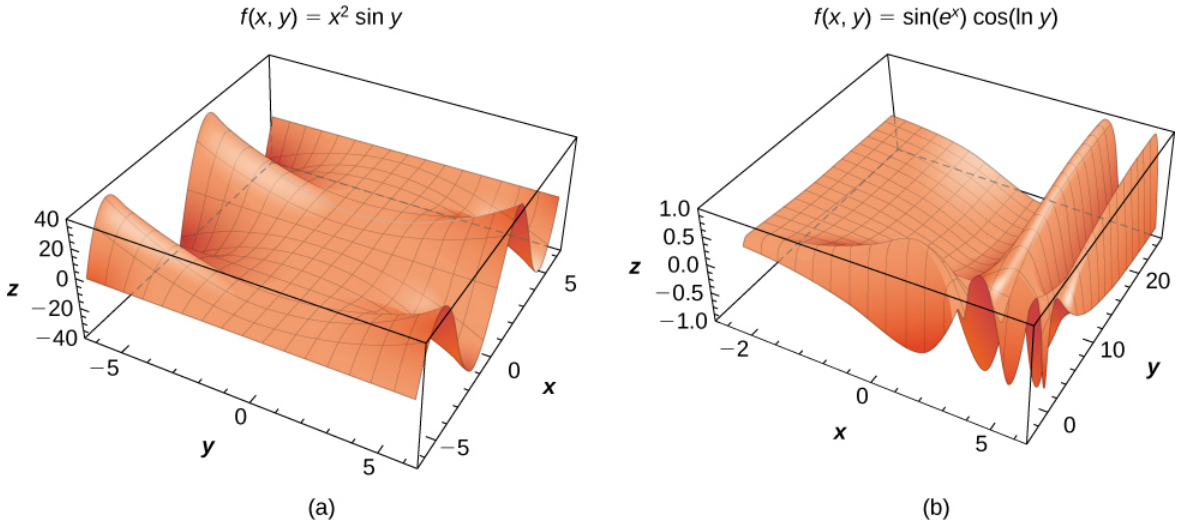


Figure 13.1.11: Examples of surfaces representing functions of two variables: (a) a combination of a power function and a sine function and (b) a combination of trigonometric, exponential, and logarithmic functions.

Functions of More Than Two Variables

So far, we have examined only functions of two variables. However, it is useful to take a brief look at functions of more than two variables. Two such examples are

$$\underbrace{f(x,y,z)=x^2-2xy+y^2+3yz-z^2+4x-2y+3x-6}_{\text{a polynomial in three variables}}$$

and

$$g(x,y,t)=(x^2-4xy+y^2)\sin t-(3x+5y)\cos t.$$

In the first function, (x,y,z) represents a point in space, and the function f maps each point in space to a fourth quantity, such as temperature or wind speed. In the second function, (x,y) can represent a point in the plane, and t can represent time. The function might map a point in the plane to a third quantity (for example, pressure) at a given time t. The method for finding the domain of a function of more than two variables is analogous to the method for functions of one or two variables.

Example 13.1.6: Domains for Functions of Three Variables

Find the domain of each of the following functions:

a. $f(x,y,z)=\dfrac{3x-4y+2z}{\sqrt{9-x^2-y^2-z^2}}$

b. $g(x,y,t)=\dfrac{\sqrt{2t-4}}{x^2-y^2}$

Solution:

a. For the function $f(x,y,z)=\dfrac{3x-4y+2z}{\sqrt{9-x^2-y^2-z^2}}$ to be defined (and be a real value), two conditions must hold:

1. The denominator **cannot** be zero.
2. The radicand **cannot** be negative.

Combining these conditions leads to the inequality

$$9-x^2-y^2-z^2>0.$$

Moving the variables to the other side and reversing the inequality gives the domain as

$$domain(f)=\{(x,y,z)\in R^3 \mid x^2+y^2+z^2<9\},$$

which describes a ball of radius 3 centered at the origin. (Note: The surface of the ball is not included in this domain.)

b. For the function $g(x,y,t)=\dfrac{\sqrt{2t-4}}{x^2-y^2}$ to be defined (and be a real value), two conditions must hold:

1. The radicand cannot be negative.
2. The denominator cannot be zero.

Since the radicand cannot be negative, this implies $2t-4\geq 0$, and therefore that $t\geq 2$. Since the denominator cannot be zero, $x^2-y^2\neq 0$, or $x^2\neq y^2$, Which can be rewritten as $y=\pm x$, which are the equations of two lines passing through the origin. Therefore, the domain of g is

$$domain(g)=\{(x,y,t)|y\neq\pm x, t\geq 2\}.$$

Exercise 13.1.4

Find the domain of the function $h(x,y,t)=(3t-6)\sqrt{y-4x^2+4}$.

Hint

Check for values that make radicands negative or denominators equal to zero.

Solution

$$domain(h)=\{(x,y,t)\in\mathbb{R}^3 \mid y\geq 4x^2-4\}$$

Functions of two variables have level curves, which are shown as curves in the $xy-plane$. However, when the function has three variables, the curves become surfaces, so we can define level surfaces for functions of three variables.

Definition: Level Surface of a Function of Three Variables

Given a function $f(x,y,z)$ and a number c in the range of f, a *level surface of a function of three variables* is defined to be the set of points satisfying the equation $f(x,y,z)=c$.

Example 13.1.7: Finding a Level Surface

Find the level surface for the function $f(x,y,z)=4x^2+9y^2-z^2$ corresponding to $c=1$.

Solution

The level surface is defined by the equation $4x^2+9y^2-z^2=1$. This equation describes a hyperboloid of one sheet as shown in Figure 13.1.12

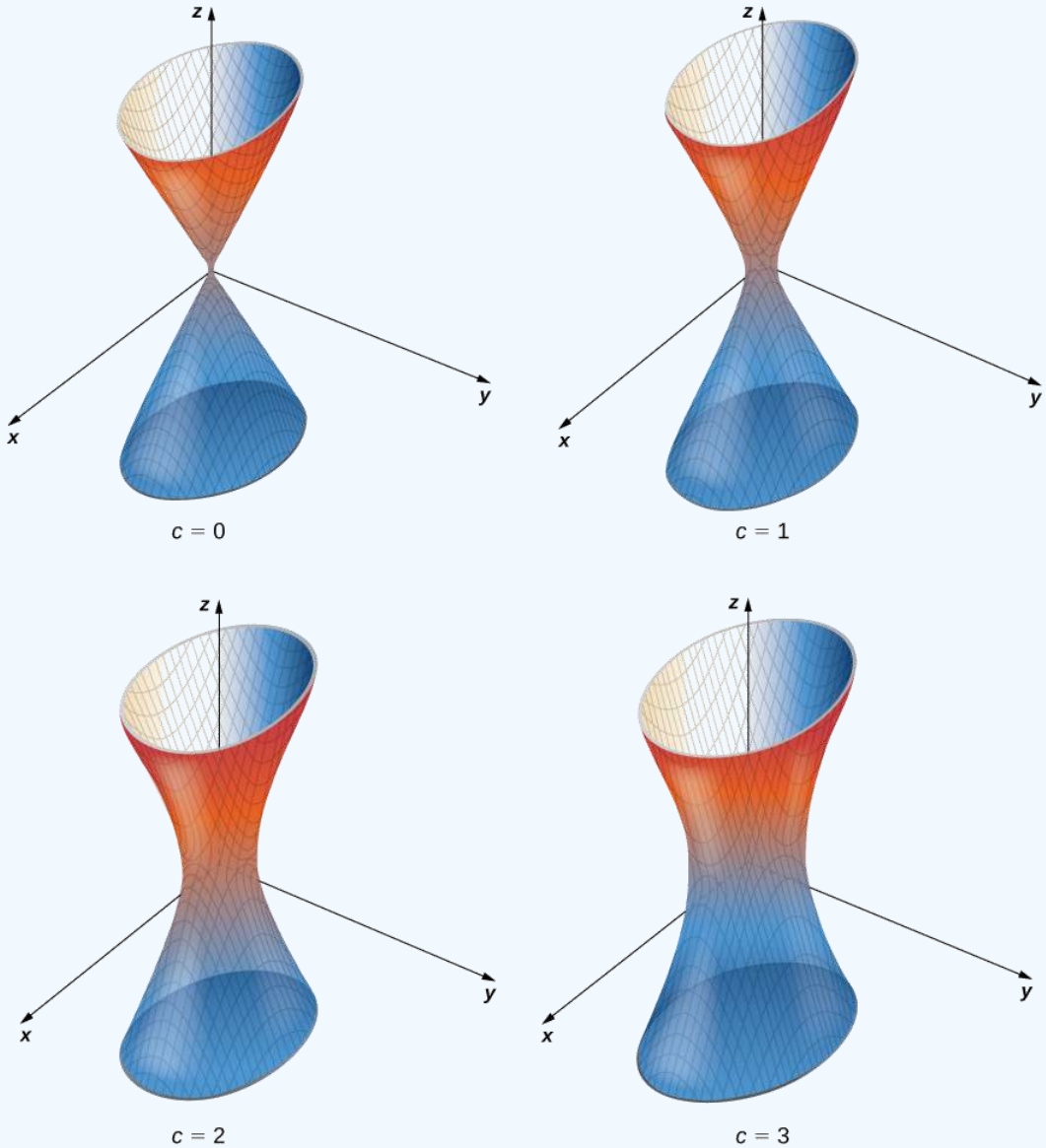


Figure 13.1.12: A hyperboloid of one sheet with some of its level surfaces.

? Exercise 13.1.5

Find the equation of the level surface of the function

$$g(x,y,z)=x^2+y^2+z^2-2x+4y-6z$$

corresponding to $c=2$, and describe the surface, if possible.

Hint

Set $g(x,y,z)=c$ and complete the square.

Solution

$(x-1)^2+(y+2)^2+(z-3)^2=16$ describes a sphere of radius 4 centered at the point $(1,-2,3)$.

Summary

- The graph of a function of two variables is a surface in $\mathbb{R}^3$ and can be studied using level curves and vertical traces.
- A set of level curves is called a contour map.

Key Equations

- **Vertical trace**

  $f(a,y)=z$ for $x=a$ or $f(x,b)=z$ for $y=b$

- **Level surface of a function of three variables**

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

Glossary

contour map
a plot of the various level curves of a given function $f(x,y)$

function of two variables
a function $z=f(x,y)$ that maps each ordered pair (x,y) in a subset D of R^2 to a unique real number z

graph of a function of two variables
a set of ordered triples (x,y,z) that satisfies the equation $z=f(x,y)$ plotted in three-dimensional Cartesian space

level curve of a function of two variables
the set of points satisfying the equation $f(x,y)=c$ for some real number c in the range of f

level surface of a function of three variables
the set of points satisfying the equation $f(x,y,z)=c$ for some real number c in the range of f

surface
the graph of a function of two variables, $z=f(x,y)$

vertical trace
the set of ordered triples (c,y,z) that solves the equation $f(c,y)=z$ for a given constant $x=c$ or the set of ordered triples (x,d,z) that solves the equation $f(x,d)=z$ for a given constant $y=d$

13.1: Functions of Several Variables is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **14.1: Functions of Several Variables** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

13.2: Limits and Continuity in Higher Dimensions

Learning Objectives

- Calculate the limit of a function of two variables.
- Learn how a function of two variables can approach different values at a boundary point, depending on the path of approach.
- State the conditions for continuity of a function of two variables.
- Verify the continuity of a function of two variables at a point.
- Calculate the limit of a function of three or more variables and verify the continuity of the function at a point.

We have now examined functions of more than one variable and seen how to graph them. In this section, we see how to take the limit of a function of more than one variable, and what it means for a function of more than one variable to be continuous at a point in its domain. It turns out these concepts have aspects that just don't occur with functions of one variable.

Limit of a Function of Two Variables

Recall from Section 2.5 that the definition of a limit of a function of one variable:

Let $f(x)$ be defined for all $x \neq a$ in an open interval containing a. Let L be a real number. Then

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

if for every $\varepsilon > 0$, there exists a $\delta > 0$, such that if $0 < |x - a| < \delta$ for all x in the domain of f, then

$$|f(x) - L| < \varepsilon.$$

Before we can adapt this definition to define a limit of a function of two variables, we first need to see how to extend the idea of an open interval in one variable to an open interval in two variables.

Definition: δ Disks

Consider a point $(a, b) \in \mathbb{R}^2$. A δ **disk** centered at point (a, b) is defined to be an open disk of radius δ centered at point (a, b) —that is,

$$\{(x, y) \in \mathbb{R}^2 \mid (x - a)^2 + (y - b)^2 < \delta^2\}$$

as shown in Figure 13.2.1.

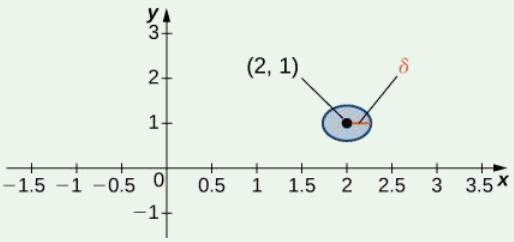


Figure 13.2.1: A δ disk centered around the point $(2, 1)$.

The idea of a δ disk appears in the definition of the limit of a function of two variables. If δ is small, then all the points (x, y) in the δ disk are close to (a, b). This is completely analogous to x being close to a in the definition of a limit of a function of one variable. In one dimension, we express this restriction as

$$a - \delta < x < a + \delta.$$

In more than one dimension, we use a δ disk.

Definition: limit of a function of two variables

Let f be a function of two variables, x and y. The limit of $f(x,y)$ as (x,y) approaches (a,b) is L, written

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

if for each $\varepsilon > 0$ there exists a small enough $\delta > 0$ such that for all points (x,y) in a δ disk around (a,b), except possibly for (a,b) itself, the value of $f(x,y)$ is no more than ε away from L (Figure 13.2.2).

Using symbols, we write the following: For any $\varepsilon > 0$, there exists a number $\delta > 0$ such that

$$|f(x,y) - L| < \varepsilon$$

whenever

$$0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta.$$

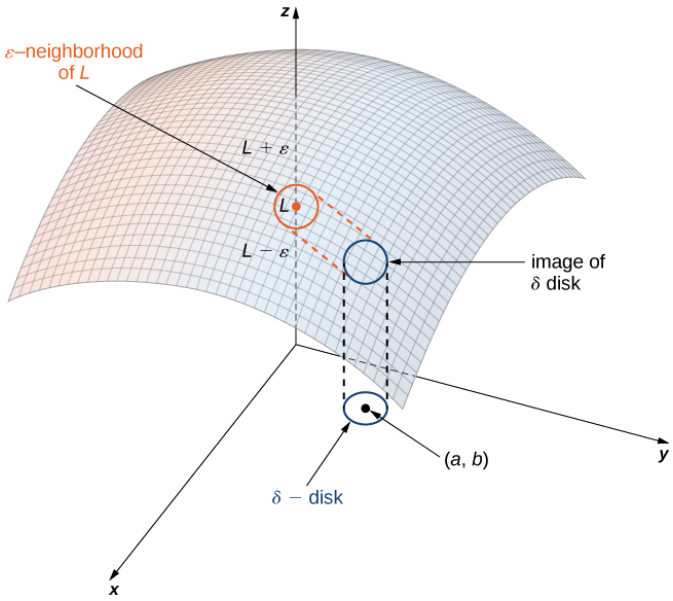


Figure 13.2.2: The limit of a function involving two variables requires that $f(x,y)$ be within ε of L whenever (x,y) is within δ of (a,b). The smaller the value of ε, the smaller the value of δ.

Proving that a limit exists using the definition of a **limit of a function of two variables** can be challenging. Instead, we use the following theorem, which gives us shortcuts to finding limits. The formulas in this theorem are an extension of the formulas in the limit laws theorem in the section titled The Limit Laws in single-variable calculus.

Theorem 13.2.1: Limit Laws for Functions of Two Variables

Let $f(x,y)$ and $g(x,y)$ be defined for all $(x,y) \neq (a,b)$ in a neighborhood around (a,b), and assume the neighborhood is contained completely inside the domain of f. Assume that L and M are real numbers such that

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

and

$$\lim_{(x,y)\to(a,b)} g(x,y) = M,$$

and let c be a constant. Then each of the following statements holds:

Constant Law:

$$\lim_{(x,y)\to(a,b)} c = c$$

Identity Laws:

$$\lim_{(x,y)\to(a,b)} x = a$$

$$\lim_{(x,y)\to(a,b)} y = b$$

Sum Law:

$$\lim_{(x,y)\to(a,b)} (f(x,y)+g(x,y)) = L+M$$

Difference Law:

$$\lim_{(x,y)\to(a,b)} (f(x,y)-g(x,y)) = L-M$$

Constant Multiple Law:

$$\lim_{(x,y)\to(a,b)} (cf(x,y)) = cL$$

Product Law:

$$\lim_{(x,y)\to(a,b)} (f(x,y)g(x,y)) = LM$$

Quotient Law:

$$\lim_{(x,y)\to(a,b)} \frac{f(x,y)}{g(x,y)} = \frac{L}{M} \text{ for } M \neq 0$$

Power Law:

$$\lim_{(x,y)\to(a,b)} (f(x,y))^n = L^n$$

for any positive integer n.

Root Law:

$$\lim_{(x,y)\to(a,b)} \sqrt[n]{f(x,y)} = \sqrt[n]{L}$$

for all L if n is odd and positive, and for $L \geq 0$ if n is even and positive.

The proofs of these properties are similar to those for the limits of functions of one variable. We can apply these laws to finding limits of various functions.

Example 13.2.1: Finding the Limit of a Function of Two Variables

Find each of the following limits:

a. $\lim_{(x,y)\to(2,-1)} (x^2-2xy+3y^2-4x+3y-6)$

b. $\lim_{(x,y)\to(2,-1)} \frac{2x+3y}{4x-3y}$

Solution

a. First use the sum and difference laws to separate the terms:

$$\lim_{(x,y)\to(2,-1)}(x^2-2xy+3y^2-4x+3y-6)$$
$$=\left(\lim_{(x,y)\to(2,-1)}x^2\right)-\left(\lim_{(x,y)\to(2,-1)}2xy\right)+\left(\lim_{(x,y)\to(2,-1)}3y^2\right)-\left(\lim_{(x,y)\to(2,-1)}4x\right)$$
$$+\left(\lim_{(x,y)\to(2,-1)}3y\right)-\left(\lim_{(x,y)\to(2,-1)}6\right).$$

Next, use the constant multiple law on the second, third, fourth, and fifth limits:

$$=(\lim_{(x,y)\to(2,-1)}x^2)-2(\lim_{(x,y)\to(2,-1)}xy)+3(\lim_{(x,y)\to(2,-1)}y^2)-4(\lim_{(x,y)\to(2,-1)}x)$$
$$+3(\lim_{(x,y)\to(2,-1)}y)-\lim_{(x,y)\to(2,-1)}6.$$

Now, use the power law on the first and third limits, and the product law on the second limit:

$$\left(\lim_{(x,y)\to(2,-1)}x\right)^2-2\left(\lim_{(x,y)\to(2,-1)}x\right)\left(\lim_{(x,y)\to(2,-1)}y\right)+3\left(\lim_{(x,y)\to(2,-1)}y\right)^2$$
$$-4\left(\lim_{(x,y)\to(2,-1)}x\right)+3\left(\lim_{(x,y)\to(2,-1)}y\right)-\lim_{(x,y)\to(2,-1)}6.$$

Last, use the identity laws on the first six limits and the constant law on the last limit:

$$\lim_{(x,y)\to(2,-1)}(x^2-2xy+3y^2-4x+3y-6)=(2)^2-2(2)(-1)+3(-1)^2-4(2)+3(-1)-6$$
$$=-6.$$

b. Before applying the quotient law, we need to verify that the limit of the denominator is nonzero. Using the difference law, constant multiple law, and identity law,

$$\lim_{(x,y)\to(2,-1)}(4x-3y)=\lim_{(x,y)\to(2,-1)}4x-\lim_{(x,y)\to(2,-1)}3y$$
$$=4(\lim_{(x,y)\to(2,-1)}x)-3(\lim_{(x,y)\to(2,-1)}y)$$
$$=4(2)-3(-1)=11.$$

Since the limit of the denominator is nonzero, the quotient law applies. We now calculate the limit of the numerator using the difference law, constant multiple law, and identity law:

$$\lim_{(x,y)\to(2,-1)}(2x+3y)=\lim_{(x,y)\to(2,-1)}2x+\lim_{(x,y)\to(2,-1)}3y$$
$$=2(\lim_{(x,y)\to(2,-1)}x)+3(\lim_{(x,y)\to(2,-1)}y)$$
$$=2(2)+3(-1)=1.$$

Therefore, according to the quotient law we have

$$\lim_{(x,y)\to(2,-1)}\frac{2x+3y}{4x-3y}=\frac{\lim_{(x,y)\to(2,-1)}(2x+3y)}{\lim_{(x,y)\to(2,-1)}(4x-3y)}$$
$$=\frac{1}{11}.$$

? Exercise 13.2.1:

Evaluate the following limit:

$$\lim_{(x,y)\to(5,-2)}\sqrt[3]{\frac{x^2-y}{y^2+x-1}}.$$

Hint

Use the limit laws.

Answer

$$\lim_{(x,y)\to(5,-2)} \sqrt[3]{\frac{x^2-y}{y^2+x-1}} = \frac{3}{2}$$

Since we are taking the limit of a function of two variables, the point (a,b) is in $\mathbb{R}^2$, and it is possible to approach this point from an infinite number of directions. Sometimes when calculating a limit, the answer varies depending on the path taken toward (a,b). If this is the case, then the limit fails to exist. In other words, the limit must be unique, regardless of path taken.

Example 13.2.2: Limits That Fail to Exist

Show that neither of the following limits exist:

a. $\lim\limits_{(x,y)\to(0,0)} \dfrac{2xy}{3x^2+y^2}$

b. $\lim\limits_{(x,y)\to(0,0)} \dfrac{4xy^2}{x^2+3y^4}$

Solution

a. The domain of the function $f(x,y) = \dfrac{2xy}{3x^2+y^2}$ consists of all points in the xy-plane except for the point $(0,0)$ (Figure 13.2.3). To show that the limit does not exist as (x,y) approaches $(0,0)$, we note that it is impossible to satisfy the definition of a limit of a function of two variables because of the fact that the function takes different values along different lines passing through point $(0,0)$. First, consider the line $y=0$ in the xy-plane. Substituting $y=0$ into $f(x,y)$ gives

$$f(x,0) = \frac{2x(0)}{3x^2+0^2} = 0$$

for any value of x. Therefore the value of f remains constant for any point on the x-axis, and as y approaches zero, the function remains fixed at zero.

Next, consider the line $y=x$. Substituting $y=x$ into $f(x,y)$ gives

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

This is true for any point on the line $y=x$. If we let x approach zero while staying on this line, the value of the function remains fixed at $\frac{1}{2}$, regardless of how small x is.

Choose a value for ε that is less than $1/2$—say, $1/4$. Then, no matter how small a δ disk we draw around $(0,0)$, the values of $f(x,y)$ for points inside that δ disk will include both 0 and $\frac{1}{2}$. Therefore, the definition of limit at a point is never satisfied and the limit fails to exist.

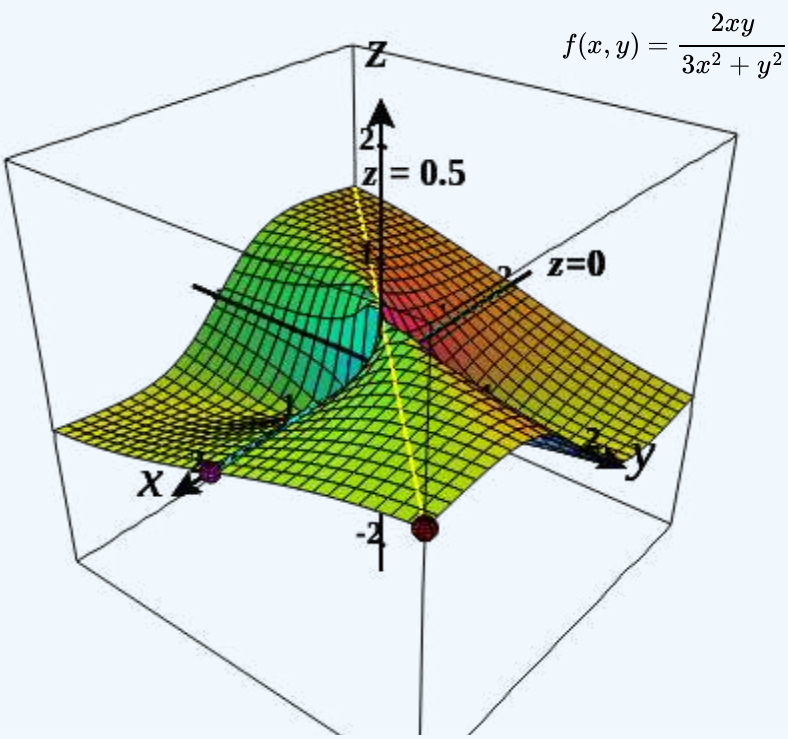


Figure 13.2.3: Graph of the function $f(x,y)=\dfrac{2xy}{3x^2+y^2}$. Along the line $y=0$, the function is equal to zero; along the line $y=x$, the function is equal to $\frac{1}{2}$.

b. In a similar fashion to a., we can approach the origin along any straight line passing through the origin. If we try the x-axis (i.e., $y=0$), then the function remains fixed at zero. The same is true for the y-axis. Suppose we approach the origin along a straight line of slope k. The equation of this line is $y=kx$. Then the limit becomes

$$\begin{aligned}\lim_{(x,y)\to(0,0)}\frac{4xy^2}{x^2+3y^4} &= \lim_{(x,y)\to(0,0)}\frac{4x(kx)^2}{x^2+3(kx)^4}\\ &= \lim_{(x,y)\to(0,0)}\frac{4k^2x^3}{x^2+3k^4x^4}\\ &= \lim_{(x,y)\to(0,0)}\frac{4k^2x}{1+3k^4x^2}\\ &= \frac{\lim\limits_{(x,y)\to(0,0)}(4k^2x)}{\lim\limits_{(x,y)\to(0,0)}(1+3k^4x^2)}\\ &= 0.\end{aligned}$$

regardless of the value of k. It would seem that the limit is equal to zero. What if we chose a curve passing through the origin instead? For example, we can consider the parabola given by the equation $x=y^2$. Substituting y^2 in place of x in $f(x,y)$ gives

$$\lim_{(x,y)\to(0,0)} \frac{4xy^2}{x^2+3y^4} = \lim_{(x,y)\to(0,0)} \frac{4(y^2)y^2}{(y^2)^2+3y^4}$$
$$= \lim_{(x,y)\to(0,0)} \frac{4y^4}{y^4+3y^4}$$
$$= \lim_{(x,y)\to(0,0)} 1$$
$$= 1.$$

By the same logic in part a, it is impossible to find a δ disk around the origin that satisfies the definition of the limit for any value of $\varepsilon < 1$. Therefore,

$$\lim_{(x,y)\to(0,0)} \frac{4xy^2}{x^2+3y^4}$$

does **not** exist.

? Exercise 13.2.2:

Show that

$$\lim_{(x,y)\to(2,1)} \frac{(x-2)(y-1)}{(x-2)^2+(y-1)^2}$$

does not exist.

Hint

Pick a line with slope k passing through point $(2,1)$.

Answer

If $y = k(x-2)+1$, then $\lim_{(x,y)\to(2,1)} \frac{(x-2)(y-1)}{(x-2)^2+(y-1)^2} = \frac{k}{1+k^2}$. Since the answer depends on k, the limit fails to exist.

Interior Points and Boundary Points

To study continuity and differentiability of a function of two or more variables, we first need to learn some new terminology.

Definition: interior and boundary points

Let S be a subset of $\mathbb{R}^2$ (Figure 13.2.4).

1. A point P_0 is called an *interior point* of S if there is a δ disk centered around P_0 contained completely in S.
2. A point P_0 is called a *boundary point* of S if every δ disk centered around P_0 contains points both inside and outside S.

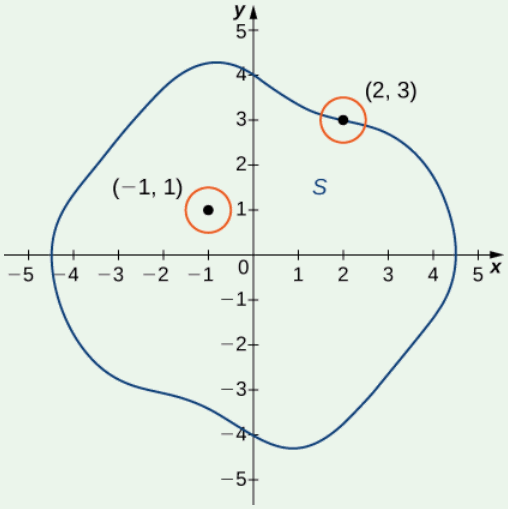


Figure 13.2.4: In the set S shown, $(-1,1)$ is an interior point and $(2,3)$ is a boundary point.

Definition: Open and closed sets

Let S be a subset of $\mathbb{R}^2$ (Figure 13.2.4).

1. S is called an *open set* if every point of S is an interior point.
2. S is called a *closed set* if it contains all its boundary points.

An example of an open set is a δ disk. If we include the boundary of the disk, then it becomes a closed set. A set that contains some, but not all, of its boundary points is neither open nor closed. For example if we include half the boundary of a δ disk but not the other half, then the set is neither open nor closed.

Definition: connected sets and Regions

Let S be a subset of $\mathbb{R}^2$ (Figure 13.2.4).

1. An open set S is a *connected set* if it cannot be represented as the union of two or more disjoint, nonempty open subsets.
2. A set S is a *region* if it is open, connected, and nonempty.

The definition of a limit of a function of two variables requires the δ disk to be contained inside the domain of the function. However, if we wish to find the limit of a function at a boundary point of the domain, the δ **disk** is not contained inside the domain. By definition, some of the points of the δ **disk** are inside the domain and some are outside. Therefore, we need only consider points that are inside both the δ disk and the domain of the function. This leads to the definition of the limit of a function at a boundary point.

Definition

Let f be a function of two variables, x and y, and suppose (a,b) is on the boundary of the domain of f. Then, the limit of $f(x,y)$ as (x,y) approaches (a,b) is L, written

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

if for any $\varepsilon > 0,$ there exists a number $\delta > 0$ such that for any point (x,y) inside the domain of f and within a suitably small distance positive δ of (a,b), the value of $f(x,y)$ is no more than ε away from L (Figure 13.2.2). Using symbols, we can write: For any $\varepsilon > 0$, there exists a number $\delta > 0$ such that

$$|f(x,y) - L| < \varepsilon \text{ whenever } 0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta.$$

Example 13.2.3: Limit of a Function at a Boundary Point

Prove

$$\lim_{(x,y)\to(4,3)} \sqrt{25-x^2-y^2} = 0.$$

Solution

The domain of the function $f(x,y)=\sqrt{25-x^2-y^2}$ is $\{(x,y)\in\mathbb{R}^2 \mid x^2+y^2\le 25\}$, which is a circle of radius 5 centered at the origin, along with its interior as shown in Figure 13.2.5.

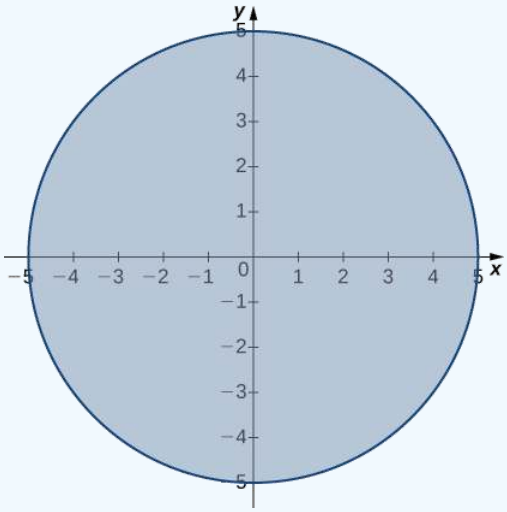


Figure 13.2.5: Domain of the function $f(x,y)=\sqrt{25-x^2-y^2}$.

We can use the limit laws, which apply to limits at the boundary of domains as well as interior points:

$$\begin{aligned}\lim_{(x,y)\to(4,3)} \sqrt{25-x^2-y^2} &= \sqrt{\lim_{(x,y)\to(4,3)}(25-x^2-y^2)}\\ &= \sqrt{\lim_{(x,y)\to(4,3)} 25 - \lim_{(x,y)\to(4,3)} x^2 - \lim_{(x,y)\to(4,3)} y^2}\\ &= \sqrt{25-4^2-3^2}\\ &= 0\end{aligned}$$

See the following graph.

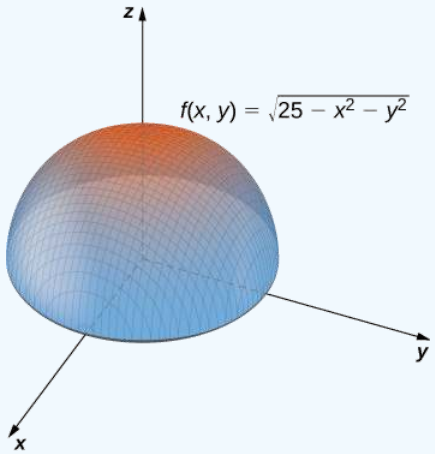


Figure 13.2.6: Graph of the function $f(x,y)=\sqrt{25-x^2-y^2}$.

Exercise 13.2.3

Evaluate the following limit:

$$\lim_{(x,y)\to(5,-2)} \sqrt{29-x^2-y^2}\,.$$

Hint

Determine the domain of $f(x,y)=\sqrt{29-x^2-y^2}$.

Answer

$$\lim_{(x,y)\to(5,-2)} \sqrt{29-x^2-y^2} = \sqrt{29-5^2-(-2)^2} = \sqrt{29-25-4} = \sqrt{0} = 0$$

Continuity of Functions of Two Variables

In Continuity, we defined the continuity of a function of one variable and saw how it relied on the limit of a function of one variable. In particular, three conditions are necessary for $f(x)$ to be continuous at point $x=a$

1. $f(a)$ exists.
2. $\lim\limits_{x\to a} f(x)$ exists.
3. $\lim\limits_{x\to a} f(x) = f(a)$.

These three conditions are necessary for continuity of a function of two variables as well.

Definition: continuous Functions

A function $f(x,y)$ is continuous at a point (a,b) in its domain if the following conditions are satisfied:

1. $f(a,b)$ exists.
2. $\lim\limits_{(x,y)\to(a,b)} f(x,y)$ exists.
3. $\lim\limits_{(x,y)\to(a,b)} f(x,y) = f(a,b)$.

Example 13.2.4: Demonstrating Continuity for a Function of Two Variables

Show that the function

$$f(x,y)=\frac{3x+2y}{x+y+1}$$

is continuous at point $(5,-3)$.

Solution

There are three conditions to be satisfied, per the definition of continuity. In this example, $a=5$ and $b=-3$.

1. $f(a,b)$ exists. This is true because the domain of the function f consists of those ordered pairs for which the denominator is nonzero (i.e., $x+y+1\neq 0$). Point $(5,-3)$ satisfies this condition. Furthermore,

$$f(a,b)=f(5,-3)=\frac{3(5)+2(-3)}{5+(-3)+1}=\frac{15-6}{2+1}=3.$$

2. $\lim\limits_{(x,y)\to(a,b)} f(x,y)$ exists. This is also true:

$$\begin{aligned}\lim_{(x,y)\to(a,b)} f(x,y) &= \lim_{(x,y)\to(5,-3)} \frac{3x+2y}{x+y+1} \\ &= \frac{\lim\limits_{(x,y)\to(5,-3)}(3x+2y)}{\lim\limits_{(x,y)\to(5,-3)}(x+y+1)} \\ &= \frac{15-6}{5-3+1} \\ &= 3.\end{aligned}$$

3. $\lim\limits_{(x,y)\to(a,b)} f(x,y) = f(a,b)$. This is true because we have just shown that both sides of this equation equal three.

? Exercise 13.2.4

Show that the function

$$f(x,y) = \sqrt{26-2x^2-y^2}$$

is continuous at point $(2,-3)$.

Hint

Use the three-part definition of continuity.

Answer

1. The domain of f contains the ordered pair $(2,-3)$ because $f(a,b) = f(2,-3) = \sqrt{16-2(2)^2-(-3)^2} = 3$
2. $\lim\limits_{(x,y)\to(a,b)} f(x,y) = 3$
3. $\lim\limits_{(x,y)\to(a,b)} f(x,y) = f(a,b) = 3$

Continuity of a function of any number of variables can also be defined in terms of delta and epsilon. A function of two variables is continuous at a point (x_0,y_0) in its domain if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that, whenever $\sqrt{(x-x_0)^2+(y-y_0)^2} < \delta$ it is true, $|f(x,y)-f(a,b)| < \varepsilon$. This definition can be combined with the formal definition (that is, the *epsilon–delta definition*) of continuity of a function of one variable to prove the following theorems:

Theorem 13.2.2: The Sum of Continuous Functions is Continuous

If $f(x,y)$ is continuous at (x_0,y_0), and $g(x,y)$ is continuous at (x_0,y_0), then $f(x,y)+g(x,y)$ is continuous at (x_0,y_0).

Theorem 13.2.3: The Product of Continuous Functions is Continuous

If $g(x)$ is continuous at x_0 and $h(y)$ is continuous at y_0, then $f(x,y) = g(x)h(y)$ is continuous at (x_0,y_0).

Theorem 13.2.4: The Composition of Continuous Functions is Continuous

Let g be a function of two variables from a domain $D \subseteq \mathbb{R}^2$ to a range $R \subseteq R$. Suppose g is continuous at some point $(x_0,y_0) \in D$ and define $z_0 = g(x_0,y_0)$. Let f be a function that maps R to R such that z_0 is in the domain of f. Last, assume f is continuous at z_0. Then $f \circ g$ is continuous at (x_0,y_0) as shown in Figure 13.2.7.

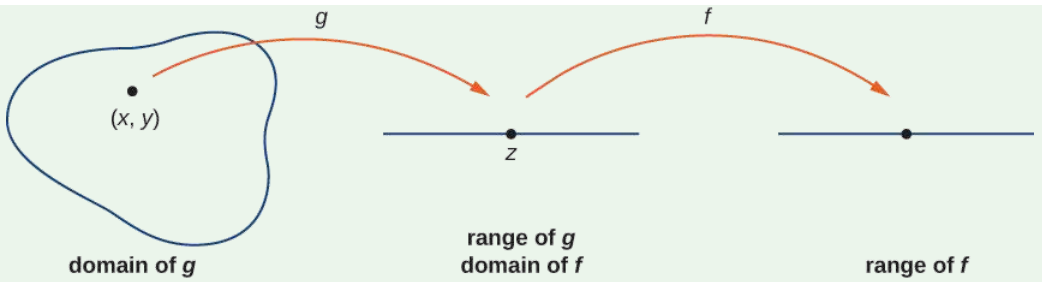


Figure 13.2.7: The composition of two continuous functions is continuous.

Let's now use the previous theorems to show continuity of functions in the following examples.

Example 13.2.5: More Examples of Continuity of a Function of Two Variables

Show that the functions $f(x,y) = 4x^3y^2$ and $g(x,y) = \cos(4x^3y^2)$ are continuous everywhere.

Solution

The polynomials $g(x) = 4x^3$ and $h(y) = y^2$ are continuous at every real number, and therefore by the product of continuous functions theorem, $f(x,y) = 4x^3y^2$ is continuous at every point (x,y) in the xy-plane. Since $f(x,y) = 4x^3y^2$ is continuous at every point (x,y) in the xy-plane and $g(x) = \cos x$ is continuous at every real number x, the continuity of the composition of functions tells us that $g(x,y) = \cos(4x^3y^2)$ is continuous at every point (x,y) in the xy-plane.

Exercise 13.2.5

Show that the functions $f(x,y) = 2x^2y^3 + 3$ and $g(x,y) = (2x^2y^3 + 3)^4$ are continuous everywhere.

Hint

Use the continuity of the sum, product, and composition of two functions.

Answer

The polynomials $g(x) = 2x^2$ and $h(y) = y^3$ are continuous at every real number; therefore, by the product of continuous functions theorem, $f(x,y) = 2x^2y^3$ is continuous at every point (x,y) in the xy-plane. Furthermore, any constant function is continuous everywhere, so $g(x,y) = 3$ is continuous at every point (x,y) in the xy-plane. Therefore, $f(x,y) = 2x^2y^3 + 3$ is continuous at every point (x,y) in the xy-plane. Last, $h(x) = x^4$ is continuous at every real number x, so by the continuity of composite functions theorem $g(x,y) = (2x^2y^3 + 3)^4$ is continuous at every point (x,y) in the xy-plane.

Functions of Three or More Variables

The limit of a function of three or more variables occurs readily in applications. For example, suppose we have a function $f(x,y,z)$ that gives the temperature at a physical location (x,y,z) in three dimensions. Or perhaps a function $g(x,y,z,t)$ can indicate air pressure at a location (x,y,z) at time t. How can we take a limit at a point in $\mathbb{R}^3$? What does it mean to be continuous at a point in four dimensions?

The answers to these questions rely on extending the concept of a δ disk into more than two dimensions. Then, the ideas of the limit of a function of three or more variables and the continuity of a function of three or more variables are very similar to the definitions given earlier for a function of two variables.

Definition: δ-balls

Let (x_0, y_0, z_0) be a point in $\mathbb{R}^3$. Then, a δ-ball in three dimensions consists of all points in $\mathbb{R}^3$ lying at a distance of less than δ from (x_0, y_0, z_0) —that is,

$$\{(x,y,z) \in \mathbb{R}^3 \mid \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2} < \delta\}.$$

To define a δ-ball in higher dimensions, add additional terms under the radical to correspond to each additional dimension. For example, given a point $P=(w_0,x_0,y_0,z_0)$ in $\mathbb{R}^4$, a δ ball around P can be described by

$$\{(w,x,y,z)\in\mathbb{R}^4 \mid \sqrt{(w-w_0)^2+(x-x_0)^2+(y-y_0)^2+(z-z_0)^2} < \delta\}.$$

To show that a limit of a function of three variables exists at a point (x_0,y_0,z_0), it suffices to show that for any point in a δ ball centered at (x_0,y_0,z_0), the value of the function at that point is arbitrarily close to a fixed value (the limit value). All the limit laws for functions of two variables hold for functions of more than two variables as well.

Example 13.2.6: Finding the Limit of a Function of Three Variables

Find

$$\lim_{(x,y,z)\to(4,1,-3)} \frac{x^2y-3z}{2x+5y-z}.$$

Solution

Before we can apply the quotient law, we need to verify that the limit of the denominator is nonzero. Using the difference law, the identity law, and the constant law,

$$\begin{aligned}\lim_{(x,y,z)\to(4,1,-3)}(2x+5y-z) &= 2\left(\lim_{(x,y,z)\to(4,1,-3)}x\right)+5\left(\lim_{(x,y,z)\to(4,1,-3)}y\right)-\left(\lim_{(x,y,z)\to(4,1,-3)}z\right)\\ &= 2(4)+5(1)-(-3)\\ &= 16.\end{aligned}$$

Since this is nonzero, we next find the limit of the numerator. Using the product law, power law, difference law, constant multiple law, and identity law,

$$\begin{aligned}\lim_{(x,y,z)\to(4,1,-3)}(x^2y-3z) &= \left(\lim_{(x,y,z)\to(4,1,-3)}x\right)^2\left(\lim_{(x,y,z)\to(4,1,-3)}y\right)-3\lim_{(x,y,z)\to(4,1,-3)}z\\ &= (4^2)(1)-3(-3)\\ &= 16+9\\ &= 25\end{aligned}$$

Last, applying the quotient law:

$$\lim_{(x,y,z)\to(4,1,-3)}\frac{x^2y-3z}{2x+5y-z} = \frac{\lim\limits_{(x,y,z)\to(4,1,-3)}(x^2y-3z)}{\lim\limits_{(x,y,z)\to(4,1,-3)}(2x+5y-z)} = \frac{25}{16}$$

Exercise 13.2.6

Find

$$\lim_{(x,y,z)\to(4,-1,3)}\sqrt{13-x^2-2y^2+z^2}$$

Hint

Use the limit laws and the continuity of the composition of functions.

Answer

$$\lim_{(x,y,z)\to(4,-1,3)}\sqrt{13-x^2-2y^2+z^2} = 2$$

Key Concepts

- To study limits and continuity for functions of two variables, we use a δ disk centered around a given point.
- A function of several variables has a limit if for any point in a δ ball centered at a point P, the value of the function at that point is arbitrarily close to a fixed value (the limit value).
- The limit laws established for a function of one variable have natural extensions to functions of more than one variable.
- A function of two variables is continuous at a point if the limit exists at that point, the function exists at that point, and the limit and function are equal at that point.

Glossary

boundary point
a point P_0 of R is a boundary point if every δ disk centered around P_0 contains points both inside and outside R

closed set
a set S that contains all its boundary points

connected set
an open set S that cannot be represented as the union of two or more disjoint, nonempty open subsets

δ disk
an open disk of radius δ centered at point (a, b)

δ ball
all points in $\mathbb{R}^3$ lying at a distance of less than δ from (x_0, y_0, z_0)

interior point
a point P_0 of $\mathbb{R}$ is a boundary point if there is a δ disk centered around P_0 contained completely in $\mathbb{R}$

open set
a set S that contains none of its boundary points

region
an open, connected, nonempty subset of $\mathbb{R}^2$

13.2: Limits and Continuity in Higher Dimensions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **14.2: Limits and Continuity** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

13.3: Partial Derivatives

Learning Objectives

- Calculate the partial derivatives of a function of two variables.
- Calculate the partial derivatives of a function of more than two variables.
- Determine the higher-order derivatives of a function of two variables.
- Explain the meaning of a partial differential equation and give an example.

Now that we have examined limits and continuity of functions of two variables, we can proceed to study derivatives. Finding derivatives of functions of two variables is the key concept in this chapter, with as many applications in mathematics, science, and engineering as differentiation of single-variable functions. However, we have already seen that limits and continuity of multivariable functions have new issues and require new terminology and ideas to deal with them. This carries over into differentiation as well.

Derivatives of a Function of Two Variables

When studying derivatives of functions of one variable, we found that one interpretation of the derivative is an instantaneous rate of change of y as a function of x. Leibniz notation for the derivative is dy/dx, which implies that y is the dependent variable and x is the independent variable. For a function $z=f(x,y)$ of two variables, x and y are the independent variables and z is the dependent variable. This raises two questions right away: How do we adapt Leibniz notation for functions of two variables? Also, what is an interpretation of the derivative? The answer lies in partial derivatives.

Definition: Partial Derivatives

Let $f(x,y)$ be a function of two variables. Then the *partial derivative* of f with respect to x, written as $\partial f/\partial x$, or f_x, is defined as

$$\frac{\partial f}{\partial x}=f_x(x,y)=\lim_{h\to 0}\frac{f(x+h,y)-f(x,y)}{h} \tag{13.3.1}$$

The partial derivative of f with respect to y, written as $\partial f/\partial y$, or f_y, is defined as

$$\frac{\partial f}{\partial y}=f_y(x,y)=\lim_{k\to 0}\frac{f(x,y+k)-f(x,y)}{k}. \tag{13.3.2}$$

This definition shows two differences already. First, the notation changes, in the sense that we still use a version of Leibniz notation, but the d in the original notation is replaced with the symbol ∂. (This rounded "d" is usually called "partial," so $\partial f/\partial x$ is spoken as the "partial of f with respect to x.") This is the first hint that we are dealing with partial derivatives. Second, we now have two different derivatives we can take, since there are two different independent variables. Depending on which variable we choose, we can come up with different partial derivatives altogether, and often do.

Example 13.3.1: Calculating Partial Derivatives from the Definition

Use the definition of the partial derivative as a limit to calculate $\partial f/\partial x$ and $\partial f/\partial y$ for the function

$$f(x,y)=x^2-3xy+2y^2-4x+5y-12.$$

Solution

First, calculate $f(x+h,y)$.

$$\begin{aligned} f(x+h,y) &= (x+h)^2-3(x+h)y+2y^2-4(x+h)+5y-12 \\ &= x^2+2xh+h^2-3xy-3hy+2y^2-4x-4h+5y-12. \end{aligned}$$

Next, substitute this into Equation 13.3.1 and simplify:

$$\begin{aligned}\frac{\partial f}{\partial x} &= \lim_{h\to 0}\frac{f(x+h,y)-f(x,y)}{h}\\ &= \lim_{h\to 0}\frac{(x^2+2xh+h^2-3xy-3hy+2y^2-4x-4h+5y-12)-(x^2-3xy+2y^2-4x+5y-12)}{h}\\ &= \lim_{h\to 0}\frac{x^2+2xh+h^2-3xy-3hy+2y^2-4x-4h+5y-12-x^2+3xy-2y^2+4x-5y+12}{h}\\ &= \lim_{h\to 0}\frac{2xh+h^2-3hy-4h}{h}\\ &= \lim_{h\to 0}\frac{h(2x+h-3y-4)}{h}\\ &= \lim_{h\to 0}(2x+h-3y-4)\\ &= 2x-3y-4.\end{aligned}$$

To calculate $\frac{\partial f}{\partial y}$, first calculate $f(x,y+h)$:

$$\begin{aligned}f(x+h,y) &= x^2-3x(y+h)+2(y+h)^2-4x+5(y+h)-12\\ &= x^2-3xy-3xh+2y^2+4yh+2h^2-4x+5y+5h-12.\end{aligned}$$

Next, substitute this into Equation 13.3.2 and simplify:

$$\begin{aligned}\frac{\partial f}{\partial y} &= \lim_{h\to 0}\frac{f(x,y+h)-f(x,y)}{h}\\ &= \lim_{h\to 0}\frac{(x^2-3xy-3xh+2y^2+4yh+2h^2-4x+5y+5h-12)-(x^2-3xy+2y^2-4x+5y-12)}{h}\\ &= \lim_{h\to 0}\frac{x^2-3xy-3xh+2y^2+4yh+2h^2-4x+5y+5h-12-x^2+3xy-2y^2+4x-5y+12}{h}\\ &= \lim_{h\to 0}\frac{-3xh+4yh+2h^2+5h}{h}\\ &= \lim_{h\to 0}\frac{h(-3x+4y+2h+5)}{h}\\ &= \lim_{h\to 0}(-3x+4y+2h+5)\\ &= -3x+4y+5\end{aligned}$$

? Exercise 13.3.1

Use the definition of the partial derivative as a limit to calculate $\partial f/\partial x$ and $\partial f/\partial y$ for the function

$$f(x,y)=4x^2+2xy-y^2+3x-2y+5.$$

Hint

Use Equations 13.3.1 and 13.3.2 from the definition of partial derivatives.

Answer

$$\frac{\partial f}{\partial x}=8x+2y+3$$

$$\frac{\partial f}{\partial y}=2x-2y-2$$

The idea to keep in mind when calculating partial derivatives is to treat all independent variables, other than the variable with respect to which we are differentiating, as constants. Then proceed to differentiate as with a function of a single variable. To see why this is true, first fix y and define $g(x)=f(x,y)$ as a function of x. Then

$$\begin{aligned} g'(x) &= \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} \\ &= \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h} \\ &= \frac{\partial f}{\partial x}. \end{aligned}$$

The same is true for calculating the partial derivative of f with respect to y. This time, fix x and define $h(y) = f(x, y)$ as a function of y. Then

$$\begin{aligned} h'(x) &= \lim_{k \to 0} \frac{h(x+k) - h(x)}{k} \\ &= \lim_{k \to 0} \frac{f(x, y+k) - f(x, y)}{k} \\ &= \frac{\partial f}{\partial y}. \end{aligned}$$

All differentiation rules apply.

Example 13.3.2: Calculating Partial Derivatives

Calculate $\partial f / \partial x$ and $\partial f / \partial y$ for the following functions by holding the opposite variable constant then differentiating:

a. $f(x, y) = x^2 - 3xy + 2y^2 - 4x + 5y - 12$
b. $g(x, y) = \sin(x^2 y - 2x + 4)$

Solution:

a. To calculate $\partial f / \partial x$, treat the variable y as a constant. Then differentiate $f(x, y)$ with respect to x using the sum, difference, and power rules:

$$\begin{aligned} \frac{\partial f}{\partial x} &= \frac{\partial}{\partial x}\left[x^2 - 3xy + 2y^2 - 4x + 5y - 12\right] \\ &= \frac{\partial}{\partial x}[x^2] - \frac{\partial}{\partial x}[3xy] + \frac{\partial}{\partial x}[2y^2] - \frac{\partial}{\partial x}[4x] + \frac{\partial}{\partial x}[5y] - \frac{\partial}{\partial x}[12] \\ &= 2x - 3y + 0 - 4 + 0 - 0 \\ &= 2x - 3y - 4. \end{aligned}$$

The derivatives of the third, fifth, and sixth terms are all zero because they do not contain the variable x, so they are treated as constant terms. The derivative of the second term is equal to the coefficient of x, which is $-3y$. Calculating $\partial f / \partial y$:

$$\begin{aligned} \frac{\partial f}{\partial y} &= \frac{\partial}{\partial y}\left[x^2 - 3xy + 2y^2 - 4x + 5y - 12\right] \\ &= \frac{\partial}{\partial y}[x^2] - \frac{\partial}{\partial y}[3xy] + \frac{\partial}{\partial y}[2y^2] - \frac{\partial}{\partial y}[4x] + \frac{\partial}{\partial y}[5y] - \frac{\partial}{\partial y}[12] \\ &= 0 - 3x + 4y - 0 + 5 - 0 \\ &= -3x + 4y + 5. \end{aligned}$$

These are the same answers obtained in Example 13.3.1.

b. To calculate $\partial g / \partial x$, treat the variable y as a constant. Then differentiate $g(x, y)$ with respect to x using the chain rule and power rule:

$$\begin{aligned}\frac{\partial g}{\partial x} &= \frac{\partial}{\partial x}\left[\sin(x^2y-2x+4)\right] \\ &= \cos(x^2y-2x+4)\frac{\partial}{\partial x}[x^2y-2x+4] \\ &= (2xy-2)\cos(x^2y-2x+4).\end{aligned}$$

To calculate $\partial g/\partial y$, treat the variable x as a constant. Then differentiate $g(x,y)$ with respect to y using the chain rule and power rule:

$$\begin{aligned}\frac{\partial g}{\partial y} &= \frac{\partial}{\partial y}\left[\sin(x^2y-2x+4)\right] \\ &= \cos(x^2y-2x+4)\frac{\partial}{\partial y}[x^2y-2x+4] \\ &= x^2\cos(x^2y-2x+4).\end{aligned}$$

? Exercise 13.3.2

Calculate $\partial f/\partial x$ and $\partial f/\partial y$ for the function

$$f(x,y)=\tan(x^3-3x^2y^2+2y^4)$$

by holding the opposite variable constant, then differentiating.

Hint

Use Equations 13.3.1and 13.3.1from the definition of partial derivatives.

Answer

$$\frac{\partial f}{\partial x}=(3x^2-6xy^2)\sec^2(x^3-3x^2y^2+2y^4)$$
$$\frac{\partial f}{\partial y}=(-6x^2y+8y^3)\sec^2(x^3-3x^2y^2+2y^4)$$

How can we interpret these partial derivatives? Recall that the graph of a function of two variables is a surface in R^3. If we remove the limit from the definition of the partial derivative with respect to x, the difference quotient remains:

$$\frac{f(x+h,y)-f(x,y)}{h}.$$

This resembles the difference quotient for the derivative of a function of one variable, except for the presence of the y variable. Figure 13.3.1illustrates a surface described by an arbitrary function $z=f(x,y)$.

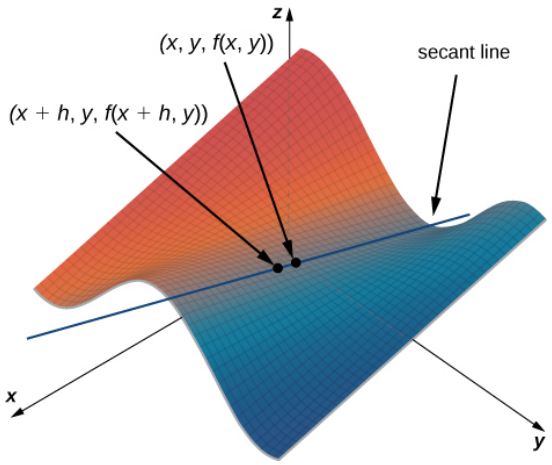


Figure 13.3.1: Secant line passing through the points $(x,y,f(x,y))$ and $(x+h,y,f(x+h,y))$.

In Figure 13.3.1, the value of h is positive. If we graph $f(x,y)$ and $f(x+h,y)$ for an arbitrary point (x,y), then the slope of the secant line passing through these two points is given by

$$\frac{f(x+h,y)-f(x,y)}{h}.$$

This line is parallel to the x-axis. Therefore, the slope of the secant line represents an average rate of change of the function f as we travel parallel to the x-axis. As h approaches zero, the slope of the secant line approaches the slope of the tangent line.

If we choose to change y instead of x by the same incremental value h, then the secant line is parallel to the y-axis and so is the tangent line. Therefore, $\partial f/\partial x$ represents the slope of the tangent line passing through the point $(x,y,f(x,y))$parallel to the x-axis and $\partial f/\partial y$ represents the slope of the tangent line passing through the point $(x,y,f(x,y))$parallel to the y-axis. If we wish to find the slope of a tangent line passing through the same point in any other direction, then we need what are called directional derivatives.

We now return to the idea of contour maps, which we introduced in Functions of Several Variables. We can use a **contour map** to estimate partial derivatives of a function $g(x,y)$.

Example 13.3.3: Partial Derivatives from a Contour Map

Use a contour map to estimate $\partial g/\partial x$ at the point $(\sqrt{5},0)$ for the function

$$g(x,y)=\sqrt{9-x^2-y^2}.$$

Solution

Figure 13.3.2represents a contour map for the function $g(x,y)$.

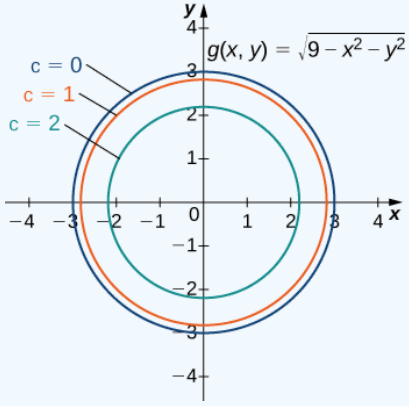


Figure 13.3.2: Contour map for the function $g(x,y)=\sqrt{9-x^2-y^2}$, using $c=0,1,2,$ and $3(c=3$ corresponds to the origin).

The inner circle on the contour map corresponds to $c=2$ and the next circle out corresponds to $c=1$. The first circle is given by the equation $2=\sqrt{9-x^2-y^2}$; the second circle is given by the equation $1=\sqrt{9-x^2-y^2}$. The first equation simplifies to $x^2+y^2=5$ and the second equation simplifies to $x^2+y^2=8$. The x-intercept of the first circle is $(\sqrt{5},0)$ and the x-intercept of the second circle is $(2\sqrt{2},0)$. We can estimate the value of $\partial g/\partial x$ evaluated at the point $(\sqrt{5},0)$ using the slope formula:

$$\begin{aligned}\left.\frac{\partial g}{\partial x}\right|_{(x,y)=(\sqrt{5},0)} &\approx \frac{g(\sqrt{5},0)-g(2\sqrt{2},0)}{\sqrt{5}-2\sqrt{2}}\\ &=\frac{2-1}{\sqrt{5}-2\sqrt{2}}\\ &=\frac{1}{\sqrt{5}-2\sqrt{2}}\approx -1.688.\end{aligned}$$

To calculate the exact value of $\partial g/\partial x$ evaluated at the point $(\sqrt{5},0)$, we start by finding $\partial g/\partial x$ using the chain rule. First, we rewrite the function as

$$g(x,y)=\sqrt{9-x^2-y^2}=(9-x^2-y^2)^{1/2}$$

and then differentiate with respect to x while holding y constant:

$$\frac{\partial g}{\partial x} = \frac{1}{2}(9 - x^2 - y^2)^{-1/2}(-2x)$$
$$= -\frac{x}{\sqrt{9 - x^2 - y^2}}.$$

Next, we evaluate this expression using $x = \sqrt{5}$ and $y = 0$:

$$\frac{\partial g}{\partial x}\Big|_{(x,y)=(\sqrt{5},0)} = -\frac{\sqrt{5}}{\sqrt{9 - (\sqrt{5})^2 - (0)^2}}$$
$$= -\frac{\sqrt{5}}{\sqrt{4}}$$
$$= -\frac{\sqrt{5}}{2} \approx -1.118.$$

The estimate for the partial derivative corresponds to the slope of the secant line passing through the points $(\sqrt{5}, 0, g(\sqrt{5}, 0))$ and $(2\sqrt{2}, 0, g(2\sqrt{2}, 0))$. It represents an approximation to the slope of the tangent line to the surface through the point $(\sqrt{5}, 0, g(\sqrt{5}, 0))$, which is parallel to the x-axis.

? Exercise 13.3.3

Use a contour map to estimate $\partial f/\partial y$ at point $(0, \sqrt{2})$ for the function

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

Compare this with the exact answer.

Hint

Create a contour map for f using values of c from -3 to 3. Which of these curves passes through point $(0, \sqrt{2})$?

Answer

Using the curves corresponding to $c = -2$ and $c = -3$, we obtain

$$\frac{\partial f}{\partial y}\Big|_{(x,y)=(0,\sqrt{2})} \approx \frac{f(0, \sqrt{3}) - f(0, \sqrt{2})}{\sqrt{3} - \sqrt{2}}$$
$$= \frac{-3 - (-2)}{\sqrt{3} - \sqrt{2}} \cdot \frac{\sqrt{3} + \sqrt{2}}{\sqrt{3} + \sqrt{2}}$$
$$= -\sqrt{3} - \sqrt{2} \approx -3.146.$$

The exact answer is

$$\frac{\partial f}{\partial y}\Big|_{(x,y)=(0,\sqrt{2})} = (-2y|_{(x,y)=(0,\sqrt{2})} = -2\sqrt{2} \approx -2.828.$$

Functions of More Than Two Variables

Suppose we have a function of three variables, such as $w = f(x, y, z)$. We can calculate partial derivatives of w with respect to any of the independent variables, simply as extensions of the definitions for partial derivatives of functions of two variables.

Definition: Partial Derivatives

Let $f(x, y, z)$ be a function of three variables. Then, the partial derivative of f with respect to x, written as $\partial f/\partial x$, or f_x, is defined to be

$$\frac{\partial f}{\partial x} = f_x(x, y, z) = \lim_{h \to 0} \frac{f(x + h, y, z) - f(x, y, z)}{h}. \tag{13.3.3}$$

The *partial derivative* of f with *respect* to y, written as $\partial f/\partial y$, or f_y, is defined to be

$$\frac{\partial f}{\partial y} = f_y(x,y,z) = \lim_{k\to 0} \frac{f(x,y+k,z) - f(x,y,z)}{k}. \tag{13.3.4}$$

The *partial derivative* of f with respect to z, written as $\partial f/\partial z$, or f_z, is defined to be

$$\frac{\partial f}{\partial z} = f_z(x,y,z) = \lim_{m\to 0} \frac{f(x,y,z+m) - f(x,y,z)}{m}. \tag{13.3.5}$$

We can calculate a partial derivative of a function of three variables using the same idea we used for a function of two variables. For example, if we have a function f of x, y, and z, and we wish to calculate $\partial f/\partial x$, then we treat the other two independent variables as if they are constants, then differentiate with respect to x.

Example 13.3.4: Calculating Partial Derivatives for a Function of Three Variables

Use the limit definition of partial derivatives to calculate $\partial f/\partial x$ for the function

$$f(x,y,z) = x^2 - 3xy + 2y^2 - 4xz + 5yz^2 - 12x + 4y - 3z.$$

Then, find $\partial f/\partial y$ and $\partial f/\partial z$ by setting the other two variables constant and differentiating accordingly.

Solution:

We first calculate $\partial f/\partial x$ using Equation 13.3.3, then we calculate the other two partial derivatives by holding the remaining variables constant. To use the equation to find $\partial f/\partial x$, we first need to calculate $f(x+h,y,z)$:

$$\begin{aligned} f(x+h,y,z) &= (x+h)^2 - 3(x+h)y + 2y^2 - 4(x+h)z + 5yz^2 - 12(x+h) + 4y - 3z \\ &= x^2 + 2xh + h^2 - 3xy - 3xh + 2y^2 - 4xz - 4hz + 5yz^2 - 12x - 12h + 4y - 3z \end{aligned}$$

and recall that $f(x,y,z) = x^2 - 3xy + 2y^2 - 4zx + 5yz^2 - 12x + 4y - 3z$. Next, we substitute these two expressions into the equation:

$$\begin{aligned} \frac{\partial f}{\partial x} &= \lim_{h\to 0}\left[\frac{\begin{array}{l} x^2 + 2xh + h^2 - 3xy - 3hy + 2y^2 - 4xz - 4hz + 5yz^2 - 12x - 12h + 4y - 3zh - x^2 - 3xy + 2y^2 - 4xz + 5yz^2 \\ - 12x + 4y - 3z \end{array}}{h}\right] \\ &= \lim_{h\to 0}\left[\frac{2xh + h^2 - 3hy - 4hz - 12h}{h}\right] \\ &= \lim_{h\to 0}\left[\frac{h(2x + h - 3y - 4z - 12)}{h}\right] \\ &= \lim_{h\to 0}(2x + h - 3y - 4z - 12) \\ &= 2x - 3y - 4z - 12. \end{aligned}$$

Then we find $\partial f/\partial y$ by holding x and z constant. Therefore, any term that does not include the variable y is constant, and its derivative is zero. We can apply the sum, difference, and power rules for functions of one variable:

$$\begin{aligned} &\frac{\partial}{\partial y}\left[x^2 - 3xy + 2y^2 - 4xz + 5yz^2 - 12x + 4y - 3z\right] \\ &= \frac{\partial}{\partial y}[x^2] - \frac{\partial}{\partial y}[3xy] + \frac{\partial}{\partial y}[2y^2] - \frac{\partial}{\partial y}[4xz] + \frac{\partial}{\partial y}[5yz^2] - \frac{\partial}{\partial y}[12x] + \frac{\partial}{\partial y}[4y] - \frac{\partial}{\partial z}[3z] \\ &= 0 - 3x + 4y - 0 + 5z^2 - 0 + 4 - 0 \\ &= -3x + 4y + 5z^2 + 4. \end{aligned}$$

To calculate $\partial f/\partial z$, we hold x and y constant and apply the sum, difference, and power rules for functions of one variable:

$$\begin{aligned}
&\frac{\partial}{\partial z}[x^2-3xy+2y^2-4xz+5yz^2-12x+4y-3z] \\
&=\frac{\partial}{\partial z}[x^2]-\frac{\partial}{\partial z}[3xy]+\frac{\partial}{\partial z}[2y^2]-\frac{\partial}{\partial z}[4xz]+\frac{\partial}{\partial z}[5yz^2]-\frac{\partial}{\partial z}[12x]+\frac{\partial}{\partial z}[4y]-\frac{\partial}{\partial z}[3z] \\
&=0-0+0-4x+10yz-0+0-3 \\
&=-4x+10yz-3
\end{aligned}$$

Exercise 13.3.4

Use the limit definition of partial derivatives to calculate $\partial f/\partial x$ for the function

$$f(x,y,z)=2x^2-4x^2y+2y^2+5xz^2-6x+3z-8.$$

Then find $\partial f/\partial y$ and $\partial f/\partial z$ by setting the other two variables constant and differentiating accordingly.

Hint

Use the strategy in the preceding example.

Answer

$$\frac{\partial f}{\partial x}=4x-8xy+5z^2-6,\ \frac{\partial f}{\partial y}=-4x^2+4y,\ \frac{\partial f}{\partial z}=10xz+3$$

Example 13.3.5: Calculating Partial Derivatives for a Function of Three Variables

Calculate the three partial derivatives of the following functions.

a. $f(x,y,z)=\dfrac{x^2y-4xz+y^2}{x-3yz}$

b. $g(x,y,z)=\sin(x^2y-z)+\cos(x^2-yz)$

Solution

In each case, treat all variables as constants except the one whose partial derivative you are calculating.

a.

$$\begin{aligned}
\frac{\partial f}{\partial x}&=\frac{\partial}{\partial x}\left[\frac{x^2y-4xz+y^2}{x-3yz}\right] \\
&=\frac{\frac{\partial}{\partial x}(x^2y-4xz+y^2)(x-3yz)-(x^2y-4xz+y^2)\frac{\partial}{\partial x}(x-3yz)}{(x-3yz)^2} \\
&=\frac{(2xy-4z)(x-3yz)-(x^2y-4xz+y^2)(1)}{(x-3yz)^2} \\
&=\frac{2x^2y-6xy^2z-4xz+12yz^2-x^2y+4xz-y^2}{(x-3yz)^2} \\
&=\frac{x^2y-6xy^2z-4xz+12yz^2+4xz-y^2}{(x-3yz)^2}
\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial y} &= \frac{\partial}{\partial y}\left[\frac{x^2y-4xz+y^2}{x-3yz}\right] \\ &= \frac{\frac{\partial}{\partial y}(x^2y-4xz+y^2)(x-3yz)-(x^2y-4xz+y^2)\frac{\partial}{\partial y}(x-3yz)}{(x-3yz)^2} \\ &= \frac{(x^2+2y)(x-3yz)-(x^2y-4xz+y^2)(-3z)}{(x-3yz)^2} \\ &= \frac{x^3-3x^2yz+2xy-6y^2z+3x^2yz-12xz^2+3y^2z}{(x-3yz)^2} \\ &= \frac{x^3+2xy-3y^2z-12xz^2}{(x-3yz)^2}\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial z} &= \frac{\partial}{\partial z}\left[\frac{x^2y-4xz+y^2}{x-3yz}\right] \\ &= \frac{\frac{\partial}{\partial z}(x^2y-4xz+y^2)(x-3yz)-(x^2y-4xz+y^2)\frac{\partial}{\partial z}(x-3yz)}{(x-3yz)^2} \\ &= \frac{(-4x)(x-3yz)-(x^2y-4xz+y^2)(-3y)}{(x-3yz)^2} \\ &= \frac{-4x^2+12xyz+3x^2y^2-12xyz+3y^3}{(x-3yz)^2} \\ &= \frac{-4x^2+3x^2y^2+3y^3}{(x-3yz)^2}\end{aligned}$$

b.

$$\begin{aligned}\frac{\partial f}{\partial x} &= \frac{\partial}{\partial x}\left[\sin(x^2y-z)+\cos(x^2-yz)\right] \\ &= (\cos(x^2y-z))\frac{\partial}{\partial x}(x^2y-z)-(\sin(x^2-yz))\frac{\partial}{\partial x}(x^2-yz) \\ &= 2xy\cos(x^2y-z)-2x\sin(x^2-yz)\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial y} &= \frac{\partial}{\partial y}[\sin(x^2y-z)+\cos(x^2-yz)] \\ &= (\cos(x^2y-z))\frac{\partial}{\partial y}(x^2y-z)-(\sin(x^2-yz))\frac{\partial}{\partial y}(x^2-yz) \\ &= x^2\cos(x^2y-z)+z\sin(x^2-yz)\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial z} &= \frac{\partial}{\partial z}[\sin(x^2y-z)+\cos(x^2-yz)] \\ &= (\cos(x^2y-z))\frac{\partial}{\partial z}(x^2y-z)-(\sin(x^2-yz))\frac{\partial}{\partial z}(x^2-yz) \\ &= -\cos(x^2y-z)+y\sin(x^2-yz)\end{aligned}$$

? Exercise 13.3.5

Calculate $\partial f/\partial x, \partial f/\partial y,$ and $\partial f/\partial z$ for the function

$$f(x,y,z)=\sec(x^2y)-\tan(x^3yz^2).$$

Hint

Use the strategy in the preceding example.

Answer

$$\frac{\partial f}{\partial x} = 2xy\sec(x^2y)\tan(x^2y) - 3x^2yz^2\sec^2(x^3yz^2)$$
$$\frac{\partial f}{\partial y} = x^2\sec(x^2y)\tan(x^2y) - x^3z^2\sec^2(x^3yz^2)$$
$$\frac{\partial f}{\partial z} = -2x^3yz\sec^2(x^3yz^2)$$

Higher-Order Partial Derivatives

Consider the function

$$f(x,y) = 2x^3 - 4xy^2 + 5y^3 - 6xy + 5x - 4y + 12.$$

Its partial derivatives are

$$\frac{\partial f}{\partial x} = 6x^2 - 4y^2 - 6y + 5$$

and

$$\frac{\partial f}{\partial y} = -8xy + 15y^2 - 6x - 4.$$

Each of these partial derivatives is a function of two variables, so we can calculate partial derivatives of these functions. Just as with derivatives of single-variable functions, we can call these *second-order derivatives*, third-order derivatives, and so on. In general, they are referred to as **higher-order partial derivatives.** There are four second-order partial derivatives for any function (provided they all exist):

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\left[\frac{\partial f}{\partial x}\right]$$
$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial x}\right]$$
$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\left[\frac{\partial f}{\partial y}\right]$$
$$\frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial y}\right].$$

An alternative notation for each is f_{xx}, f_{xy}, f_{yx}, and f_{yy}, respectively. Higher-order partial derivatives calculated with respect to different variables, such as f_{xy} and f_{yx}, are commonly called **mixed partial derivatives.**

Example 13.3.6: Calculating Second Partial Derivatives

Calculate all four second partial derivatives for the function

$$f(x,y) = xe^{-3y} + \sin(2x - 5y). \tag{13.3.6}$$

Solution:

To calculate $\frac{\partial^2 f}{\partial x^2}$ and $\frac{\partial^2 f}{\partial y \partial x}$, we first calculate $\partial f/\partial x$:

$$\frac{\partial f}{\partial x} = e^{-3y} + 2\cos(2x - 5y). \tag{13.3.7}$$

To calculate $\frac{\partial^2 f}{\partial x^2}$, differentiate $\partial f/\partial x$ (Equation 13.3.7) with respect to x:

$$\begin{aligned}\frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x}\left[\frac{\partial f}{\partial x}\right] \\ &= \frac{\partial}{\partial x}[e^{-3y} + 2\cos(2x-5y)] \\ &= -4\sin(2x-5y).\end{aligned}$$

To calculate $\dfrac{\partial^2 f}{\partial y \partial x}$, differentiate $\partial f/\partial x$ (Equation 13.3.7) with respect to y:

$$\begin{aligned}\frac{\partial^2 f}{\partial y\,\partial x} &= \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial x}\right] \\ &= \frac{\partial}{\partial y}[e^{-3y} + 2\cos(2x-5y)] \\ &= -3e^{-3y} + 10\sin(2x-5y).\end{aligned}$$

To calculate $\dfrac{\partial^2 f}{\partial x \partial y}$ and $\dfrac{\partial^2 f}{\partial y^2}$, first calculate $\partial f/\partial y$:

$$\frac{\partial f}{\partial y} = -3xe^{-3y} - 5\cos(2x-5y). \tag{13.3.8}$$

To calculate $\dfrac{\partial^2 f}{\partial x \partial y}$, differentiate $\partial f/\partial y$ (Equation 13.3.8) with respect to x:

$$\begin{aligned}\frac{\partial^2 f}{\partial x \partial y} &= \frac{\partial}{\partial x}\left[\frac{\partial f}{\partial y}\right] \\ &= \frac{\partial}{\partial x}[-3xe^{-3y} - 5\cos(2x-5y)] \\ &= -3e^{-3y} + 10\sin(2x-5y).\end{aligned}$$

To calculate $\dfrac{\partial^2 f}{\partial y^2}$, differentiate $\partial f/\partial y$ (Equation 13.3.8) with respect to y:

$$\begin{aligned}\frac{\partial^2 f}{\partial y^2} &= \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial y}\right] \\ &= \frac{\partial}{\partial y}[-3xe^{-3y} - 5\cos(2x-5y)] \\ &= 9xe^{-3y} - 25\sin(2x-5y).\end{aligned}$$

? Exercise 13.3.6

Calculate all four second partial derivatives for the function

$$f(x,y) = \sin(3x-2y) + \cos(x+4y).$$

Hint

Follow the same steps as in the previous example.

Answer

$$\frac{\partial^2 f}{\partial x^2} = -9\sin(3x-2y) - \cos(x+4y)$$

$$\frac{\partial^2 f}{\partial y \partial x} = 6\sin(3x-2y) - 4\cos(x+4y)$$

$$\frac{\partial^2 f}{\partial x \partial y} = 6\sin(3x-2y) - 4\cos(x+4y)$$

$$\frac{\partial^2 f}{\partial y^2} = -4\sin(3x-2y) - 16\cos(x+4y)$$

At this point we should notice that, in both Example 13.3.6 and the checkpoint, it was true that $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y}$. Under certain conditions, this is always true. In fact, it is a direct consequence of the following theorem.

Theorem 13.3.1: Equality of Mixed Partial Derivatives (Clairaut's Theorem)

Suppose that $f(x,y)$ is defined on an open disk D that contains the point (a,b). If the functions f_{xy} and f_{yx} are continuous on D, then $f_{xy}(a,b) = f_{yx}(a,b)$.

Clairaut's theorem guarantees that as long as mixed second-order derivatives are continuous, the order in which we choose to differentiate the functions (i.e., which variable goes first, then second, and so on) does not matter. It can be extended to higher-order derivatives as well. The proof of Clairaut's theorem can be found in most advanced calculus books.

Two other second-order partial derivatives can be calculated for any function $f(x,y)$. The partial derivative f_{xx} is equal to the partial derivative of f_x with respect to x, and f_{yy} is equal to the partial derivative of f_y with respect to y.

Partial Differential Equations

Previously, we studied differential equations in which the unknown function had one independent variable. A partial differential equation is an equation that involves an unknown function of more than one independent variable and one or more of its partial derivatives. Examples of partial differential equations are

$$u_t = c^2(u_{xx} + u_{yy})$$
heat equation in two dimensions

$$u_{tt} = c^2(u_{xx} + u_{yy})$$
wave equation in two dimensions

$$u_{xx} + u_{yy} = 0$$
Laplace's equation in two dimensions

In the heat and wave equations, the unknown function u has three independent variables: t, x, and y with c is an arbitrary constant. The independent variables x and y are considered to be spatial variables, and the variable t represents time. In Laplace's equation, the unknown function u has two independent variables x and y.

Example 13.3.7: A Solution to the Wave Equation

Verify that

$$u(x,y,t) = 5\sin(3\pi x)\sin(4\pi y)\cos(10\pi t)$$

is a solution to the wave equation

$$u_{tt} = 4(u_{xx} + u_{yy}). \tag{13.3.9}$$

Solution

First, we calculate u_{tt}, u_{xx}, and u_{yy}:

$$\begin{aligned} u_{tt}(x,y,t) &= \frac{\partial}{\partial t}\left[\frac{\partial u}{\partial t}\right] \\ &= \frac{\partial}{\partial t}[5\sin(3\pi x)\sin(4\pi y)(-10\pi\sin(10\pi t))] \\ &= \frac{\partial}{\partial t}\left[-50\pi\sin(3\pi x)\sin(4\pi y)\sin(10\pi t)\right] \\ &= -500\pi^2\sin(3\pi x)\sin(4\pi y)\cos(10\pi t) \end{aligned}$$

$$\begin{aligned} u_{xx}(x,y,t) &= \frac{\partial}{\partial x}\left[\frac{\partial u}{\partial x}\right] \\ &= \frac{\partial}{\partial x}\left[15\pi\cos(3\pi x)\sin(4\pi y)\cos(10\pi t)\right] \\ &= -45\pi^2\sin(3\pi x)\sin(4\pi y)\cos(10\pi t) \\ u_{yy}(x,y,t) &= \frac{\partial}{\partial y}\left[\frac{\partial u}{\partial y}\right] \\ &= \frac{\partial}{\partial y}\left[5\sin(3\pi x)(4\pi\cos(4\pi y))\cos(10\pi t)\right] \\ &= \frac{\partial}{\partial y}\left[20\pi\sin(3\pi x)\cos(4\pi y)\cos(10\pi t)\right] \\ &= -80\pi^2\sin(3\pi x)\sin(4\pi y)\cos(10\pi t). \end{aligned}$$

Next, we substitute each of these into the right-hand side of Equation 13.3.9 and simplify:

$$\begin{aligned} 4(u_{xx}+u_{yy}) &= 4(-45\pi^2\sin(3\pi x)\sin(4\pi y)\cos(10\pi t) + -80\pi^2\sin(3\pi x)\sin(4\pi y)\cos(10\pi t)) \\ &= 4(-125\pi^2\sin(3\pi x)\sin(4\pi y)\cos(10\pi t)) \\ &= -500\pi^2\sin(3\pi x)\sin(4\pi y)\cos(10\pi t) \\ &= u_{tt}. \end{aligned}$$

This verifies the solution.

? Exercise 13.3.7: A Solution to the Heat Equation

Verify that

$$u(x,y,t) = 2\sin\left(\frac{x}{3}\right)\sin\left(\frac{y}{4}\right)e^{-25t/16}$$

is a solution to the heat equation

$$u_t = 9(u_{xx}+u_{yy}).$$

Hint

Calculate the partial derivatives and substitute into the right-hand side.

Answer

TBA

Since the solution to the two-dimensional heat equation is a function of three variables, it is not easy to create a visual representation of the solution. We can graph the solution for fixed values of t, which amounts to snapshots of the heat distributions at fixed times. These snapshots show how the heat is distributed over a two-dimensional surface as time progresses. The graph of the preceding solution at time $t = 0$ appears in Figure 13.3.3. As time progresses, the extremes level out, approaching zero as t approaches infinity.

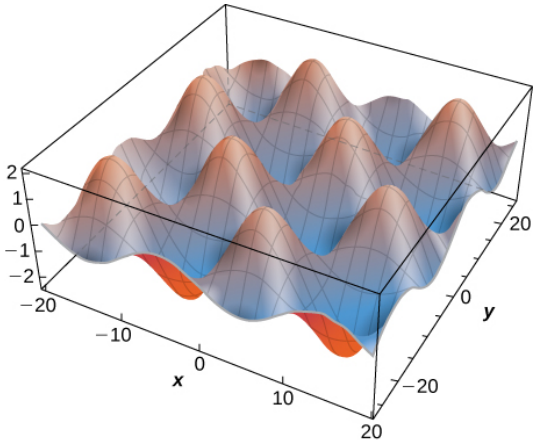


Figure 13.3.3

If we consider the heat equation in one dimension, then it is possible to graph the solution over time. The heat equation in one dimension becomes

$$u_t = c^2 u_{xx},$$

where c^2 represents the thermal diffusivity of the material in question. A solution of this differential equation can be written in the form

$$u_m(x,t) = e^{-\pi^2 m^2 c^2 t} \sin(m\pi x)$$

where m is any positive integer. A graph of this solution using $m = 1$ appears in Figure 13.3.4, where the initial temperature distribution over a wire of length 1 is given by $u(x,0) = \sin \pi x$. Notice that as time progresses, the wire cools off. This is seen because, from left to right, the highest temperature (which occurs in the middle of the wire) decreases and changes color from red to blue.

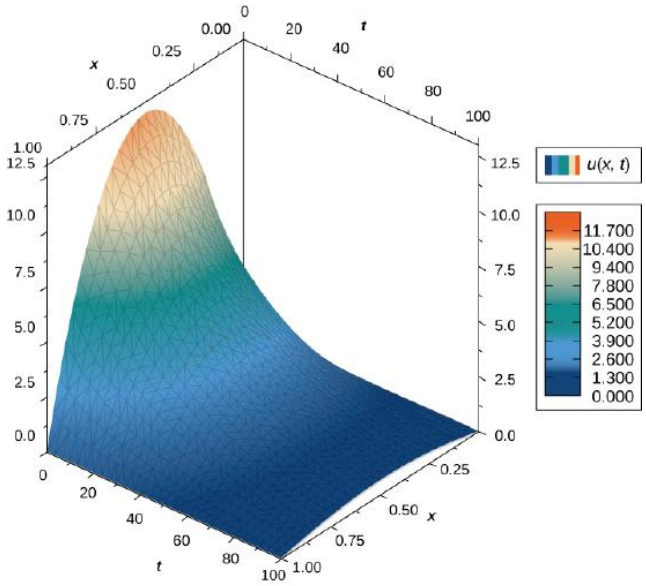
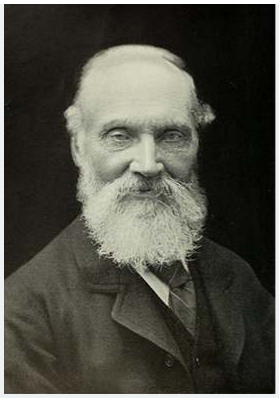


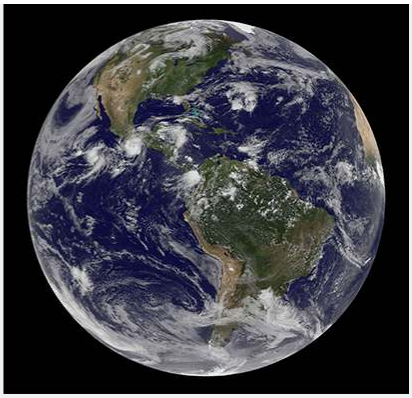
Figure 13.3.4: Graph of a solution of the heat equation in one dimension over time.

Lord Kelvin and the Age of Earth

During the late 1800s, the scientists of the new field of geology were coming to the conclusion that Earth must be “millions and millions” of years old. At about the same time, Charles Darwin had published his treatise on evolution. Darwin’s view was that evolution needed many millions of years to take place, and he made a bold claim that the Weald chalk fields, where important fossils were found, were the result of 300 million years of erosion.



(a)



(b)

Figure 13.3.5: (a) William Thomson (Lord Kelvin), 1824-1907, was a British physicist and electrical engineer; (b) Kelvin used the heat diffusion equation to estimate the age of Earth (credit: modification of work by NASA).

At that time, eminent physicist William Thomson (Lord Kelvin) used an important partial differential equation, known as the heat diffusion equation, to estimate the age of Earth by determining how long it would take Earth to cool from molten rock to what we had at that time. His conclusion was a range of **20** to **400** million years, but most likely about **50** million years. For many decades, the proclamations of this irrefutable icon of science did not sit well with geologists or with Darwin.

- Read Kelvin's paper on estimating the age of the Earth.

Kelvin made reasonable assumptions based on what was known in his time, but he also made several assumptions that turned out to be wrong. One incorrect assumption was that Earth is solid and that the cooling was therefore via conduction only, hence justifying the use of the diffusion equation. But the most serious error was a forgivable one—omission of the fact that Earth contains radioactive elements that continually supply heat beneath Earth's mantle. The discovery of radioactivity came near the end of Kelvin's life and he acknowledged that his calculation would have to be modified.

Kelvin used the simple one-dimensional model applied only to Earth's outer shell, and derived the age from graphs and the roughly known temperature gradient near Earth's surface. Let's take a look at a more appropriate version of the diffusion equation in radial coordinates, which has the form

$$\frac{\partial T}{\partial t} = K\left[\frac{\partial^2 T}{\partial^2 r} + \frac{2}{r}\frac{\partial T}{\partial r}\right] \tag{13.3.10}$$

.

Here, $T(r,t)$ is temperature as a function of r (measured from the center of Earth) and time t. K is the heat conductivity—for molten rock, in this case. The standard method of solving such a partial differential equation is by separation of variables, where we express the solution as the product of functions containing each variable separately. In this case, we would write the temperature as

$$T(r,t) = R(r)f(t).$$

1. Substitute this form into Equation 13.3.10and, noting that $f(t)$ is constant with respect to distance (r) and $R(r)$ is constant with respect to time (t), show that

$$\frac{1}{f}\frac{\partial f}{\partial t} = \frac{K}{R}\left[\frac{\partial^2 R}{\partial r^2} + \frac{2}{r}\frac{\partial R}{\partial r}\right].$$

2. This equation represents the separation of variables we want. The left-hand side is only a function of t and the right-hand side is only a function of r, and they must be equal for all values of r and t. Therefore, they both must be equal to a constant. Let's call that constant $-\lambda^2$. (The convenience of this choice is seen on substitution.) So, we have

$$\frac{1}{f}\frac{\partial f}{\partial t} = -\lambda^2 \text{and} \frac{K}{R}\left[\frac{\partial^2 R}{\partial r^2} + \frac{2}{r}\frac{\partial R}{\partial r}\right] = -\lambda^2.$$

3. Now, we can verify through direct substitution for each equation that the solutions are $f(t) = Ae^{-\lambda^2 t}$ and $R(r) = B\left(\frac{\sin \alpha r}{r}\right) + C\left(\frac{\cos \alpha r}{r}\right)$, where $\alpha = \lambda/\sqrt{K}$. Note that $f(t) = Ae^{+\lambda n^2 t}$ is also a valid solution, so we could have chosen $+\lambda^2$ for our constant. Can you see why it would not be valid for this case as time increases?
4. Let's now apply boundary conditions.
   a. The temperature must be finite at the center of Earth, $r = 0$. Which of the two constants, B or C, must therefore be zero to keep R finite at $r = 0$? (Recall that $\sin(\alpha r)/r \to \alpha =$ as $r \to 0$, but $\cos(\alpha r)/r$ behaves very differently.)
   b. Kelvin argued that when magma reaches Earth's surface, it cools very rapidly. A person can often touch the surface within weeks of the flow. Therefore, the surface reached a moderate temperature very early and remained nearly constant at a surface temperature T_s. For simplicity, let's set $T = 0$ at $r = R_E$ and find α such that this is the temperature there for all time t. (Kelvin took the value to be $300K \approx 80^\circ F$. We can add this $300K$ constant to our solution later.) For this to be true, the sine argument must be zero at $r = R_E$. Note that α has an infinite series of values that satisfies this condition. Each value of α represents a valid solution (each with its own value for A). The total or general solution is the sum of all these solutions.
   c. At $t = 0$, we assume that all of Earth was at an initial hot temperature T_0 (Kelvin took this to be about $7000K$.) The application of this boundary condition involves the more advanced application of Fourier coefficients. As noted in part b. each value of α_n represents a valid solution, and the general solution is a sum of all these solutions. This results in a series solution:

$$T(r,t) = \left(\frac{T_0 R_E}{\pi}\right) \sum_n \frac{(-1)^{n-1}}{n} e^{-\lambda n^2 t} \frac{\sin(\alpha_n r)}{r}$$

   where $\alpha_n = n\pi/R_E$.

Note how the values of α_n come from the boundary condition applied in part b. The term $\frac{-1^{n-1}}{n}$ is the constant A_n for each term in the series, determined from applying the Fourier method. Letting $\beta = \frac{\pi}{R_E}$, examine the first few terms of this solution shown here and note how λ^2 in the exponential causes the higher terms to decrease quickly as time progresses:

$$T(r,t) = \frac{T_0 R_E}{\pi r}\left(e^{-K\beta^2 t}(\sin \beta r) - \frac{1}{2}e^{-4K\beta^2 t}(\sin 2\beta r) + \frac{1}{3}e^{-9K\beta^2 t}(\sin 3\beta r) - \frac{1}{4}e^{-16K\beta^2 t}(\sin 4\beta r) + \frac{1}{5}e^{-25K\beta^2 t}(\sin 5\beta r)\ldots\right).$$

Near time $t = 0$, many terms of the solution are needed for accuracy. Inserting values for the conductivity K and $\beta = \pi/R_E$ for time approaching merely thousands of years, only the first few terms make a significant contribution. Kelvin only needed to look at the solution near Earth's surface (Figure 13.3.6) and, after a long time, determine what time best yielded the estimated temperature gradient known during his era ($1^\circ F$ increase per $50ft$). He simply chose a range of times with a gradient close to this value. In Figure 13.3.6, the solutions are plotted and scaled, with the $300 - K$ surface temperature added. Note that the center of Earth would be relatively cool. At the time, it was thought Earth must be solid.

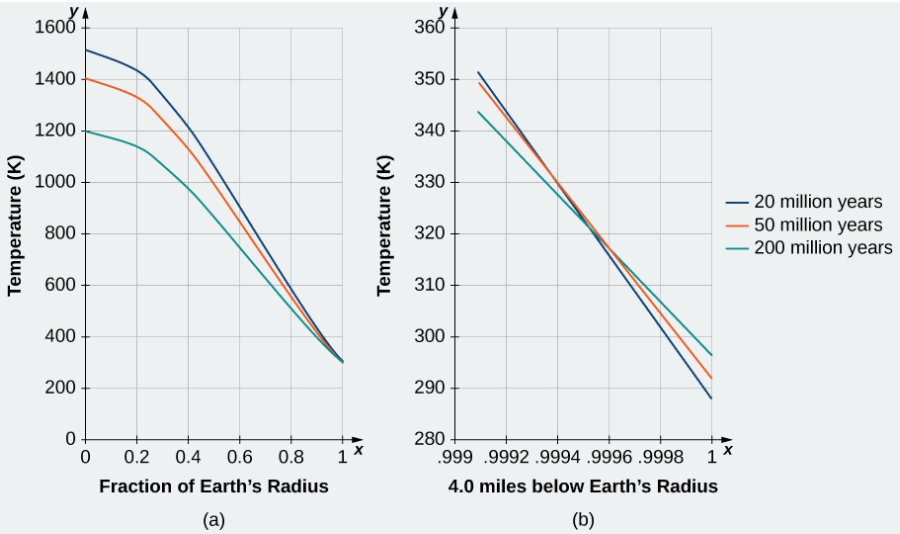


Figure 13.3.6: Temperature versus radial distance from the center of Earth. (a) Kelvin's results, plotted to scale. (b) A close-up of the results at a depth of 4.0 miles below Earth's surface.

Epilog

On May **20, 1904,** physicist **Ernest Rutherford** spoke at the Royal Institution to announce a revised calculation that included the contribution of radioactivity as a source of Earth's heat. In Rutherford's own words:

> "I came into the room, which was half-dark, and presently spotted Lord Kelvin in the audience, and realized that I was in for trouble at the last part of my speech dealing with the age of the Earth, where my views conflicted with his. To my relief, Kelvin fell fast asleep, but as I came to the important point, I saw the old bird sit up, open an eye and cock a baleful glance at me.
>
> Then a sudden inspiration came, and I said Lord Kelvin had limited the age of the Earth, provided no new source [of heat] was discovered. That prophetic utterance referred to what we are now considering tonight, radium! Behold! The old boy beamed upon me."

Rutherford calculated an age for Earth of about **500** million years. Today's accepted value of Earth's age is about **4.6** billion years.

Key Concepts

- A partial derivative is a derivative involving a function of more than one independent variable.
- To calculate a partial derivative with respect to a given variable, treat all the other variables as constants and use the usual differentiation rules.
- Higher-order partial derivatives can be calculated in the same way as higher-order derivatives.

Key Equations

Partial derivative of f with respect to x

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

Partial derivative of f with respect to y

$$\frac{\partial f}{\partial y} = \lim_{k \to 0} \frac{f(x, y+k) - f(x, y)}{k}$$

Glossary

higher-order partial derivatives
second-order or higher partial derivatives, regardless of whether they are mixed partial derivatives

mixed partial derivatives
second-order or higher partial derivatives, in which at least two of the differentiations are with respect to different variables

partial derivative

a derivative of a function of more than one independent variable in which all the variables but one are held constant

partial differential equation

an equation that involves an unknown function of more than one independent variable and one or more of its partial derivatives

---

13.3: Partial Derivatives is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **14.3: Partial Derivatives** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

13.4: The Chain Rule

Learning Objectives

- State the chain rules for one or two independent variables.
- Use tree diagrams as an aid to understanding the chain rule for several independent and intermediate variables.
- Perform implicit differentiation of a function of two or more variables.

In single-variable calculus, we found that one of the most useful differentiation rules is the chain rule, which allows us to find the derivative of the composition of two functions. The same thing is true for multivariable calculus, but this time we have to deal with more than one form of the chain rule. In this section, we study extensions of the chain rule and learn how to take derivatives of compositions of functions of more than one variable.

Chain Rules for One or Two Independent Variables

Recall that the chain rule for the derivative of a composite of two functions can be written in the form

$$\frac{d}{dx}\Big(f(g(x))\Big) = f'\big(g(x)\big)g'(x).$$

In this equation, both $f(x)$ and $g(x)$ are functions of one variable. Now suppose that f is a function of two variables and g is a function of one variable. Or perhaps they are both functions of two variables, or even more. How would we calculate the derivative in these cases? The following theorem gives us the answer for the case of one independent variable.

Theorem 13.4.1: Chain Rule for One Independent Variable

Suppose that $x = g(t)$ and $y = h(t)$ are differentiable functions of t and $z = f(x, y)$ is a differentiable function of x and y. Then $z = f(x(t), y(t))$ is a differentiable function of t and

$$\frac{dz}{dt} = \frac{\partial z}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial z}{\partial y} \cdot \frac{dy}{dt}, \tag{13.4.1}$$

where the ordinary derivatives are evaluated at t and the partial derivatives are evaluated at (x, y).

Proof

The proof of this theorem uses the definition of differentiability of a function of two variables. Suppose that f is differentiable at the point $P(x_0, y_0)$, where $x_0 = g(t_0)$ and $y_0 = h(t_0)$ for a fixed value of t_0. We wish to prove that $z = f\big(x(t), y(t)\big)$ is differentiable at $t = t_0$ and that Equation 13.4.1 holds at that point as well.

Since f is differentiable at P, we know that

$$z(t) = f(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) + E(x, y),$$

where

$$\lim_{(x,y)\to(x_0,y_0)} \frac{E(x, y)}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} = 0.$$

We then subtract $z_0 = f(x_0, y_0)$ from both sides of this equation:

$$\begin{aligned} z(t) - z(t_0) &= f(x(t), y(t)) - f(x(t_0), y(t_0)) \\ &= f_x(x_0, y_0)(x(t) - x(t_0)) + f_y(x_0, y_0)(y(t) - y(t_0)) + E(x(t), y(t)). \end{aligned}$$

Next, we divide both sides by $t - t_0$:

$$\frac{z(t) - z(t_0)}{t - t_0} = f_x(x_0, y_0)\frac{x(t) - x(t_0)}{t - t_0} + f_y(x_0, y_0)\frac{y(t) - y(t_0)}{t - t_0} + \frac{E(x(t), y(t))}{t - t_0}.$$

Then we take the limit as t approaches t_0:

$$\lim_{t\to t_0}\frac{z(t)-z(t_0)}{t-t_0} = f_x(x_0,y_0)\lim_{t\to t_0}\left(\frac{x(t)-x(t_0)}{t-t_0}\right) + f_y(x_0,y_0)\lim_{t\to t_0}\left(\frac{y(t)-y(t_0)}{t-t_0}\right) + \lim_{t\to t_0}\frac{E(x(t),y(t))}{t-t_0}.$$

The left-hand side of this equation is equal to dz/dt, which leads to

$$\frac{dz}{dt} = f_x(x_0,y_0)\frac{dx}{dt} + f_y(x_0,y_0)\frac{dy}{dt} + \lim_{t\to t_0}\frac{E(x(t),y(t))}{t-t_0}.$$

The last term can be rewritten as

$$\lim_{t\to t_0}\frac{E(x(t),y(t))}{t-t_0} = \lim_{t\to t_0}\frac{E(x,y)}{\sqrt{(x-x_0)^2+(y-y_0)^2}}\frac{\sqrt{(x-x_0)^2+(y-y_0)^2}}{t-t_0})$$
$$= \lim_{t\to t_0}\left(\frac{E(x,y)}{\sqrt{(x-x_0)^2+(y-y_0)^2}}\right)\lim_{t\to t_0}\left(\frac{\sqrt{(x-x_0)^2+(y-y_0)^2}}{t-t_0}\right).$$

As t approaches t_0, $\big(x(t),y(t)\big)$ approaches $\big(x(t_0),y(t_0)\big)$, so we can rewrite the last product as

$$\lim_{(x,y)\to(x_0,y_0)}\frac{E(x,y)}{\sqrt{(x-x_0)^2+(y-y_0)^2}}\lim_{(x,y)\to(x_0,y_0)}\left(\frac{\sqrt{(x-x_0)^2+(y-y_0)^2}}{t-t_0}\right).$$

Since the first limit is equal to zero, we need only show that the second limit is finite:

$$\lim_{(x,y)\to(x_0,y_0)}\frac{\sqrt{(x-x_0)^2+(y-y_0)^2}}{t-t_0} = \lim_{(x,y)\to(x_0,y_0)}\sqrt{\frac{(x-x_0)^2+(y-y_0)^2}{(t-t_0)^2}}$$
$$= \lim_{(x,y)\to(x_0,y_0)}\sqrt{\left(\frac{x-x_0}{t-t_0}\right)^2+\left(\frac{y-y_0}{t-t_0}\right)^2}$$
$$= \sqrt{\left[\lim_{(x,y)\to(x_0,y_0)}\left(\frac{x-x_0}{t-t_0}\right)\right]^2+\left[\lim_{(x,y)\to(x_0,y_0)}\left(\frac{y-y_0}{t-t_0}\right)\right]^2}.$$

Since $x(t)$ and $y(t)$ are both differentiable functions of t, both limits inside the last radical exist. Therefore, this value is finite. This proves the chain rule at $t=t_0$; the rest of the theorem follows from the assumption that all functions are differentiable over their entire domains.

□

Closer examination of Equation 13.4.1 reveals an interesting pattern. The first term in the equation is $\frac{\partial f}{\partial x}\cdot\frac{dx}{dt}$ and the second term is $\frac{\partial f}{\partial y}\cdot\frac{dy}{dt}$. Recall that when multiplying fractions, cancelation can be used. If we treat these derivatives as fractions, then each product "simplifies" to something resembling $\partial f/dt$. The variables x and y that disappear in this simplification are often called **intermediate variables**: they are independent variables for the function f, but are dependent variables for the variable t. Two terms appear on the right-hand side of the formula, and f is a function of two variables. This pattern works with functions of more than two variables as well, as we see later in this section.

Example 13.4.1: Using the Chain Rule

Calculate dz/dt for each of the following functions:

a. $z=f(x,y)=4x^2+3y^2, \quad x=x(t)=\sin t, \quad y=y(t)=\cos t$
b. $z=f(x,y)=\sqrt{x^2-y^2}, \quad x=x(t)=e^{2t}, \quad y=y(t)=e^{-t}$

Solution

a. To use the chain rule, we need four quantities—$\partial z/\partial x$, $\partial z/\partial y$, dx/dt, and dy/dt:

- $\dfrac{\partial z}{\partial x} = 8x$
- $\dfrac{dx}{dt} = \cos t$
- $\dfrac{\partial z}{\partial y} = 6y$
- $\dfrac{dy}{dt} = -\sin t$

Now, we substitute each of these into Equation 13.4.1:

$$\begin{aligned}\frac{dz}{dt} &= \frac{\partial z}{\partial x}\cdot\frac{dx}{dt} + \frac{\partial z}{\partial y}\cdot\frac{dy}{dt} \\ &= (8x)(\cos t) + (6y)(-\sin t) \\ &= 8x\cos t - 6y\sin t.\end{aligned}$$

This answer has three variables in it. To reduce it to one variable, use the fact that $x(t) = \sin t$ and $y(t) = \cos t$. We obtain

$$\begin{aligned}\frac{dz}{dt} &= 8x\cos t - 6y\sin t \\ &= 8(\sin t)\cos t - 6(\cos t)\sin t \\ &= 2\sin t\cos t.\end{aligned}$$

This derivative can also be calculated by first substituting $x(t)$ and $y(t)$ into $f(x,y)$, then differentiating with respect to t:

$$\begin{aligned}z = f(x,y) &= f\big(x(t), y(t)\big) \\ &= 4(x(t))^2 + 3(y(t))^2 \\ &= 4\sin^2 t + 3\cos^2 t.\end{aligned}$$

Then

$$\begin{aligned}\frac{dz}{dt} &= 2(4\sin t)(\cos t) + 2(3\cos t)(-\sin t) \\ &= 8\sin t\cos t - 6\sin t\cos t \\ &= 2\sin t\cos t,\end{aligned}$$

which is the same solution. However, it may not always be this easy to differentiate in this form.

b. To use the chain rule, we again need four quantities—$\partial z/\partial x, \partial z/dy, dx/dt$, and dy/dt :

- $\dfrac{\partial z}{\partial x} = \dfrac{x}{\sqrt{x^2 - y^2}}$
- $\dfrac{dx}{dt} = 2e^{2t}$
- $\dfrac{\partial z}{\partial y} = \dfrac{-y}{\sqrt{x^2 - y^2}}$
- $\dfrac{dy}{dt} = -e^{-t}$.

We substitute each of these into Equation 13.4.1:

$$\begin{aligned}\frac{dz}{dt} &= \frac{\partial z}{\partial x}\cdot\frac{dx}{dt}+\frac{\partial z}{\partial y}\cdot\frac{dy}{dt}\\ &= \left(\frac{x}{\sqrt{x^2-y^2}}\right)(2e^{2t})+\left(\frac{-y}{\sqrt{x^2-y^2}}\right)(-e^{-t})\\ &= \frac{2xe^{2t}-ye^{-t}}{\sqrt{x^2-y^2}}.\end{aligned}$$

To reduce this to one variable, we use the fact that $x(t)=e^{2t}$ and $y(t)=e^{-t}$. Therefore,

$$\begin{aligned}\frac{dz}{dt} &= \frac{2xe^2t+ye^{-t}}{\sqrt{x^2-y^2}}\\ &= \frac{2(e^{2t})e^{2t}+(e^{-t})e^{-t}}{\sqrt{e^{4t}-e^{-2t}}}\\ &= \frac{2e^{4t}+e^{-2t}}{\sqrt{e^{4t}-e^{-2t}}}.\end{aligned}$$

To eliminate negative exponents, we multiply the top by e^{2t} and the bottom by $\sqrt{e^{4t}}$:

$$\begin{aligned}\frac{dz}{dt} &= \frac{2e^{4t}+e^{-2t}}{\sqrt{e^{4t}-e^{-2t}}}\cdot\frac{e^{2t}}{\sqrt{e^{4t}}}\\ &= \frac{2e^{6t}+1}{\sqrt{e^{8t}-e^{2t}}}\\ &= \frac{2e^{6t}+1}{\sqrt{e^{2t}(e^{6t}-1)}}\\ &= \frac{2e^{6t}+1}{e^t\sqrt{e^{6t}-1}}.\end{aligned}$$

Again, this derivative can also be calculated by first substituting $x(t)$ and $y(t)$ into $f(x,y)$, then differentiating with respect to t:

$$\begin{aligned}z &= f(x,y)\\ &= f(x(t),y(t))\\ &= \sqrt{(x(t))^2-(y(t))^2}\\ &= \sqrt{e^{4t}-e^{-2t}}\\ &= (e^{4t}-e^{-2t})^{1/2}.\end{aligned}$$

Then

$$\begin{aligned}\frac{dz}{dt} &= \frac{1}{2}(e^{4t}-e^{-2t})^{-1/2}\left(4e^{4t}+2e^{-2t}\right)\\ &= \frac{2e^{4t}+e^{-2t}}{\sqrt{e^{4t}-e^{-2t}}}.\end{aligned}$$

This is the same solution.

Exercise 13.4.1

Calculate dz/dt given the following functions. Express the final answer in terms of t.

$$\begin{aligned} z = f(x,y) &= x^2 - 3xy + 2y^2, \\ x = x(t) &= 3\sin 2t, \\ y = y(t) &= 4\cos 2t \end{aligned}$$

Hint

Calculate $\partial z/\partial x, \partial z/dy, dx/dt$,and dy/dt, then use Equation 13.4.1.

Answer

$$\begin{aligned} \frac{dz}{dt} &= \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} \\ &= (2x-3y)(6\cos 2t) + (-3x+4y)(-8\sin 2t) \\ &= -92\sin 2t\cos 2t - 72(\cos^2 2t - \sin^2 2t) \\ &= -46\sin 4t - 72\cos 4t. \end{aligned}$$

It is often useful to create a visual representation of Equation 13.4.1 for the chain rule. This is called a **tree diagram** for the chain rule for functions of one variable and it provides a way to remember the formula (Figure 13.4.1). This diagram can be expanded for functions of more than one variable, as we shall see very shortly.

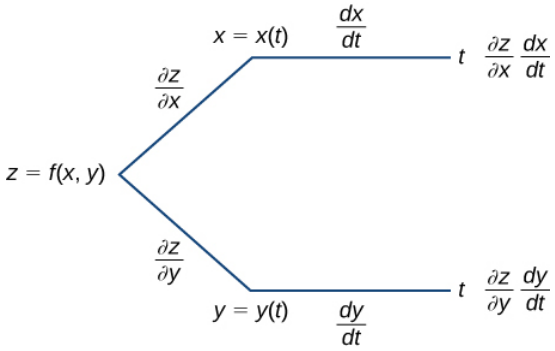


Figure 13.4.1: Tree diagram for the case $\dfrac{dz}{dt} = \dfrac{\partial z}{\partial x}\cdot\dfrac{dx}{dt} + \dfrac{\partial z}{\partial y}\cdot\dfrac{dy}{dt}$.

In this diagram, the leftmost corner corresponds to $z = f(x,y)$. Since f has two **independent variables**, there are two lines coming from this corner. The upper branch corresponds to the variable x and the lower branch corresponds to the variable y. Since each of these variables is then dependent on one variable t, one branch then comes from x and one branch comes from y. Last, each of the branches on the far right has a label that represents the path traveled to reach that branch. The top branch is reached by following the x branch, then the t branch; therefore, it is labeled $(\partial z/\partial x)\times(dx/dt)$. The bottom branch is similar: first the y branch, then the t branch. This branch is labeled $(\partial z/\partial y)\times(dy/dt)$. To get the formula for dz/dt, add all the terms that appear on the rightmost side of the diagram. This gives us Equation.

In Chain Rule for Two Independent Variables, $z = f(x,y)$ is a function of x and y, and both $x = g(u,v)$ and $y = h(u,v)$ are functions of the independent variables u and v.

Theorem 13.4.2: Chain Rule for Two Independent Variables

Suppose $x = g(u,v)$ and $y = h(u,v)$ are differentiable functions of u and v, and $z = f(x,y)$ is a differentiable function of x and y. Then, $z = f(g(u,v),h(u,v))$ is a differentiable function of u and v, and

$$\frac{\partial z}{\partial u} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial u} \tag{13.4.2}$$

and

$$\frac{\partial z}{\partial v} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial v}. \tag{13.4.3}$$

We can draw a tree diagram for each of these formulas as well as follows.

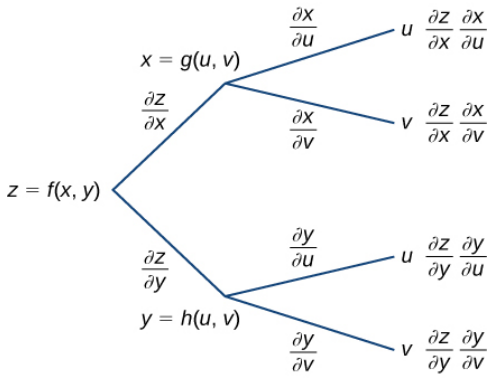


Figure 13.4.2: Tree diagram for $\frac{\partial z}{\partial u} = \frac{\partial z}{\partial x}\cdot\frac{\partial x}{\partial u} + \frac{\partial z}{\partial y}\cdot\frac{\partial y}{\partial u}$ and $\frac{\partial z}{\partial v} = \frac{\partial z}{\partial x}\cdot\frac{\partial x}{\partial v} + \frac{\partial z}{\partial y}\cdot\frac{\partial y}{\partial v}$.

To derive the formula for $\partial z/\partial u$, start from the left side of the diagram, then follow only the branches that end with u and add the terms that appear at the end of those branches. For the formula for $\partial z/\partial v$, follow only the branches that end with v and add the terms that appear at the end of those branches.

There is an important difference between these two chain rule theorems. In Chain Rule for One Independent Variable, the left-hand side of the formula for the derivative is not a partial derivative, but in Chain Rule for Two Independent Variables it is. The reason is that, in Chain Rule for One Independent Variable, z is ultimately a function of t alone, whereas in Chain Rule for Two Independent Variables, z is a function of both u and v.

Example 13.4.2: Using the Chain Rule for Two Variables

Calculate $\partial z/\partial u$ and $\partial z/\partial v$ using the following functions:

$$z = f(x,y) = 3x^2 - 2xy + y^2,\ x = x(u,v) = 3u + 2v,\ y = y(u,v) = 4u - v.$$

Solution

To implement the chain rule for two variables, we need six partial derivatives—$\partial z/\partial x,\ \partial z/\partial y,\ \partial x/\partial u,\ \partial x/\partial v,\ \partial y/\partial u$,and $\partial y/\partial v$:

$$\frac{\partial z}{\partial x} = 6x - 2y \qquad \frac{\partial z}{\partial y} = -2x + 2y$$
$$\frac{\partial x}{\partial u} = 3 \qquad \frac{\partial x}{\partial v} = 2$$
$$\frac{\partial y}{\partial u} = 4 \qquad \frac{\partial y}{\partial v} = -1.$$

To find $\partial z/\partial u$, we use Equation 13.4.2:

$$\begin{aligned}\frac{\partial z}{\partial u} &= \frac{\partial z}{\partial x}\cdot\frac{\partial x}{\partial u} + \frac{\partial z}{\partial y}\cdot\frac{\partial y}{\partial u}\\ &= 3(6x - 2y) + 4(-2x + 2y)\\ &= 10x + 2y.\end{aligned}$$

Next, we substitute $x(u,v) = 3u + 2v$ and $y(u,v) = 4u - v$:

$$\begin{aligned}\frac{\partial z}{\partial u} &= 10x + 2y \\ &= 10(3u + 2v) + 2(4u - v) \\ &= 38u + 18v.\end{aligned}$$

To find $\partial z/\partial v$, we use Equation 13.4.3:

$$\begin{aligned}\frac{\partial z}{\partial v} &= \frac{\partial z}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial v} \\ &= 2(6x - 2y) + (-1)(-2x + 2y) \\ &= 14x - 6y.\end{aligned}$$

Then we substitute $x(u, v) = 3u + 2v$ and $y(u, v) = 4u - v$:

$$\begin{aligned}\frac{\partial z}{\partial v} &= 14x - 6y \\ &= 14(3u + 2v) - 6(4u - v) \\ &= 18u + 34v\end{aligned}$$

? Exercise 13.4.2

Calculate $\partial z/\partial u$ and $\partial z/\partial v$ given the following functions:

$$z = f(x, y) = \frac{2x - y}{x + 3y}, \quad x(u, v) = e^{2u}\cos 3v, \quad y(u, v) = e^{2u}\sin 3v.$$

Hint

Calculate $\partial z/\partial x$, $\partial z/\partial y$, $\partial x/\partial u$, $\partial x/\partial v$, $\partial y/\partial u$, and $\partial y/\partial v$, then use Equation 13.4.2 and Equation 13.4.3

Answer

$$\frac{\partial z}{\partial u} = 0, \quad \frac{\partial z}{\partial v} = \frac{-21}{(3\sin 3v + \cos 3v)^2}$$

The Generalized Chain Rule

Now that we've see how to extend the original chain rule to functions of two variables, it is natural to ask: Can we extend the rule to more than two variables? The answer is yes, as the *generalized chain rule* states.

Theorem 13.4.3: Generalized Chain Rule

Let $w = f(x_1, x_2, \ldots, x_m)$ be a differentiable function of m independent variables, and for each $i \in 1, \ldots, m$, let $x_i = x_i(t_1, t_2, \ldots, t_n)$ be a differentiable function of n independent variables. Then

$$\frac{\partial w}{\partial t_j} = \frac{\partial w}{\partial x_1}\frac{\partial x_1}{\partial t_j} + \frac{\partial w}{\partial x_2}\frac{\partial x_2}{\partial t_j} + \cdots + \frac{\partial w}{\partial x_m}\frac{\partial x_m}{\partial t_j}$$

for any $j \in 1, 2, \ldots, n$.

In the next example we calculate the derivative of a function of three independent variables in which each of the three variables is dependent on two other variables.

Example 13.4.3: Using the Generalized Chain Rule

Calculate $\partial w/\partial u$ and $\partial w/\partial v$ using the following functions:

$$\begin{aligned} w &= f(x,y,z) = 3x^2 - 2xy + 4z^2 \\ x &= x(u,v) = e^u \sin v \\ y &= y(u,v) = e^u \cos v \\ z &= z(u,v) = e^u. \end{aligned}$$

Solution

The formulas for $\partial w/\partial u$ and $\partial w/\partial v$ are

$$\begin{aligned} \frac{\partial w}{\partial u} &= \frac{\partial w}{\partial x}\cdot\frac{\partial x}{\partial u} + \frac{\partial w}{\partial y}\cdot\frac{\partial y}{\partial u} + \frac{\partial w}{\partial z}\cdot\frac{\partial z}{\partial u} \\ \frac{\partial w}{\partial v} &= \frac{\partial w}{\partial x}\cdot\frac{\partial x}{\partial v} + \frac{\partial w}{\partial y}\cdot\frac{\partial y}{\partial v} + \frac{\partial w}{\partial z}\cdot\frac{\partial z}{\partial v}. \end{aligned}$$

Therefore, there are nine different partial derivatives that need to be calculated and substituted. We need to calculate each of them:

$$\begin{aligned} \frac{\partial w}{\partial x} &= 6x - 2y \quad \frac{\partial w}{\partial y} = -2x \quad \frac{\partial w}{\partial z} = 8z \\ \frac{\partial x}{\partial u} &= e^u \sin v \quad \frac{\partial y}{\partial u} = e^u \cos v \quad \frac{\partial z}{\partial u} = e^u \\ \frac{\partial x}{\partial v} &= e^u \cos v \quad \frac{\partial y}{\partial v} = -e^u \sin v \quad \frac{\partial z}{\partial v} = 0. \end{aligned}$$

Now, we substitute each of them into the first formula to calculate $\partial w/\partial u$:

$$\begin{aligned} \frac{\partial w}{\partial u} &= \frac{\partial w}{\partial x}\cdot\frac{\partial x}{\partial u} + \frac{\partial w}{\partial y}\cdot\frac{\partial y}{\partial u} + \frac{\partial w}{\partial z}\cdot\frac{\partial z}{\partial u} \\ &= (6x - 2y)e^u \sin v - 2xe^u \cos v + 8ze^u, \end{aligned}$$

then substitute $x(u,v) = e^u \sin v,\ y(u,v) = e^u \cos v,$ and $z(u,v) = e^u$ into this equation:

$$\begin{aligned} \frac{\partial w}{\partial u} &= (6x - 2y)e^u \sin v - 2xe^u \cos v + 8ze^u \\ &= (6e^u \sin v - 2eu \cos v)e^u \sin v - 2(e^u \sin v)e^u \cos v + 8e^{2u} \\ &= 6e^{2u}\sin^2 v - 4e^{2u}\sin v \cos v + 8e^{2u} \\ &= 2e^{2u}(3\sin^2 v - 2\sin v \cos v + 4). \end{aligned}$$

Next, we calculate $\partial w/\partial v$:

$$\begin{aligned} \frac{\partial w}{\partial v} &= \frac{\partial w}{\partial x}\cdot\frac{\partial x}{\partial v} + \frac{\partial w}{\partial y}\cdot\frac{\partial y}{\partial v} + \frac{\partial w}{\partial z}\cdot\frac{\partial z}{\partial v} \\ &= (6x - 2y)e^u \cos v - 2x(-e^u \sin v) + 8z(0), \end{aligned}$$

then we substitute $x(u,v) = e^u \sin v,\ y(u,v) = e^u \cos v,$ and $z(u,v) = e^u$ into this equation:

$$\begin{aligned} \frac{\partial w}{\partial v} &= (6x - 2y)e^u \cos v - 2x(-e^u \sin v) \\ &= (6e^u \sin v - 2e^u \cos v)e^u \cos v + 2(e^u \sin v)(e^u \sin v) \\ &= 2e^{2u}\sin^2 v + 6e^{2u}\sin v \cos v - 2e^{2u}\cos^2 v \\ &= 2e^{2u}(\sin^2 v + \sin v \cos v - \cos^2 v). \end{aligned}$$

Exercise 13.4.3

Calculate $\partial w/\partial u$ and $\partial w/\partial v$ given the following functions:

$$w = f(x,y,z) = \frac{x+2y-4z}{2x-y+3z}$$
$$x = x(u,v) = e^{2u}\cos 3v$$
$$y = y(u,v) = e^{2u}\sin 3v$$
$$z = z(u,v) = e^{2u}.$$

Hint

Calculate nine partial derivatives, then use the same formulas from Example 13.4.3

Answer

$$\frac{\partial w}{\partial u} = 0$$
$$\frac{\partial w}{\partial v} = \frac{15-33\sin 3v+6\cos 3v}{(3+2\cos 3v-\sin 3v)^2}$$

Example 13.4.4: Drawing a Tree Diagram

Create a tree diagram for the case when

$$w = f(x,y,z), \quad x = x(t,u,v), \quad y = y(t,u,v), \quad z = z(t,u,v)$$

and write out the formulas for the three partial derivatives of w.

Solution

Starting from the left, the function f has three independent variables: x, y, and z. Therefore, three branches must be emanating from the first node. Each of these three branches also has three branches, for each of the variables t, u, and v.

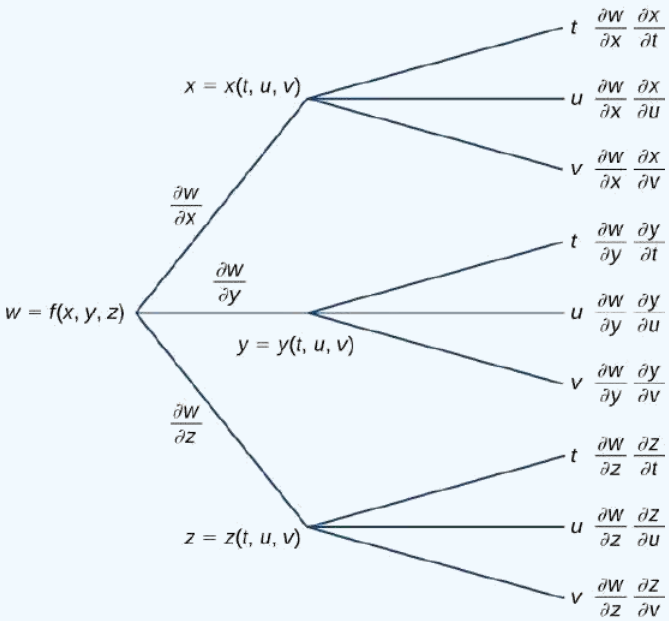


Figure 13.4.3: Tree diagram for a function of three variables, each of which is a function of three independent variables.

The three formulas are

$$\frac{\partial w}{\partial t} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial w}{\partial z}\frac{\partial z}{\partial t}$$
$$\frac{\partial w}{\partial u} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial u} + \frac{\partial w}{\partial z}\frac{\partial z}{\partial u}$$
$$\frac{\partial w}{\partial v} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial v} + \frac{\partial w}{\partial z}\frac{\partial z}{\partial v}.$$

? Exercise 13.4.4

Create a tree diagram for the case when

$$w = f(x, y), \quad x = x(t, u, v), \quad y = y(t, u, v)$$

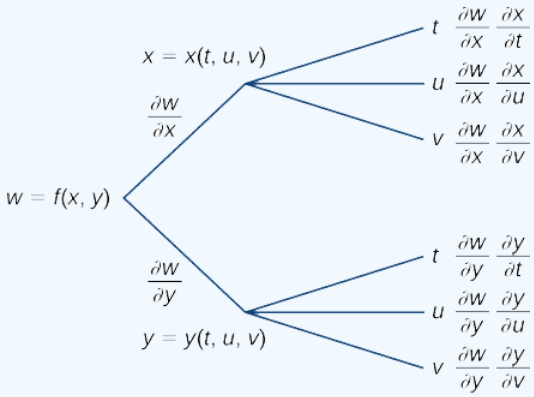
and write out the formulas for the three partial derivatives of w.

Hint

Determine the number of branches that emanate from each node in the tree.

Answer

$$\frac{\partial w}{\partial t} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial t}$$
$$\frac{\partial w}{\partial u} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial u}$$
$$\frac{\partial w}{\partial v} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial v}$$



Implicit Differentiation

Recall from implicit differentiation provides a method for finding dy/dx when y is defined implicitly as a function of x. The method involves differentiating both sides of the equation defining the function with respect to x, then solving for dy/dx. Partial derivatives provide an alternative to this method.

Consider the ellipse defined by the equation $x^2 + 3y^2 + 4y - 4 = 0$ as follows.

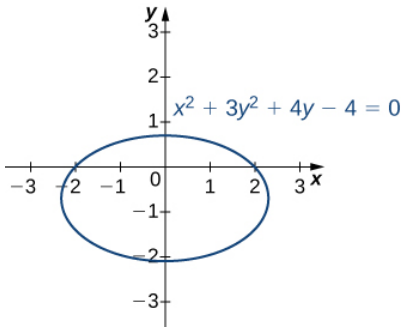


Figure 13.4.4: Graph of the ellipse defined by $x^2+3y^2+4y-4=0$.

This equation implicitly defines y as a function of x. As such, we can find the derivative dy/dx using the method of implicit differentiation:

$$\begin{aligned}\frac{d}{dx}(x^2+3y^2+4y-4) &= \frac{d}{dx}(0)\\ 2x+6y\frac{dy}{dx}+4\frac{dy}{dx} &= 0\\ (6y+4)\frac{dy}{dx} &= -2x\\ \frac{dy}{dx} &= -\frac{x}{3y+2}\end{aligned}$$

We can also define a function $z=f(x,y)$ by using the left-hand side of the equation defining the ellipse. Then $f(x,y)=x^2+3y^2+4y-4$. The ellipse $x^2+3y^2+4y-4=0$ can then be described by the equation $f(x,y)=0$. Using this function and the following theorem gives us an alternative approach to calculating dy/dx.

Theorem 13.4.4: Implicit Differentiation of a Function of Two or More Variables

Suppose the function $z=f(x,y)$ defines y implicitly as a function $y=g(x)$ of x via the equation $f(x,y)=0$. Then

$$\frac{dy}{dx}=-\frac{\partial f/\partial x}{\partial f/\partial y} \tag{13.4.4}$$

provided $f_y(x,y)\neq 0$.

If the equation $f(x,y,z)=0$ defines z implicitly as a differentiable function of x and y, then

$$\frac{dz}{dx}=-\frac{\partial f/\partial x}{\partial f/\partial z} \quad \text{and} \quad \frac{dz}{dy}=-\frac{\partial f/\partial y}{\partial f/\partial z} \tag{13.4.5}$$

as long as $f_z(x,y,z)\neq 0$.

Equation 13.4.4 is a direct consequence of Equation 13.4.2. In particular, if we assume that y is defined implicitly as a function of x via the equation $f(x,y)=0$, we can apply the chain rule to find dy/dx :

$$\begin{aligned}\frac{d}{dx}f(x,y) &= \frac{d}{dx}(0)\\ \frac{\partial f}{\partial x}\cdot\frac{dx}{dx}+\frac{\partial f}{\partial y}\cdot\frac{dy}{dx} &= 0\\ \frac{\partial f}{\partial x}+\frac{\partial f}{\partial y}\cdot\frac{dy}{dx} &= 0.\end{aligned}$$

Solving this equation for dy/dx gives Equation 13.4.4. Equation 13.4.4 can be derived in a similar fashion.

Let's now return to the problem that we started before the previous theorem. Using Theorem 13.4.4 and the function $f(x,y)=x^2+3y^2+4y-4$, we obtain

$$\frac{\partial f}{\partial x} = 2x$$
$$\frac{\partial f}{\partial y} = 6y + 4.$$

Then Equation 13.4.4 gives

$$\frac{dy}{dx} = -\frac{\partial f/\partial x}{\partial f/\partial y} = -\frac{2x}{6y+4} = -\frac{x}{3y+2},$$

which is the same result obtained by the earlier use of implicit differentiation.

Example 13.4.5: Implicit Differentiation by Partial Derivatives

a. Calculate dy/dx if y is defined implicitly as a function of x via the equation $3x^2 - 2xy + y^2 + 4x - 6y - 11 = 0$. What is the equation of the tangent line to the graph of this curve at point $(2, 1)$?

b. Calculate $\partial z/\partial x$ and $\partial z/\partial y$, given $x^2 e^y - yze^x = 0$.

Solution

a. Set $f(x, y) = 3x^2 - 2xy + y^2 + 4x - 6y - 11 = 0$, then calculate f_x and $f_y : f_x(x, y) = 6x - 2y + 4$ and $f_y(x, y) = -2x + 2y - 6$.

The derivative is given by

$$\frac{dy}{dx} = -\frac{\partial f/\partial x}{\partial f/\partial y} = \frac{6x - 2y + 4}{-2x + 2y - 6} = \frac{3x - y + 2}{x - y + 3}.$$

The slope of the tangent line at point $(2, 1)$ is given by

$$\left.\frac{dy}{dx}\right|_{(x,y)=(2,1)} = \frac{3(2) - 1 + 2}{2 - 1 + 3} = \frac{7}{4}$$

To find the equation of the tangent line, we use the point-slope form (Figure 13.4.5):

$$\begin{aligned} y - y_0 &= m(x - x_0) \\ y - 1 &= \frac{7}{4}(x - 2) \\ y &= \frac{7}{4}x - \frac{7}{2} + 1 \\ y &= \frac{7}{4}x - \frac{5}{2}. \end{aligned}$$

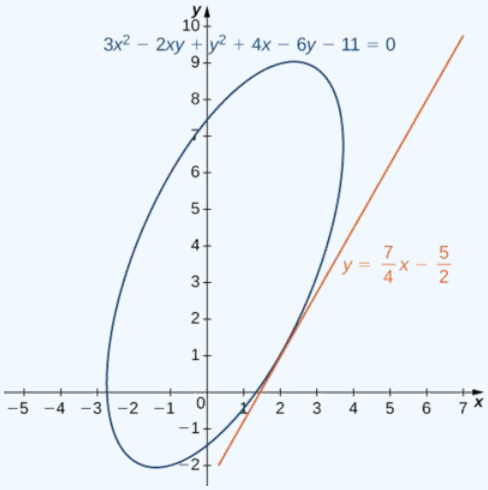


Figure 13.4.5: Graph of the rotated ellipse defined by $3x^2 - 2xy + y^2 + 4x - 6y - 11 = 0$.

b. We have $f(x,y,z)=x^2e^y-yze^x$. Therefore,

$$\frac{\partial f}{\partial x}=2xe^y-yze^x$$
$$\frac{\partial f}{\partial y}=x^2e^y-ze^x$$
$$\frac{\partial f}{\partial z}=-ye^x$$

Using Equation 13.4.5,

$$\frac{\partial z}{\partial x}=-\frac{\partial f/\partial x}{\partial f/\partial y}=-\frac{2xe^y-yze^x}{-ye^x}=\frac{2xe^y-yze^x}{ye^x} \quad \text{and} \quad \frac{\partial z}{\partial y}=-\frac{\partial f/\partial y}{\partial f/\partial z}=-\frac{x^2e^y-ze^x}{-ye^x}=\frac{x^2e^y-ze^x}{ye^x}$$

? Exercise 13.4.5

Find dy/dx if y is defined implicitly as a function of x by the equation $x^2+xy-y^2+7x-3y-26=0$. What is the equation of the tangent line to the graph of this curve at point $(3,-2)$?

Hint

Calculate $\partial f/dx$ and $\partial f/dy$, then use Equation 13.4.4

Solution

$$\frac{dy}{dx}=\frac{2x+y+7}{2y-x+3}\Big|_{(3,-2)}=\frac{2(3)+(-2)+7}{2(-2)-(3)+3}=-\frac{11}{4}$$

Equation of the tangent line: $y=-\frac{11}{4}x+\frac{25}{4}$

Key Concepts

- The chain rule for functions of more than one variable involves the partial derivatives with respect to all the independent variables.
- Tree diagrams are useful for deriving formulas for the chain rule for functions of more than one variable, where each independent variable also depends on other variables.

Key Equations

- **Chain rule, one independent variable**

$$\frac{dz}{dt}=\frac{\partial z}{\partial x}\cdot\frac{dx}{dt}+\frac{\partial z}{\partial y}\cdot\frac{dy}{dt}$$

- **Chain rule, two independent variables**

$$\frac{dz}{du}=\frac{\partial z}{\partial x}\cdot\frac{\partial x}{\partial u}+\frac{\partial z}{\partial y}\cdot\frac{\partial y}{\partial u}\quad\frac{dz}{dv}=\frac{\partial z}{\partial x}\cdot\frac{\partial x}{\partial v}+\frac{\partial z}{\partial y}\cdot\frac{\partial y}{\partial v}$$

- **Generalized chain rule**

$$\frac{\partial w}{\partial t_j}=\frac{\partial w}{\partial x_1}\frac{\partial x_1}{\partial t_j}+\frac{\partial w}{\partial x_2}\frac{\partial x_1}{\partial t_j}+\cdots+\frac{\partial w}{\partial x_m}\frac{\partial x_m}{\partial t_j}$$

Glossary

generalized chain rule

the chain rule extended to functions of more than one independent variable, in which each independent variable may depend on one or more other variables

intermediate variable

given a composition of functions (e.g., $f(x(t), y(t))$), the intermediate variables are the variables that are independent in the outer function but dependent on other variables as well; in the function $f(x(t), y(t))$, the variables x and y are examples of intermediate variables

tree diagram

illustrates and derives formulas for the generalized chain rule, in which each independent variable is accounted for

---

13.4: The Chain Rule is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **14.5: The Chain Rule for Multivariable Functions** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

13.5: Directional Derivatives and Gradient Vectors

Learning Objectives

- Determine the directional derivative in a given direction for a function of two variables.
- Determine the gradient vector of a given real-valued function.
- Explain the significance of the gradient vector with regard to direction of change along a surface.
- Use the gradient to find the tangent to a level curve of a given function.
- Calculate directional derivatives and gradients in three dimensions.

A function $z = f(x, y)$ has two partial derivatives: $\partial z/\partial x$ and $\partial z/\partial y$. These derivatives correspond to each of the independent variables and can be interpreted as instantaneous rates of change (that is, as slopes of a tangent line). For example, $\partial z/\partial x$ represents the slope of a tangent line passing through a given point on the surface defined by $z = f(x, y)$, assuming the tangent line is parallel to the x**-axis**. Similarly, $\partial z/\partial y$ represents the slope of the tangent line parallel to the y**-axis.** Now we consider the possibility of a tangent line parallel to neither axis.

Directional Derivatives

We start with the graph of a surface defined by the equation $z = f(x, y)$. Given a point (a, b) in the domain of f, we choose a direction to travel from that point. We measure the direction using an angle θ, which is measured counterclockwise in the xy-plane, starting at zero from the positive x-axis (Figure 13.5.1). The distance we travel is h and the direction we travel is given by the unit vector $\vec{\mathbf{u}} = (\cos\theta)\,\hat{\mathbf{i}} + (\sin\theta)\,\hat{\mathbf{j}}$. Therefore, the z-coordinate of the second point on the graph is given by $z = f(a + h\cos\theta, b + h\sin\theta)$.

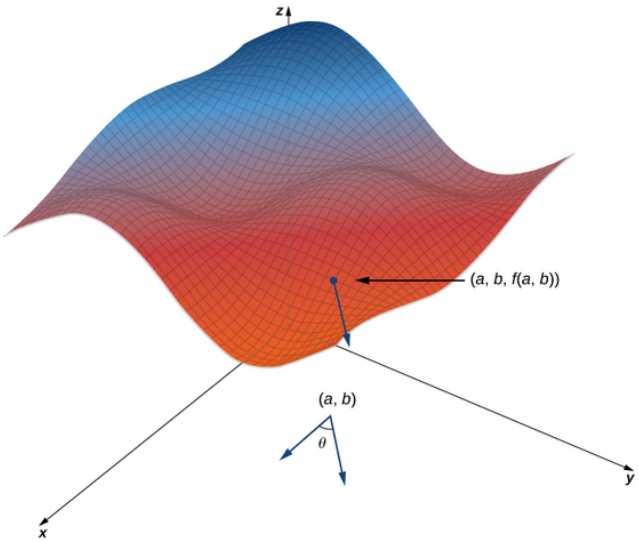


Figure 13.5.1: Finding the directional derivative at a point on the graph of $z = f(x, y)$. The slope of the blue arrow on the graph indicates the value of the directional derivative at that point.

We can calculate the slope of the secant line by dividing the difference in z-values by the length of the line segment connecting the two points in the domain. The length of the line segment is h. Therefore, the slope of the secant line is

$$m_{sec} = \frac{f(a + h\cos\theta, b + h\sin\theta) - f(a, b)}{h}$$

To find the slope of the tangent line in the same direction, we take the limit as h approaches zero.

Definition: Directional Derivatives

Suppose $z = f(x, y)$ is a function of two variables with a domain of D. Let $(a, b) \in D$ and define $\vec{\mathbf{u}} = (\cos\theta)\,\hat{\mathbf{i}} + (\sin\theta)\,\hat{\mathbf{j}}$. Then the directional derivative of f in the direction of $\vec{\mathbf{u}}$ is given by

$$D_{\vec{\mathbf{u}}} f(a, b) = \lim_{h \to 0} \frac{f(a + h\cos\theta, b + h\sin\theta) - f(a, b)}{h} \tag{13.5.1}$$

provided the limit exists.

Equation 13.5.1 provides a formal definition of the directional derivative that can be used in many cases to calculate a directional derivative.

Note that since the point (a,b) is chosen randomly from the domain D of the function f, we can use this definition to find the directional derivative as a function of x and y.

That is,

$$D_{\vec{\mathbf{u}}} f(x,y) = \lim_{h\to 0} \frac{f(x+h\cos\theta, y+h\sin\theta) - f(x,y)}{h} \tag{13.5.2}$$

Example 13.5.1: Finding a Directional Derivative from the Definition

Let $\theta = \arccos(3/5)$. Find the directional derivative $D_{\vec{\mathbf{u}}} f(x,y)$ of $f(x,y) = x^2 - xy + 3y^2$ in the direction of $\vec{\mathbf{u}} = (\cos\theta)\,\hat{\mathbf{i}} + (\sin\theta)\,\hat{\mathbf{j}}$.

Then determine $D_{\vec{\mathbf{u}}} f(-1,2)$.

Solution

First of all, since $\cos\theta = 3/5$ and θ is acute, this implies

$$\sin\theta = \sqrt{1 - \left(\frac{3}{5}\right)^2} = \sqrt{\frac{16}{25}} = \frac{4}{5}.$$

Using $f(x,y) = x^2 - xy + 3y^2$, we first calculate $f(x+h\cos\theta, y+h\sin\theta)$:

$$\begin{aligned} f(x+h\cos\theta, y+h\sin\theta) &= (x+h\cos\theta)^2 - (x+h\cos\theta)(y+h\sin\theta) + 3(y+h\sin\theta)^2 \\ &= x^2 + 2xh\cos\theta + h^2\cos^2\theta - xy - xh\sin\theta - yh\cos\theta - h^2\sin\theta\cos\theta + 3y^2 + 6yh\sin\theta + 3h^2\sin^2\theta \\ &= x^2 + 2xh(\frac{3}{5}) + \frac{9h^2}{25} - xy - \frac{4xh}{5} - \frac{3yh}{5} - \frac{12h^2}{25} + 3y^2 + 6yh(\frac{4}{5}) + 3h^2(\frac{16}{25}) \\ &= x^2 - xy + 3y^2 + \frac{2xh}{5} + \frac{9h^2}{5} + \frac{21yh}{5}. \end{aligned}$$

We substitute this expression into Equation 13.5.1 with $a = x$ and $b = y$:

$$\begin{aligned} D_{\vec{\mathbf{u}}} f(x,y) &= \lim_{h\to 0} \frac{f(x+h\cos\theta, y+h\sin\theta) - f(x,y)}{h} \\ &= \lim_{h\to 0} \frac{(x^2 - xy + 3y^2 + \frac{2xh}{5} + \frac{9h^2}{5} + \frac{21yh}{5}) - (x^2 - xy + 3y^2)}{h} \\ &= \lim_{h\to 0} \frac{\frac{2xh}{5} + \frac{9h^2}{5} + \frac{21yh}{5}}{h} \\ &= \lim_{h\to 0} \frac{2x}{5} + \frac{9h}{5} + \frac{21y}{5} \\ &= \frac{2x+21y}{5}. \end{aligned}$$

To calculate $D_{\vec{\mathbf{u}}} f(-1,2)$, we substitute $x = -1$ and $y = 2$ into this answer (Figure 13.5.2):

$$D_{\vec{\mathbf{u}}} f(-1,2) = \frac{2(-1)+21(2)}{5} = \frac{-2+42}{5} = 8.$$

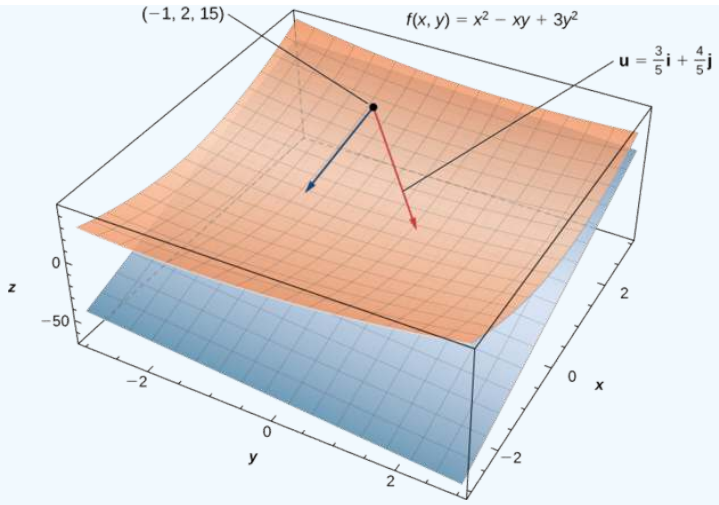


Figure 13.5.2: Finding the directional derivative in a given direction $\vec{\mathbf{u}}$ at a given point on a surface. The plane is tangent to the surface at the given point $(-1, 2, 15)$.

An easier approach to calculating directional derivatives that involves partial derivatives is outlined in the following theorem.

Theorem 13.5.1: Directional Derivative of a Function of Two Variables

Let $z = f(x, y)$ be a function of two variables x and y, and assume that f_x and f_y exist. Then the directional derivative of f in the direction of $\vec{\mathbf{u}} = (\cos\theta)\,\hat{\mathbf{i}} + (\sin\theta)\,\hat{\mathbf{j}}$ is given by

$$D_{\vec{\mathbf{u}}} f(x, y) = f_x(x, y)\cos\theta + f_y(x, y)\sin\theta. \tag{13.5.3}$$

Proof

Applying the definition of a directional derivative stated above in Equation 13.5.1, the directional derivative of f in the direction of $\vec{\mathbf{u}} = (\cos\theta)\,\hat{\mathbf{i}} + (\sin\theta)\,\hat{\mathbf{j}}$ at a point (x_0, y_0) in the domain of f can be written

$$D_{\vec{\mathbf{u}}} f((x_0, y_0)) = \lim_{t\to 0} \frac{f(x_0 + t\cos\theta, y_0 + t\sin\theta) - f(x_0, y_0)}{t}.$$

Let $x = x_0 + t\cos\theta$ and $y = y_0 + t\sin\theta$, and define $g(t) = f(x, y)$. Since f_x and f_y both exist, we can use the chain rule for functions of two variables to calculate $g'(t)$:

$$g'(t) = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} = f_x(x, y)\cos\theta + f_y(x, y)\sin\theta.$$

If $t = 0$, then $x = x_0$ and $y = y_0$, so

$$g'(0) = f_x(x_0, y_0)\cos\theta + f_y(x_0, y_0)\sin\theta$$

By the definition of $g'(t)$, it is also true that

$$g'(0) = \lim_{t\to 0} \frac{g(t) - g(0)}{t} = \lim_{t\to 0} \frac{f(x_0 + t\cos\theta, y_0 + t\sin\theta) - f(x_0, y_0)}{t}.$$

Therefore, $D_{\vec{\mathbf{u}}} f(x_0, y_0) = f_x(x_0, y_0)\cos\theta + f_y(x_0, y_0)\sin\theta$.

Since the point (x_0, y_0) is an arbitrary point from the domain of f, this result holds for all points in the domain of f for which the partials f_x and f_y exist.

Therefore,

$$D_{\vec{\mathbf{u}}} f(x, y) = f_x(x, y)\cos\theta + f_y(x, y)\sin\theta.$$

□

Example 13.5.2: Finding a Directional Derivative: Alternative Method

Let $\theta = \arccos(3/5)$. Find the directional derivative $D_{\vec{\mathbf{u}}} f(x,y)$ of $f(x,y) = x^2 - xy + 3y^2$ in the direction of $\vec{\mathbf{u}} = (\cos\theta)\,\hat{\mathbf{i}} + (\sin\theta)\,\hat{\mathbf{j}}$.

Then determine $D_{\vec{\mathbf{u}}} f(-1,2)$.

Solution

First, we must calculate the partial derivatives of f:

$$\begin{aligned} f_x(x,y) &= 2x - y \\ f_y(x,y) &= -x + 6y, \end{aligned}$$

Then we use Equation 13.5.3 with $\theta = \arccos(3/5)$:

$$\begin{aligned} D_{\vec{\mathbf{u}}} f(x,y) &= f_x(x,y)\cos\theta + f_y(x,y)\sin\theta \\ &= (2x-y)\frac{3}{5} + (-x+6y)\frac{4}{5} \\ &= \frac{6x}{5} - \frac{3y}{5} - \frac{4x}{5} + \frac{24y}{5} \\ &= \frac{2x+21y}{5}. \end{aligned}$$

To calculate $D_{\vec{\mathbf{u}}} f(-1,2)$, let $x = -1$ and $y = 2$:

$$D_{\vec{\mathbf{u}}} f(-1,2) = \frac{2(-1)+21(2)}{5} = \frac{-2+42}{5} = 8.$$

This is the same answer obtained in Example 13.5.1.

Exercise 13.5.1:

Find the directional derivative $D_{\vec{\mathbf{u}}} f(x,y)$ of $f(x,y) = 3x^2y - 4xy^3 + 3y^2 - 4x$ in the direction of $\vec{\mathbf{u}} = (\cos\frac{\pi}{3})\,\hat{\mathbf{i}} + (\sin\frac{\pi}{3})\,\hat{\mathbf{j}}$ using Equation 13.5.3.

What is $D_{\vec{\mathbf{u}}} f(3,4)$?

Hint

Calculate the partial derivatives and determine the value of θ.

Answer

$$D_{\vec{\mathbf{u}}} f(x,y) = \frac{(6xy - 4y^3 - 4)(1)}{2} + \frac{(3x^2 - 12xy^2 + 6y)\sqrt{3}}{2}$$

$$D_{\vec{\mathbf{u}}} f(3,4) = \frac{72 - 256 - 4}{2} + \frac{(27 - 576 + 24)\sqrt{3}}{2} = -94 - \frac{525\sqrt{3}}{2}$$

If the vector that is given for the direction of the derivative is not a unit vector, then it is only necessary to divide by the norm of the vector. For example, if we wished to find the directional derivative of the function in Example 13.5.2 in the direction of the vector $\langle -5, 12\rangle$, we would first divide by its magnitude to get $\vec{\mathbf{u}}$. This gives us $\vec{\mathbf{u}} = \langle -\frac{5}{13}, \frac{12}{13}\rangle$.

Then

$$\begin{aligned} D_{\vec{\mathbf{u}}} f(x,y) &= f_x(x,y)\cos\theta + f_y(x,y)\sin\theta \\ &= -\frac{5}{13}(2x-y) + \frac{12}{13}(-x+6y) \\ &= -\frac{22}{13}x + \frac{17}{13}y \end{aligned}$$

Gradient

The right-hand side of Equation 13.5.3 is equal to $f_x(x,y)\cos\theta + f_y(x,y)\sin\theta$, which can be written as the dot product of two vectors. Define the first vector as $\vec{\boldsymbol{\nabla}} f(x,y) = f_x(x,y)\,\hat{\mathbf{i}} + f_y(x,y)\,\hat{\mathbf{j}}$ and the second vector as $\vec{\mathbf{u}} = (\cos\theta)\,\hat{\mathbf{i}} + (\sin\theta)\,\hat{\mathbf{j}}$. Then the right-hand

side of the equation can be written as the dot product of these two vectors:

$$D_{\vec{\mathbf{u}}} f(x,y) = \vec{\nabla} f(x,y) \cdot \vec{\mathbf{u}}. \tag{13.5.4}$$

The first vector in Equation 13.5.4 has a special name: the gradient of the function f. The symbol ∇ is called **nabla** and the vector $\vec{\nabla} f$ is read "**del** f."

Definition: The Gradient

Let $z = f(x,y)$ be a function of x and y such that f_x and f_y exist. The vector $\vec{\nabla} f(x,y)$ is called the **gradient** of f and is defined as

$$\vec{\nabla} f(x,y) = f_x(x,y)\,\hat{\mathbf{i}} + f_y(x,y)\,\hat{\mathbf{j}}. \tag{13.5.5}$$

The vector $\vec{\nabla} f(x,y)$ is also written as "**grad** f."

Example 13.5.3: Finding Gradients

Find the gradient $\vec{\nabla} f(x,y)$ of each of the following functions:

a. $f(x,y) = x^2 - xy + 3y^2$
b. $f(x,y) = \sin 3x \cos 3y$

Solution

For both parts a. and b., we first calculate the partial derivatives f_x and f_y, then use Equation 13.5.5.

a. $f_x(x,y) = 2x - y$ and $f_y(x,y) = -x + 6y$, so

$$\begin{aligned} \vec{\nabla} f(x,y) &= f_x(x,y)\,\hat{\mathbf{i}} + f_y(x,y)\,\hat{\mathbf{j}} \\ &= (2x - y)\,\hat{\mathbf{i}} + (-x + 6y)\,\hat{\mathbf{j}}. \end{aligned}$$

b. $f_x(x,y) = 3\cos 3x \cos 3y$ and $f_y(x,y) = -3\sin 3x \sin 3y$, so

$$\begin{aligned} \vec{\nabla} f(x,y) &= f_x(x,y)\,\hat{\mathbf{i}} + f_y(x,y)\,\hat{\mathbf{j}} \\ &= (3\cos 3x \cos 3y)\,\hat{\mathbf{i}} - (3\sin 3x \sin 3y)\,\hat{\mathbf{j}}. \end{aligned}$$

Exercise 13.5.2

Find the gradient $\vec{\nabla} f(x,y)$ of $f(x,y) = \dfrac{x^2 - 3y^2}{2x + y}$.

Hint

Calculate the partial derivatives, then use Equation 13.5.5.

Answer

$$\vec{\nabla} f(x,y) = \frac{2x^2 + 2xy + 6y^2}{(2x+y)^2}\,\hat{\mathbf{i}} - \frac{x^2 + 12xy + 3y^2}{(2x+y)^2}\,\hat{\mathbf{j}}$$

The gradient has some important properties. We have already seen one formula that uses the gradient: the formula for the directional derivative. Recall from The Dot Product that if the angle between two vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ is φ, then $\vec{\mathbf{a}} \cdot \vec{\mathbf{b}} = \|\vec{\mathbf{a}}\|\|\vec{\mathbf{b}}\| \cos\varphi$. Therefore, if the angle between $\vec{\nabla} f(x_0, y_0)$ and $\vec{\mathbf{u}} = (cos\theta)\,\hat{\mathbf{i}} + (sin\theta)\,\hat{\mathbf{j}}$ is φ, we have

$$D_{\vec{\mathbf{u}}} f(x_0,y_0) = \vec{\nabla} f(x_0,y_0) \cdot \vec{\mathbf{u}} = \|\vec{\nabla} f(x_0,y_0)\|\|\vec{\mathbf{u}}\| \cos\varphi = \|\vec{\nabla} f(x_0,y_0)\| \cos\varphi.$$

The $\|\vec{\mathbf{u}}\|$ disappears because $\vec{\mathbf{u}}$ is a unit vector. Therefore, the directional derivative is equal to the magnitude of the gradient evaluated at (x_0, y_0) multiplied by $\cos\varphi$. Recall that $\cos\varphi$ ranges from -1 to 1.

If $\varphi = 0$, then $\cos\varphi = 1$ and $\vec{\nabla} f(x_0, y_0)$ and $\vec{\mathbf{u}}$ both point in the same direction.

If $\varphi = \pi$, then $\cos\varphi = -1$ and $\vec{\nabla} f(x_0, y_0)$ and $\vec{\mathbf{u}}$ point in opposite directions.

In the first case, the value of $D_{\vec{\mathbf{u}}} f(x_0, y_0)$ is maximized; in the second case, the value of $D_{\vec{\mathbf{u}}} f(x_0, y_0)$ is minimized.

We can also see that if $\vec{\nabla} f(x_0, y_0) = \vec{0}$, then

$$D_{\vec{u}} f(x_0, y_0) = \vec{\nabla} f(x_0, y_0) \cdot \vec{u} = 0$$

for any vector $\vec{u}$. These three cases are outlined in the following theorem.

Theorem 13.5.2: Properties of the Gradient

Suppose the function $z = f(x, y)$ is differentiable at (x_0, y_0) (Figure 13.5.3).

i. If $\vec{\nabla} f(x_0, y_0) = \vec{0}$, then $D_{\vec{u}} f(x_0, y_0) = 0$ for any unit vector $\vec{u}$.
ii. If $\vec{\nabla} f(x_0, y_0) \neq \vec{0}$, then $D_{\vec{u}} f(x_0, y_0)$ is **maximized** when $\vec{u}$ points in the same direction as $\vec{\nabla} f(x_0, y_0)$. The maximum value of $D_{\vec{u}} f(x_0, y_0)$ is $\|\vec{\nabla} f(x_0, y_0)\|$.
iii. If $\vec{\nabla} f(x_0, y_0) \neq \vec{0}$, then $D_{\vec{u}} f(x_0, y_0)$ is **minimized** when $\vec{u}$ points in the opposite direction from $\vec{\nabla} f(x_0, y_0)$. The minimum value of $D_{\vec{u}} f(x_0, y_0)$ is $-\|\vec{\nabla} f(x_0, y_0)\|$.

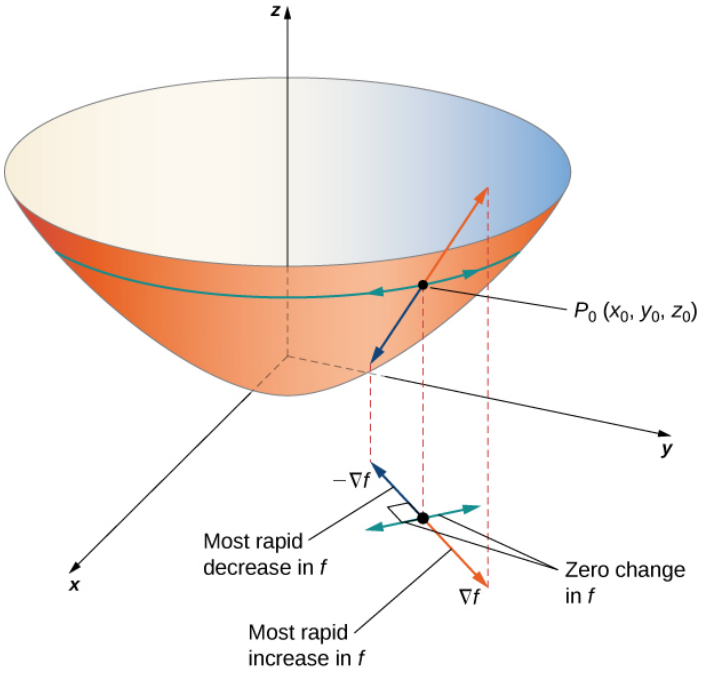


Figure 13.5.3: The gradient indicates the maximum and minimum values of the directional derivative at a point.

Example 13.5.4: Finding a Maximum Directional Derivative

Find the direction for which the directional derivative of $f(x, y) = 3x^2 - 4xy + 2y^2$ at $(-2, 3)$ is a maximum. What is the maximum value?

Solution

The maximum value of the directional derivative occurs when $\vec{\nabla} f$ and the unit vector point in the same direction. Therefore, we start by calculating $\vec{\nabla} f(x, y)$:

$$f_x(x, y) = 6x - 4y \text{ and } f_y(x, y) = -4x + 4y$$

so

$$\vec{\nabla} f(x, y) = f_x(x, y)\,\hat{\mathbf{i}} + f_y(x, y)\,\hat{\mathbf{j}} = (6x - 4y)\,\hat{\mathbf{i}} + (-4x + 4y)\,\hat{\mathbf{j}}.$$

Next, we evaluate the gradient at $(-2, 3)$:

$$\vec{\nabla} f(-2, 3) = (6(-2) - 4(3))\,\hat{\mathbf{i}} + (-4(-2) + 4(3))\,\hat{\mathbf{j}} = -24\,\hat{\mathbf{i}} + 20\,\hat{\mathbf{j}}.$$

We need to find a unit vector that points in the same direction as $\vec{\nabla} f(-2,3)$, so the next step is to divide $\vec{\nabla} f(-2,3)$ by its magnitude, which is $\sqrt{(-24)^2+(20)^2} = \sqrt{976} = 4\sqrt{61}$. Therefore,

$$\frac{\vec{\nabla} f(-2,3)}{\|\vec{\nabla} f(-2,3)\|} = \frac{-24}{4\sqrt{61}} i + \frac{20}{4\sqrt{61}} j = -\frac{6\sqrt{61}}{61}\,\hat{\mathbf{i}} + \frac{5\sqrt{61}}{61}\,\hat{\mathbf{j}}.$$

This is the unit vector that points in the same direction as $\vec{\nabla} f(-2,3)$. To find the angle corresponding to this unit vector, we solve the equations

$$\cos\theta = \frac{-6\sqrt{61}}{61} \text{ and } \sin\theta = \frac{5\sqrt{61}}{61}$$

for θ. Since cosine is negative and sine is positive, the angle must be in the second quadrant. Therefore, $\theta = \pi - \arcsin((5\sqrt{61})/61) \approx 2.45$ rad.

The maximum value of the directional derivative at $(-2,3)$ is $\|\vec{\nabla} f(-2,3)\| = 4\sqrt{61}$ (Figure 13.5.4).

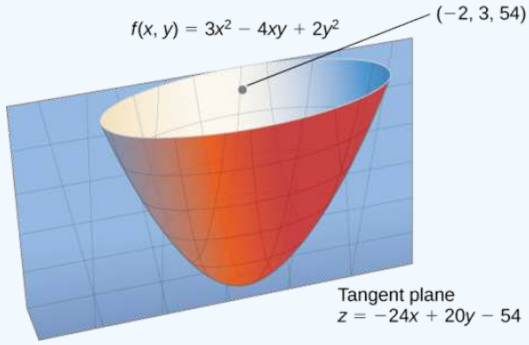


Figure 13.5.4: The maximum value of the directional derivative at $(-2,3)$ is in the direction of the gradient.

? Exercise 13.5.3

Find the direction for which the directional derivative of $g(x,y) = 4x - xy + 2y^2$ at $(-2,3)$ is a maximum. What is the maximum value?

Hint

Evaluate the gradient of g at point $(-2,3)$.

Answer

The gradient of g at $(-2,3)$ is $\vec{\nabla} g(-2,3) = \hat{\mathbf{i}} + 14\,\hat{\mathbf{j}}$. The unit vector that points in the same direction as $\vec{\nabla} g(-2,3)$ is

$$\frac{\vec{\nabla} g(-2,3)}{\|\vec{\nabla} g(-2,3)\|} = \frac{1}{\sqrt{197}}\,\hat{\mathbf{i}} + \frac{14}{\sqrt{197}}\,\hat{\mathbf{j}} = \frac{\sqrt{197}}{197}\,\hat{\mathbf{i}} + \frac{14\sqrt{197}}{197}\,\hat{\mathbf{j}},$$

which gives an angle of $\theta = \arcsin((14\sqrt{197})/197) \approx 1.499$ rad.

The maximum value of the directional derivative is $\|\vec{\nabla} g(-2,3)\| = \sqrt{197}$.

Figure 13.5.5 shows a portion of the graph of the function $f(x,y) = 3 + \sin x \sin y$. Given a point (a,b) in the domain of f, the maximum value of the directional derivative at that point is given by $\|\vec{\nabla} f(a,b)\|$. This would equal the rate of greatest ascent if the surface represented a topographical map. If we went in the opposite direction, it would be the rate of greatest descent.

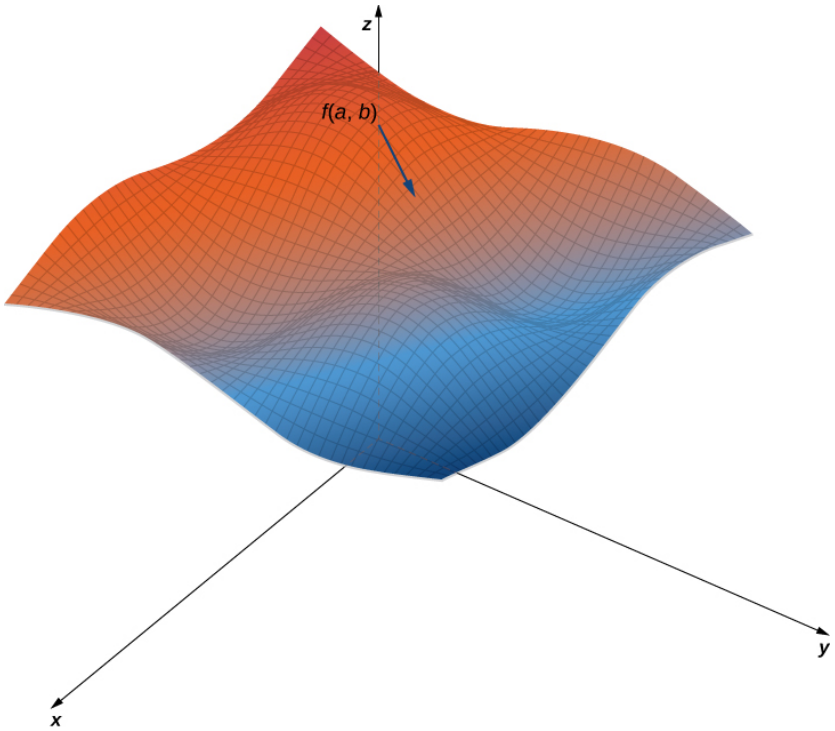


Figure 13.5.5: A typical surface in $\mathbb{R}^3$. Given a point on the surface, the directional derivative can be calculated using the gradient.

When using a topographical map, the steepest slope is always in the direction where the contour lines are closest together (Figure 13.5.6). This is analogous to the contour map of a function, assuming the level curves are obtained for equally spaced values throughout the range of that function.

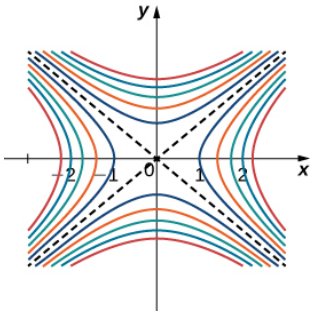


Figure 13.5.6: Contour map for the function $f(x,y) = x^2 - y^2$ using level values between -5 and 5.

Gradients and Level Curves

Recall that if a curve is defined parametrically by the function pair $(x(t), y(t))$, then the vector $x'(t)\,\hat{\mathbf{i}} + y'(t)\,\hat{\mathbf{j}}$ is tangent to the curve for every value of t in the domain. Now let's assume $z = f(x,y)$ is a differentiable function of x and y, and (x_0, y_0) is in its domain. Let's suppose further that $x_0 = x(t_0)$ and $y_0 = y(t_0)$ for some value of t, and consider the level curve $f(x,y) = k$. Define $g(t) = f(x(t), y(t))$ and calculate $g'(t)$ on the level curve. By the chain Rule,

$$g'(t) = f_x(x(t), y(t))x'(t) + f_y(x(t), y(t))y'(t).$$

But $g'(t) = 0$ because $g(t) = k$ for all t. Therefore, on the one hand,

$$f_x(x(t), y(t))x'(t) + f_y(x(t), y(t))y'(t) = 0;$$

on the other hand,

$$f_x(x(t), y(t))x'(t) + f_y(x(t), y(t))y'(t) = \vec{\nabla} f(x,y) \cdot \langle x'(t), y'(t) \rangle.$$

Therefore,

$$\vec{\nabla} f(x,y) \cdot \langle x'(t), y'(t) \rangle = 0.$$

Thus, the dot product of these vectors is equal to zero, which implies they are orthogonal. However, the second vector is tangent to the level curve, which implies the gradient must be normal to the level curve, which gives rise to the following theorem.

Theorem 13.5.3: Gradient Is Normal to the Level Curve

Suppose the function $z = f(x, y)$ has continuous first-order partial derivatives in an open disk centered at a point (x_0, y_0). If $\vec{\nabla} f(x_0, y_0) \neq 0$, then $\vec{\nabla} f(x_0, y_0)$ is normal to the level curve of f at (x_0, y_0).

We can use this theorem to find tangent and normal vectors to level curves of a function.

Example 13.5.5: Finding Tangents to Level Curves

For the function $f(x, y) = 2x^2 - 3xy + 8y^2 + 2x - 4y + 4$, find a tangent vector to the level curve at point $(-2, 1)$. Graph the level curve corresponding to $f(x, y) = 18$ and draw in $\vec{\nabla} f(-2, 1)$ and a tangent vector.

Solution

First, we must calculate $\vec{\nabla} f(x, y)$:

$$f_x(x, y) = 4x - 3y + 2 \text{ and } f_y = -3x + 16y - 4 \text{ so } \vec{\nabla} f(x, y) = (4x - 3y + 2)\,\hat{\mathbf{i}} + (-3x + 16y - 4)\,\hat{\mathbf{j}}.$$

Next, we evaluate $\vec{\nabla} f(x, y)$ at $(-2, 1)$:

$$\vec{\nabla} f(-2, 1) = (4(-2) - 3(1) + 2)\,\hat{\mathbf{i}} + (-3(-2) + 16(1) - 4)\,\hat{\mathbf{j}} = -9\,\hat{\mathbf{i}} + 18\,\hat{\mathbf{j}}.$$

This vector is orthogonal to the curve at point $(-2, 1)$. We can obtain a tangent vector by reversing the components and multiplying either one by -1. Thus, for example, $-18\,\hat{\mathbf{i}} - 9\,\hat{\mathbf{j}}$ is a tangent vector (Figure 13.5.7).

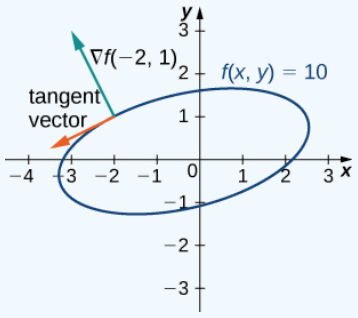


Figure 13.5.7: Tangent and normal vectors to $2x^2 - 3xy + 8y^2 + 2x - 4y + 4 = 18$ at point $(-2, 1)$.

Exercise 13.5.4

For the function $f(x, y) = x^2 - 2xy + 5y^2 + 3x - 2y + 3$, find the tangent to the level curve at point $(1, 1)$. Draw the graph of the level curve corresponding to $f(x, y) = 8$ and draw $\vec{\nabla} f(1, 1)$ and a tangent vector.

Hint

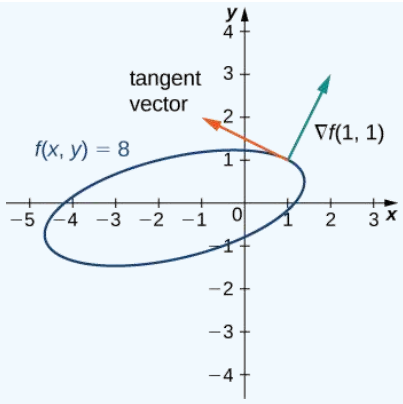
Calculate the gradient at point $(1, 1)$.

Answer

$\vec{\nabla} f(x, y) = (2x - 2y + 3)\,\hat{\mathbf{i}} + (-2x + 10y - 2)\,\hat{\mathbf{j}}$

$\vec{\nabla} f(1, 1) = 3\,\hat{\mathbf{i}} + 6\,\hat{\mathbf{j}}$

Tangent vector: $6\,\hat{\mathbf{i}} - 3\,\hat{\mathbf{j}}$ or $-6\,\hat{\mathbf{i}} + 3\,\hat{\mathbf{j}}$



Three-Dimensional Gradients and Directional Derivatives

The definition of a gradient can be extended to functions of more than two variables.

Definition: Gradients in 3D

Let $w = f(x, y, z)$ be a function of three variables such that f_x, f_y, and f_z exist. The vector $\vec{\nabla} f(x, y, z)$ is called the gradient of f and is defined as

$$\vec{\nabla} f(x, y, z) = f_x(x, y, z)\,\hat{\mathbf{i}} + f_y(x, y, z)\,\hat{\mathbf{j}} + f_z(x, y, z)\,\hat{\mathbf{k}}. \tag{13.5.6}$$

$\vec{\nabla} f(x, y, z)$ can also be written as **grad** $f(x, y, z)$.

Calculating the gradient of a function in three variables is very similar to calculating the gradient of a function in two variables. First, we calculate the partial derivatives f_x, f_y, and f_z, and then we use Equation 13.5.6.

Example 13.5.6: Finding Gradients in Three Dimensions

Find the gradient $\vec{\nabla} f(x, y, z)$ of each of the following functions:

a. $f(x, y, z) = 5x^2 - 2xy + y^2 - 4yz + z^2 + 3xz$
b. $f(x, y, z) = e^{-2z} \sin 2x \cos 2y$

Solution

For both parts a. and b., we first calculate the partial derivatives f_x, f_y, and f_z, then use Equation 13.5.6.

a. $f_x(x, y, z) = 10x - 2y + 3z$, $f_y(x, y, z) = -2x + 2y - 4z$, and $f_z(x, y, z) = 3x - 4y + 2z$, so

$$\begin{aligned}\vec{\nabla} f(x, y, z) &= f_x(x, y, z)\,\hat{\mathbf{i}} + f_y(x, y, z)\,\hat{\mathbf{j}} + f_z(x, y, z)\,\hat{\mathbf{k}} \\ &= (10x - 2y + 3z)\,\hat{\mathbf{i}} + (-2x + 2y - 4z)\,\hat{\mathbf{j}} + (3x - 4y + 2z)\,\hat{\mathbf{k}}.\end{aligned}$$

b. $f_x(x, y, z) = 2e^{-2z} \cos 2x \cos 2y$, $f_y(x, y, z) = -2e^{-2z} \sin 2x \sin 2y$, and $f_z(x, y, z) = -2e^{-2z} \sin 2x \cos 2y$, so

$$\begin{aligned}\vec{\nabla} f(x, y, z) &= f_x(x, y, z)\,\hat{\mathbf{i}} + f_y(x, y, z)\,\hat{\mathbf{j}} + f_z(x, y, z)\,\hat{\mathbf{k}} \\ &= (2e^{-2z} \cos 2x \cos 2y)\,\hat{\mathbf{i}} + (-2e^{-2z} \sin 2x \sin 2y)\,\hat{\mathbf{j}} + (-2e^{-2z} \sin 2x \cos 2y)\,\hat{\mathbf{k}} \\ &= 2e^{-2z}(\cos 2x \cos 2y\,\hat{\mathbf{i}} - \sin 2x \sin 2y\,\hat{\mathbf{j}} - \sin 2x \cos 2y\,\hat{\mathbf{k}}).\end{aligned}$$

Exercise 13.5.5:

Find the gradient $\vec{\nabla} f(x, y, z)$ of $f(x, y, z) = \dfrac{x^2 - 3y^2 + z^2}{2x + y - 4z}$.

Answer

$$\vec{\nabla} f(x, y, z) = \frac{2x^2 + 2xy + 6y^2 - 8xz - 2z^2}{(2x + y - 4z)^2}\,\hat{\mathbf{i}} - \frac{x^2 + 12xy + 3y^2 - 24yz + z^2}{(2x + y - 4z)^2}\,\hat{\mathbf{j}} + \frac{4x^2 - 12y^2 - 4z^2 + 4xz + 2yz}{(2x + y - 4z)^2}\,\hat{\mathbf{k}}$$

The directional derivative can also be generalized to functions of three variables. To determine a direction in three dimensions, a vector with three components is needed. This vector is a unit vector, and the components of the unit vector are called *directional cosines***.** Given a three-dimensional unit vector $\vec{\mathbf{u}}$ in standard form (i.e., the initial point is at the origin), this vector forms three different angles with the positive x-, y-, and z-axes. Let's call these angles α, β, and γ. Then the directional cosines are given by $\cos\alpha, \cos\beta$, and $\cos\gamma$. These are the components of the unit vector $\vec{\mathbf{u}}$; since $\vec{\mathbf{u}}$ is a unit vector, it is true that $\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1$.

Definition: Directional Derivative of a Function of Three variables

Suppose $w = f(x,y,z)$ is a function of three variables with a domain of D. Let $(x_0, y_0, z_0) \in D$ and let $\vec{\mathbf{u}} = \cos\alpha\,\hat{\mathbf{i}} + \cos\beta\,\hat{\mathbf{j}} + \cos\gamma\,\hat{\mathbf{k}}$ be a unit vector. Then, the directional derivative of f in the direction of u is given by

$$D_{\vec{\mathbf{u}}} f(x_0, y_0, z_0) = \lim_{t\to 0} \frac{f(x_0 + t\cos\alpha, y_0 + t\cos\beta, z_0 + t\cos\gamma) - f(x_0, y_0, z_0)}{t}$$

provided the limit exists.

We can calculate the directional derivative of a function of three variables by using the gradient, leading to a formula that is analogous to Equation 13.5.3.

Theorem 13.5.4: Directional Derivative of a Function of Three Variables

Let $f(x,y,z)$ be a differentiable function of three variables and let $\vec{\mathbf{u}} = \cos\alpha\,\hat{\mathbf{i}} + \cos\beta\,\hat{\mathbf{j}} + \cos\gamma\,\hat{\mathbf{k}}$ be a unit vector. Then, the directional derivative of f in the direction of $\vec{\mathbf{u}}$ is given by

$$D_{\vec{\mathbf{u}}} f(x,y,z) = \vec{\nabla} f(x,y,z) \cdot \vec{\mathbf{u}} = f_x(x,y,z)\cos\alpha + f_y(x,y,z)\cos\beta + f_z(x,y,z)\cos\gamma. \tag{13.5.7}$$

The three angles α, β, and γ determine the unit vector $\vec{\mathbf{u}}$. In practice, we can use an arbitrary (nonunit) vector, then divide by its magnitude to obtain a unit vector in the desired direction.

Example 13.5.7: Finding a Directional Derivative in Three Dimensions

Calculate $D_{\vec{\mathbf{v}}} f(1,-2,3)$ in the direction of $\vec{\mathbf{v}} = -\hat{\mathbf{i}} + 2\hat{\mathbf{j}} + 2\hat{\mathbf{k}}$ for the function

$$f(x,y,z) = 5x^2 - 2xy + y^2 - 4yz + z^2 + 3xz.$$

Solution:

First, we find the magnitude of v:

$$\|\vec{\mathbf{v}}\| = \sqrt{(-1)^2 + (2)^2 + (2)^2} = \sqrt{9} = 3.$$

Therefore, $\dfrac{\vec{\mathbf{v}}}{\|\vec{\mathbf{v}}\|} = \dfrac{-\hat{\mathbf{i}} + 2\hat{\mathbf{j}} + 2\hat{\mathbf{k}}}{3} = -\dfrac{1}{3}\hat{\mathbf{i}} + \dfrac{2}{3}\hat{\mathbf{j}} + \dfrac{2}{3}\hat{\mathbf{k}}$ is a unit vector in the direction of $\vec{\mathbf{v}}$, so $\cos\alpha = -\dfrac{1}{3}, \cos\beta = \dfrac{2}{3}$, and $\cos\gamma = \dfrac{2}{3}$. Next, we calculate the partial derivatives of f:

$$\begin{aligned} f_x(x,y,z) &= 10x - 2y + 3z \\ f_y(x,y,z) &= -2x + 2y - 4z \\ f_z(x,y,z) &= -4y + 2z + 3x, \end{aligned}$$

then substitute them into Equation 13.5.7:

$$\begin{aligned} D_{\vec{\mathbf{v}}} f(x,y,z) &= f_x(x,y,z)\cos\alpha + f_y(x,y,z)\cos\beta + f_z(x,y,z)\cos\gamma \\ &= (10x - 2y + 3z)\left(-\tfrac{1}{3}\right) + (-2x + 2y - 4z)\left(\tfrac{2}{3}\right) + (-4y + 2z + 3x)\left(\tfrac{2}{3}\right) \\ &= -\frac{10x}{3} + \frac{2y}{3} - \frac{3z}{3} - \frac{4x}{3} + \frac{4y}{3} - \frac{8z}{3} - \frac{8y}{3} + \frac{4z}{3} + \frac{6x}{3} \\ &= -\frac{8x}{3} - \frac{2y}{3} - \frac{7z}{3}. \end{aligned}$$

Last, to find $D_{\vec{\mathbf{v}}} f(1,-2,3)$, we substitute $x = 1,\ y = -2$, and $z = 3$:

$$\begin{aligned} D_{\vec{\mathbf{v}}} f(1,-2,3) &= -\frac{8(1)}{3} - \frac{2(-2)}{3} - \frac{7(3)}{3} \\ &= -\frac{8}{3} + \frac{4}{3} - \frac{21}{3} \\ &= -\frac{25}{3}. \end{aligned}$$

? Exercise 13.5.6:

Calculate $D_{\vec{\mathbf{v}}} f(x,y,z)$ and $D_{\vec{\mathbf{v}}} f(0,-2,5)$ in the direction of $\vec{\mathbf{v}} = -3\,\hat{\mathbf{i}} + 12\,\hat{\mathbf{j}} - 4\,\hat{\mathbf{k}}$ for the function

$$f(x,y,z) = 3x^2 + xy - 2y^2 + 4yz - z^2 + 2xz.$$

Hint

First, divide $\vec{\mathbf{v}}$ by its magnitude, calculate the partial derivatives of f, then use Equation 13.5.7.

Answer

$$D_{\vec{\mathbf{v}}} f(x,y,z) = -\frac{3}{13}(6x + y + 2z) + \frac{12}{13}(x - 4y + 4z) - \frac{4}{13}(2x + 4y - 2z)$$

$$D_{\vec{\mathbf{v}}} f(0,-2,5) = \frac{384}{13}$$

Summary

- A directional derivative represents a rate of change of a function in any given direction.
- The gradient can be used in a formula to calculate the directional derivative.
- The gradient indicates the direction of greatest change of a function of more than one variable.

Key Equations

- **directional derivative (two dimensions)**

$$D_{\vec{\mathbf{u}}} f(a,b) = \lim_{h \to 0} \frac{f(a + h\cos\theta, b + h\sin\theta) - f(a,b)}{h}$$

or

$$D_{\vec{\mathbf{u}}} f(x,y) = f_x(x,y)\cos\theta + f_y(x,y)\sin\theta$$

- **gradient (two dimensions)**

$$\vec{\nabla} f(x,y) = f_x(x,y)\,\hat{\mathbf{i}} + f_y(x,y)\,\hat{\mathbf{j}}$$

- **gradient (three dimensions)**

$$\vec{\nabla} f(x,y,z) = f_x(x,y,z)\,\hat{\mathbf{i}} + f_y(x,y,z)\,\hat{\mathbf{j}} + f_z(x,y,z)\,\hat{\mathbf{k}}$$

- **directional derivative (three dimensions)**

$$D_{\vec{\mathbf{u}}} f(x,y,z) = \vec{\nabla} f(x,y,z) \cdot \vec{\mathbf{u}} = f_x(x,y,z)\cos\alpha + f_y(x,y,z)\cos\beta + f_x(x,y,z)\cos\gamma$$

Glossary

directional derivative

the derivative of a function in the direction of a given unit vector

gradient

the gradient of the function $f(x,y)$ is defined to be $\vec{\nabla} f(x,y) = (\partial f/\partial x)\,\hat{\mathbf{i}} + (\partial f/\partial y)\,\hat{\mathbf{j}}$, which can be generalized to a function of any number of independent variables

13.5: Directional Derivatives and Gradient Vectors is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **14.6: Directional Derivatives and the Gradient** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

13.6: Tangent Planes and Differentials

Learning Objectives

- Determine the equation of a plane tangent to a given surface at a point.
- Use the tangent plane to approximate a function of two variables at a point.
- Explain when a function of two variables is differentiable.
- Use the total differential to approximate the change in a function of two variables.

In this section, we consider the problem of finding the tangent plane to a surface, which is analogous to finding the equation of a tangent line to a curve when the curve is defined by the graph of a function of one variable, $y = f(x)$. The slope of the tangent line at the point $x = a$ is given by $m = f'(a)$; what is the slope of a tangent plane? We learned about the equation of a plane in Equations of Lines and Planes in Space; in this section, we see how it can be applied to the problem at hand.

Tangent Planes

Intuitively, it seems clear that, in a plane, only one line can be tangent to a curve at a point. However, in three-dimensional space, many lines can be tangent to a given point. If these lines lie in the same plane, they determine the tangent plane at that point. A more intuitive way to think of a tangent plane is to assume the surface is smooth at that point (no corners). Then, a tangent line to the surface at that point in any direction does not have any abrupt changes in slope because the direction changes smoothly. Therefore, in a small-enough neighborhood around the point, a tangent plane touches the surface at that point only.

Definition: Tangent Lines

Let $P_0 = (x_0, y_0, z_0)$ be a point on a surface S, and let C be any curve passing through P_0 and lying entirely in S. If the tangent lines to all such curves C at P_0 lie in the same plane, then this plane is called the *tangent plane* to S at P_0 (Figure 13.6.1).

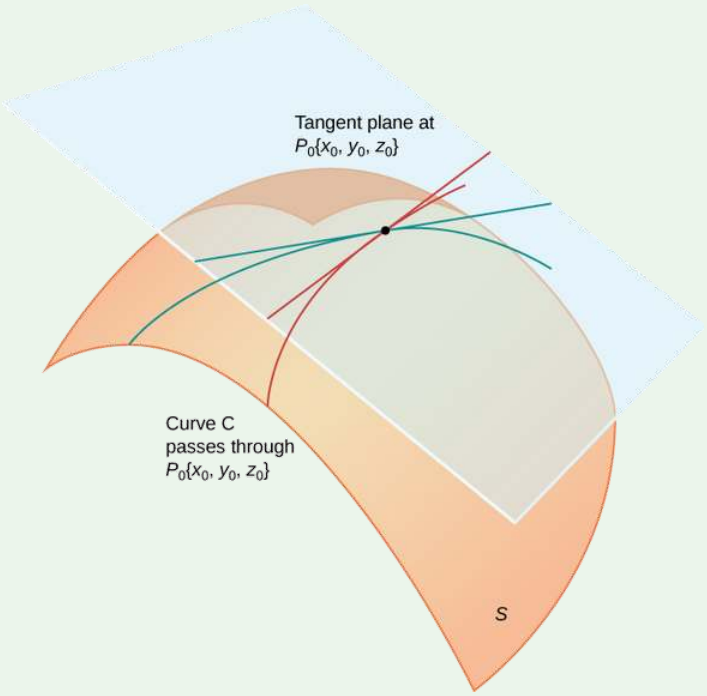


Figure 13.6.1: The tangent plane to a surface S at a point P_0 contains all the tangent lines to curves in S *that pass through* P_0.

For a tangent plane to a surface to exist at a point on that surface, it is sufficient for the function that defines the surface to be differentiable at that point. We define the term tangent plane here and then explore the idea intuitively.

Definition: Tangent Planes

Let S be a surface defined by a differentiable function $z = f(x, y)$, and let $P_0 = (x_0, y_0)$ be a point in the domain of f. Then, the equation of the tangent plane to S at P_0 is given by

$$z = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0). \tag{13.6.1}$$

To see why this formula is correct, let's first find two tangent lines to the surface S. The equation of the tangent line to the curve that is represented by the intersection of S with the vertical trace given by $x = x_0$ is $z = f(x_0, y_0) + f_y(x_0, y_0)(y - y_0)$. Similarly, the equation of the tangent line to the curve that is represented by the intersection of S with the vertical trace given by $y = y_0$ is $z = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0)$. A parallel vector to the first tangent line is $\vec{\mathbf{a}} = \hat{\mathbf{j}} + f_y(x_0, y_0)\,\hat{\mathbf{k}}$; a parallel vector to the second tangent line is $\vec{\mathbf{b}} = \hat{\mathbf{i}} + f_x(x_0, y_0)\,\hat{\mathbf{k}}$. We can take the cross product of these two vectors:

$$\begin{aligned}
\vec{\mathbf{a}} \times \vec{\mathbf{b}} &= (\hat{\mathbf{j}} + f_y(x_0, y_0)\,\hat{\mathbf{k}}) \times (\hat{\mathbf{i}} + f_x(x_0, y_0)\,\hat{\mathbf{k}}) \\
&= \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ 0 & 1 & f_y(x_0, y_0) \\ 1 & 0 & f_x(x_0, y_0) \end{vmatrix} \\
&= f_x(x_0, y_0)\,\hat{\mathbf{i}} + f_y(x_0, y_0)\,\hat{\mathbf{j}} - \hat{\mathbf{k}}.
\end{aligned}$$

This vector is perpendicular to both lines and is therefore perpendicular to the tangent plane. We can use this vector as a normal vector to the tangent plane, along with the point $P_0 = (x_0, y_0, f(x_0, y_0))$ in the equation for a plane:

$$\begin{aligned}
\vec{\mathbf{n}} \cdot ((x - x_0)\,\hat{\mathbf{i}} + (y - y_0)\,\hat{\mathbf{j}} + (z - f(x_0, y_0))\,\hat{\mathbf{k}}) &= 0 \\
(f_x(x_0, y_0)\,\hat{\mathbf{i}} + f_y(x_0, y_0)\,\hat{\mathbf{j}} - \hat{\mathbf{k}}) \cdot ((x - x_0)\,\hat{\mathbf{i}} + (y - y_0)\,\hat{\mathbf{j}} + (z - f(x_0, y_0))\,\hat{\mathbf{k}}) &= 0 \\
f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) - (z - f(x_0, y_0)) &= 0.
\end{aligned}$$

Solving this equation for z gives Equation 13.6.1.

Example 13.6.1: Finding a Tangent Plane

Find the equation of the tangent plane to the surface defined by the function $f(x, y) = 2x^2 - 3xy + 8y^2 + 2x - 4y + 4$ at point $(2, -1)$.

Solution

First, we must calculate $f_x(x, y)$ and $f_y(x, y)$, then use Equation with $x_0 = 2$ and $y_0 = -1$:

$$\begin{aligned}
f_x(x, y) &= 4x - 3y + 2 \\
f_y(x, y) &= -3x + 16y - 4 \\
f(2, -1) &= 2(2)^2 - 3(2)(-1) + 8(-1)^2 + 2(2) - 4(-1) + 4 = 34 \\
f_x(2, -1) &= 4(2) - 3(-1) + 2 = 13 \\
f_y(2, -1) &= -3(2) + 16(-1) - 4 = -26.
\end{aligned}$$

Then Equation 13.6.1 becomes

$$\begin{aligned}
z &= f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) \\
z &= 34 + 13(x - 2) - 26(y - (-1)) \\
z &= 34 + 13x - 26 - 26y - 26 \\
z &= 13x - 26y - 18.
\end{aligned}$$

(See the following figure).

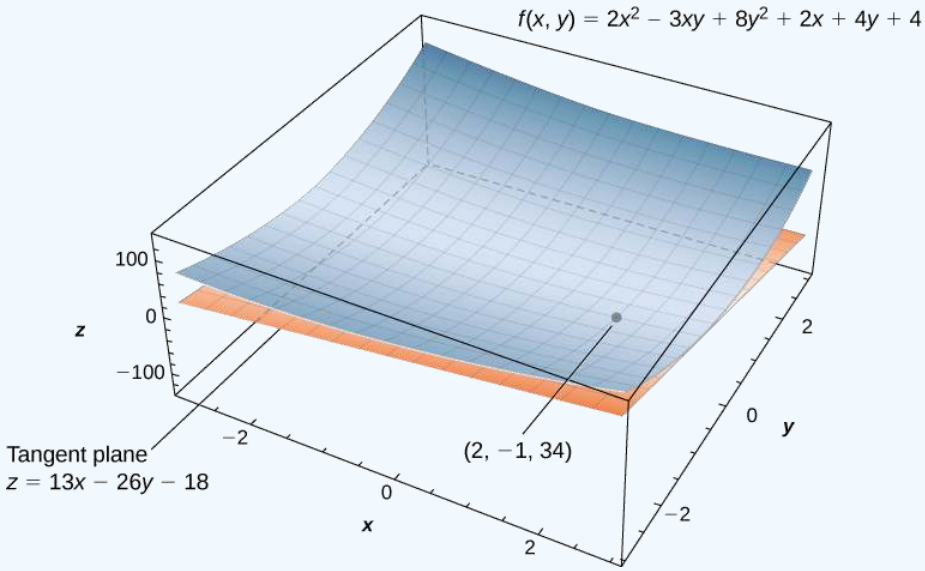


Figure 13.6.2: Calculating the equation of a tangent plane to a given surface at a given point.

? Exercise 13.6.1

Find the equation of the tangent plane to the surface defined by the function $f(x,y) = x^3 - x^2y + y^2 - 2x + 3y - 2$ at point $(-1, 3)$.

Hint

First, calculate $f_x(x,y)$ and $f_y(x,y)$, then use Equation 13.6.1.

Answer

$z = 7x + 8y - 3$

✓ Example 13.6.2: Finding Another Tangent Plane

Find the equation of the tangent plane to the surface defined by the function $f(x,y) = \sin(2x)\cos(3y)$ at the point $(\pi/3, \pi/4)$.

Solution

First, calculate $f_x(x,y)$ and $f_y(x,y)$, then use Equation 13.6.1 with $x_0 = \pi/3$ and $y_0 = \pi/4$:

$$f_x(x,y) = 2\cos(2x)\cos(3y)$$
$$f_y(x,y) = -3\sin(2x)\sin(3y)$$
$$f\left(\frac{\pi}{3}, \frac{\pi}{4}\right) = \sin\left(2\left(\frac{\pi}{3}\right)\right)\cos\left(3\left(\frac{\pi}{4}\right)\right) = \left(\frac{\sqrt{3}}{2}\right)\left(-\frac{\sqrt{2}}{2}\right) = -\frac{\sqrt{6}}{4}$$
$$f_x\left(\frac{\pi}{3}, \frac{\pi}{4}\right) = 2\cos\left(2\left(\frac{\pi}{3}\right)\right)\cos\left(3\left(\frac{\pi}{4}\right)\right) = 2\left(-\frac{1}{2}\right)\left(-\frac{\sqrt{2}}{2}\right) = \frac{\sqrt{2}}{2}$$
$$f_y\left(\frac{\pi}{3}, \frac{\pi}{4}\right) = -3\sin\left(2\left(\frac{\pi}{3}\right)\right)\sin\left(3\left(\frac{\pi}{4}\right)\right) = -3\left(\frac{\sqrt{3}}{2}\right)\left(\frac{\sqrt{2}}{2}\right) = -\frac{3\sqrt{6}}{4}.$$

Then Equation 13.6.1 becomes

$$\begin{aligned} z &= f(x_0,y_0)+f_x(x_0,y_0)(x-x_0)+f_y(x_0,y_0)(y-y_0) \\ &= -\frac{\sqrt{6}}{4}+\frac{\sqrt{2}}{2}\left(x-\frac{\pi}{3}\right)-\frac{3\sqrt{6}}{4}\left(y-\frac{\pi}{4}\right) \\ &= \frac{\sqrt{2}}{2}x-\frac{3\sqrt{6}}{4}y-\frac{\sqrt{6}}{4}-\frac{\pi\sqrt{2}}{6}+\frac{3\pi\sqrt{6}}{16} \end{aligned}$$

A tangent plane to a surface does not always exist at every point on the surface. Consider the piecewise function

$$f(x,y)=\begin{cases} \dfrac{xy}{\sqrt{x^2+y^2}}, & (x,y)\neq(0,0) \\ 0, & (x,y)=(0,0) \end{cases}. \tag{13.6.2}$$

The graph of this function follows.

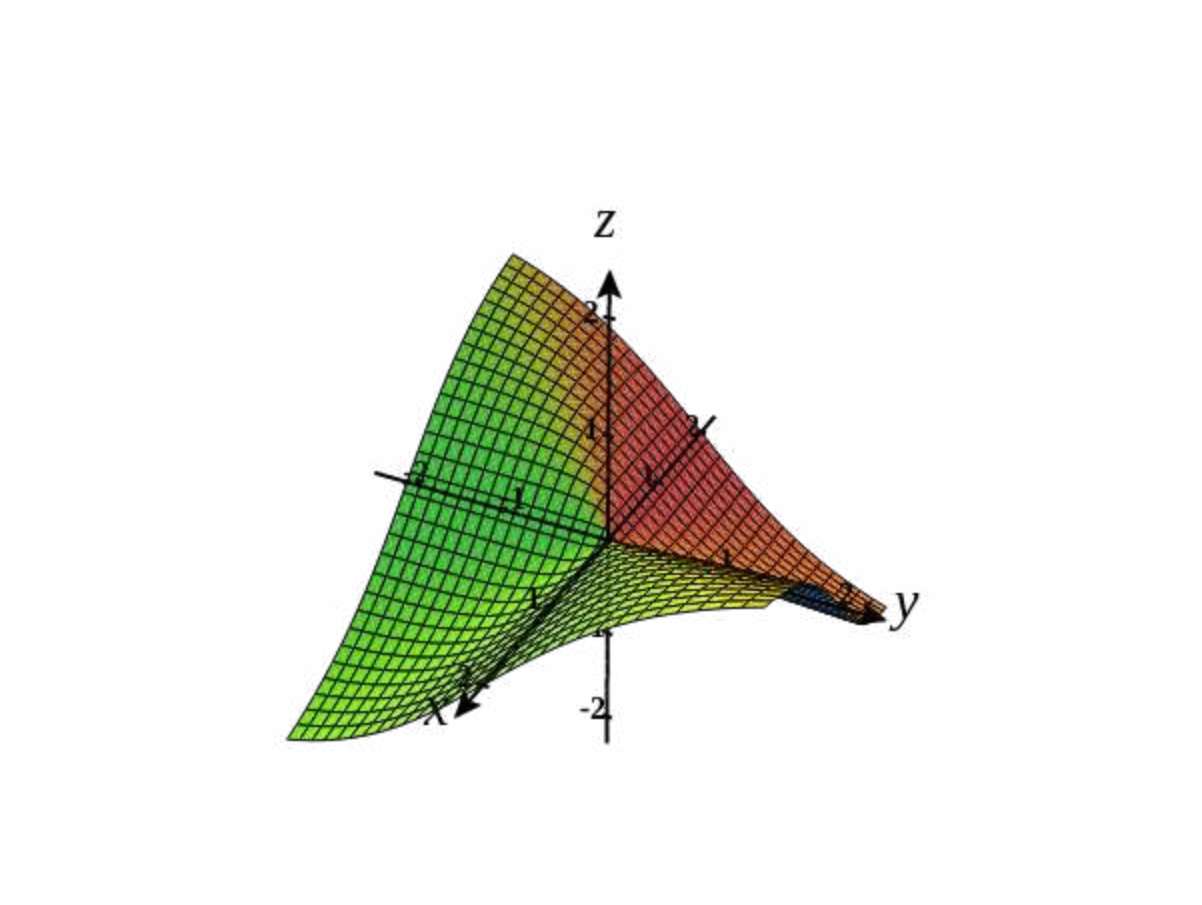


Figure* 13.6.3*: *Graph of a function that does not have a tangent plane at the origin. Dynamic figure powered by CalcPlot3D.*

If either $x=0$ or $y=0$, then $f(x,y)=0$, so the value of the function does not change on either the x- or y-axis. Therefore, $f_x(x,0)=f_y(0,y)=0$, so as either x or y approach zero, these partial derivatives stay equal to zero. Substituting them into Equation gives $z=0$ as the equation of the tangent line. However, if we approach the origin from a different direction, we get a different story. For example, suppose we approach the origin along the line $y=x$. If we put $y=x$ into the original function, it becomes

$$f(x,x)=\frac{x(x)}{\sqrt{x^2+(x)^2}}=\frac{x^2}{\sqrt{2x^2}}=\frac{|x|}{\sqrt{2}}.$$

When $x > 0$, the slope of this curve is equal to $\sqrt{2}/2$; when $x < 0$, the slope of this curve is equal to $-(\sqrt{2}/2)$. This presents a problem. In the definition of tangent plane, we presumed that all tangent lines through point P (in this case, the origin) lay in the same plane. This is clearly not the case here. When we study differentiable functions, we will see that this function is not differentiable at the origin.

Linear Approximations

Recall from Linear Approximations and Differentials that the formula for the linear approximation of a function $f(x)$ at the point $x = a$ is given by

$$y \approx f(a) + f'(a)(x - a).$$

The diagram for the linear approximation of a function of one variable appears in the following graph.

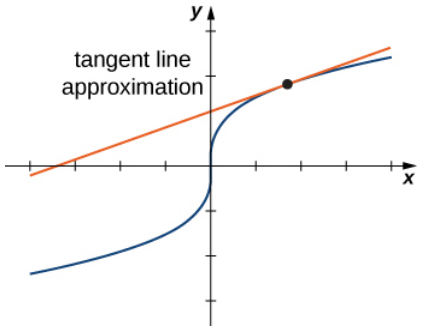


Figure 13.6.4: Linear approximation of a function in one variable.

The tangent line can be used as an approximation to the function $f(x)$ for values of x reasonably close to $x = a$. When working with a function of two variables, the tangent line is replaced by a tangent plane, but the approximation idea is much the same.

Definition: Linear Approximation

Given a function $z = f(x, y)$ with continuous partial derivatives that exist at the point (x_0, y_0), the linear approximation of f at the point (x_0, y_0) is given by the equation

$$L(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0). \tag{13.6.3}$$

Notice that this equation also represents the tangent plane to the surface defined by $z = f(x, y)$ at the point (x_0, y_0). The idea behind using a linear approximation is that, if there is a point (x_0, y_0) at which the precise value of $f(x, y)$ is known, then for values of (x, y) reasonably close to (x_0, y_0), the linear approximation (i.e., tangent plane) yields a value that is also reasonably close to the exact value of $f(x, y)$ (Figure 13.6.5). Furthermore the plane that is used to find the linear approximation is also the tangent plane to the surface at the point (x_0, y_0).

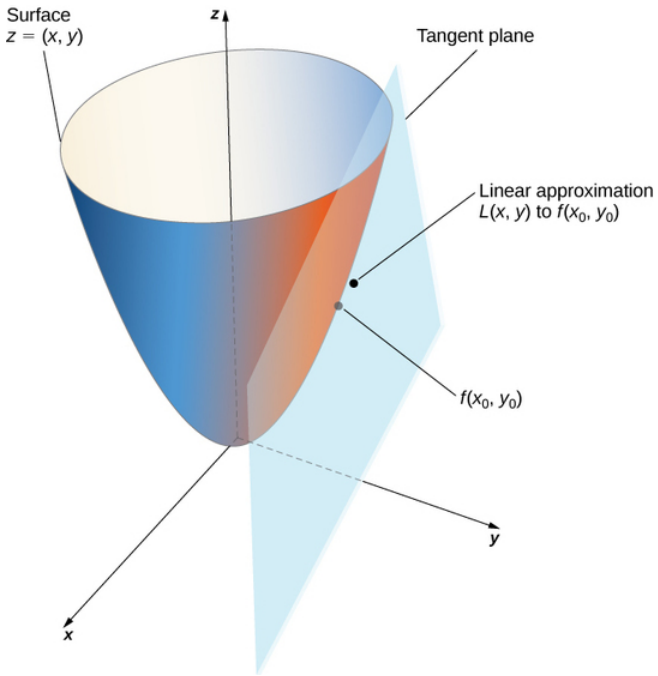


Figure 13.6.5: Using a tangent plane for linear approximation at a point.

Example 13.6.3: Using a Tangent Plane Approximation

Given the function $f(x,y)=\sqrt{41-4x^2-y^2}$, approximate $f(2.1,2.9)$ using point $(2,3)$ for (x_0,y_0). What is the approximate value of $f(2.1,2.9)$to four decimal places?

Solution

To apply Equation 13.6.3, we first must calculate $f(x_0,y_0)$, $f_x(x_0,y_0)$, and $f_y(x_0,y_0)$ using $x_0=2$ and $y_0=3$:

$$\begin{aligned} f(x_0,y_0) &= f(2,3)=\sqrt{41-4(2)^2-(3)^2}=\sqrt{41-16-9}=\sqrt{16}=4 \\ f_x(x,y) &= -\frac{4x}{\sqrt{41-4x^2-y^2}} \text{ so } f_x(x_0,y_0)=-\frac{4(2)}{\sqrt{41-4(2)^2-(3)^2}}=-2 \\ f_y(x,y) &= -\frac{y}{\sqrt{41-4x^2-y^2}} \text{ so } f_y(x_0,y_0)=-\frac{3}{\sqrt{41-4(2)^2-(3)^2}}=-\frac{3}{4}. \end{aligned}$$

Now we substitute these values into Equation 13.6.3:

$$\begin{aligned} L(x,y) &= f(x_0,y_0)+f_x(x_0,y_0)(x-x_0)+f_y(x_0,y_0)(y-y_0) \\ &= 4-2(x-2)-\frac{3}{4}(y-3) \\ &= \frac{41}{4}-2x-\frac{3}{4}y. \end{aligned}$$

Last, we substitute $x=2.1$ and $y=2.9$ into $L(x,y)$:

$$L(2.1,2.9)=\frac{41}{4}-2(2.1)-\frac{3}{4}(2.9)=10.25-4.2-2.175=3.875.$$

The approximate value of $f(2.1,2.9)$to four decimal places is

$$f(2.1,2.9)=\sqrt{41-4(2.1)^2-(2.9)^2}=\sqrt{14.95}\approx 3.8665,$$

which corresponds to a 0.2 error in approximation.

Exercise 13.6.2

Given the function $f(x,y)=e^{5-2x+3y}$, approximate $f(4.1,0.9)$using point $(4,1)$ for (x_0,y_0). What is the approximate value of $f(4.1,0.9)$to four decimal places?

Hint

First calculate $f(x_0,y_0)$, $f_x(x_0,y_0)$, and $f_y(x_0,y_0)$ using $x_0=4$ and $y_0=1$, then use Equation 13.6.3

Answer

$L(x,y)=6-2x+3y,$ so $L(4.1,0.9)=6-2(4.1)+3(0.9)=0.5\, f(4.1,0.9)=e^{5-2(4.1)+3(0.9)}=e^{-0.5}\approx 0.6065.$

Differentiability

When working with a function $y=f(x)$ of one variable, the function is said to be differentiable at a point $x=a$ if $f'(a)$ exists. Furthermore, if a function of one variable is differentiable at a point, the graph is "smooth" at that point (i.e., no corners exist) and a tangent line is well-defined at that point.

The idea behind differentiability of a function of two variables is connected to the idea of smoothness at that point. In this case, a surface is considered to be smooth at point P if a tangent plane to the surface exists at that point. If a function is differentiable at a point, then a tangent plane to the surface exists at that point. Recall the formula (Equation 13.6.1) for a tangent plane at a point (x_0,y_0) is given by

$$z=f(x_0,y_0)+f_x(x_0,y_0)(x-x_0)+f_y(x_0,y_0)(y-y_0)$$

For a tangent plane to exist at the point (x_0,y_0), the partial derivatives must therefore exist at that point. However, this is not a sufficient condition for smoothness, as was illustrated in Figure 13.6.3. In that case, the partial derivatives existed at the origin, but the function also had a corner on the graph at the origin.

Definition: Differentiable Functions

A function $f(x,y)$is *differentiable* at a point $P(x_0,y_0)$ if, for all points (x,y) in a δ disk around P, we can write

$$f(x,y)=f(x_0,y_0)+f_x(x_0,y_0)(x-x_0)+f_y(x_0,y_0)(y-y_0)+E(x,y), \tag{13.6.4}$$

where the error term E satisfies

$$\lim_{(x,y)\to(x_0,y_0)}\frac{E(x,y)}{\sqrt{(x-x_0)^2+(y-y_0)^2}}=0. \tag{13.6.5}$$

The last term in Equation 13.6.4 is referred to as the **error term** and it represents how closely the tangent plane comes to the surface in a small neighborhood (δ disk) of point P. For the function f to be differentiable at P, the function must be smooth—that is, the graph of f must be close to the tangent plane for points near P.

Example 13.6.4: Demonstrating Differentiability

Show that the function $f(x,y)=2x^2-4y$ is differentiable at point $(2,-3)$.

Solution

First, we calculate $f(x_0,y_0)$, $f_x(x_0,y_0)$, and $f_y(x_0,y_0)$ using $x_0=2$ and $y_0=-3,$ then we use Equation 13.6.4:

$$\begin{aligned} f(2,-3) &= 2(2)^2-4(-3)=8+12=20\\ f_x(2,-3) &= 4(2)=8\\ f_y(2,-3) &= -4. \end{aligned}$$

Therefore $m_1=8$ and $m_2=-4,$ and Equation 13.6.4becomes

$$
\begin{aligned}
f(x,y) &= f(2,-3)+f_x(2,-3)(x-2)+f_y(2,-3)(y+3)+E(x,y)\\
2x^2-4y &= 20+8(x-2)-4(y+3)+E(x,y)\\
2x^2-4y &= 20+8x-16-4y-12+E(x,y)\\
2x^2-4y &= 8x-4y-8+E(x,y)\\
E(x,y) &= 2x^2-8x+8.
\end{aligned}
$$

Next, we calculate the limit in Equation 13.6.5:

$$
\begin{aligned}
\lim_{(x,y)\to(x_0,y_0)} \frac{E(x,y)}{\sqrt{(x-x_0)^2+(y-y_0)^2}} &= \lim_{(x,y)\to(2,-3)} \frac{2x^2-8x+8}{\sqrt{(x-2)^2+(y+3)^2}}\\
&= \lim_{(x,y)\to(2,-3)} \frac{2(x^2-4x+4)}{\sqrt{(x-2)^2+(y+3)^2}}\\
&= \lim_{(x,y)\to(2,-3)} \frac{2(x-2)^2}{\sqrt{(x-2)^2+(y+3)^2}}\\
&\le \lim_{(x,y)\to(2,-3)} \frac{2((x-2)^2+(y+3)^2)}{\sqrt{(x-2)^2+(y+3)^2}}\\
&= \lim_{(x,y)\to(2,-3)} 2\sqrt{(x-2)^2+(y+3)^2}\\
&= 0.
\end{aligned}
$$

Since $E(x,y)\ge 0$ for any value of x or y, the original limit must be equal to zero. Therefore, $f(x,y)=2x^2-4y$ is differentiable at point $(2,-3)$.

? Exercise 13.6.3

Show that the function $f(x,y)=3x-4y^2$ is differentiable at point $(-1,2)$.

Hint

First, calculate $f(x_0,y_0)$, $f_x(x_0,y_0)$, and $f_y(x_0,y_0)$ using $x_0=-1$ and $y_0=2$, then use Equation 13.6.5to find $E(x,y)$. Last, calculate the limit.

Answer

$$
f(-1,2) = -19, \quad f_x(-1,2)=3, \quad f_y(-1,2)=-16, \quad E(x,y)=-4(y-2)^2.
$$

$$
\begin{aligned}
\lim_{(x,y)\to(x_0,y_0)} \frac{E(x,y)}{\sqrt{(x-x_0)^2+(y-y_0)^2}} &= \lim_{(x,y)\to(-1,2)} \frac{-4(y-2)^2}{\sqrt{(x+1)^2+(y-2)^2}}\\
&\le \lim_{(x,y)\to(-1,2)} \frac{-4((x+1)^2+(y-2)^2)}{\sqrt{(x+1)^2+(y-2)^2}}\\
&= \lim_{(x,y)\to(2,-3)} -4\sqrt{(x+1)^2+(y-2)^2}\\
&= 0.
\end{aligned}
$$

This function from (Equation 13.6.2)

$$
f(x,y)=\begin{cases} \dfrac{xy}{\sqrt{x^2+y^2}}, & (x,y)\neq(0,0)\\ 0, & (x,y)=(0,0)\end{cases}
$$

is not differentiable at the origin (Figure 13.6.3). We can see this by calculating the partial derivatives. This function appeared earlier in the section, where we showed that $f_x(0,0) = f_y(0,0) = 0$. Substituting this information into Equations 13.6.4 and 13.6.5 using $x_0 = 0$ and $y_0 = 0$, we get

$$\begin{aligned} f(x,y) &= f(0,0) + f_x(0,0)(x-0) + f_y(0,0)(y-0) + E(x,y) \\ E(x,y) &= \frac{xy}{\sqrt{x^2+y^2}}. \end{aligned}$$

Calculating

$$\lim_{(x,y)\to(x_0,y_0)} \frac{E(x,y)}{\sqrt{(x-x_0)^2+(y-y_0)^2}}$$

gives

$$\begin{aligned} \lim_{(x,y)\to(x_0,y_0)} \frac{E(x,y)}{\sqrt{(x-x_0)^2+(y-y_0)^2}} &= \lim_{(x,y)\to(0,0)} \frac{\frac{xy}{\sqrt{x^2+y^2}}}{\sqrt{x^2+y^2}} \\ &= \lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2}. \end{aligned}$$

Depending on the path taken toward the origin, this limit takes different values. Therefore, the limit does not exist and the function f is not differentiable at the origin as shown in the following figure.

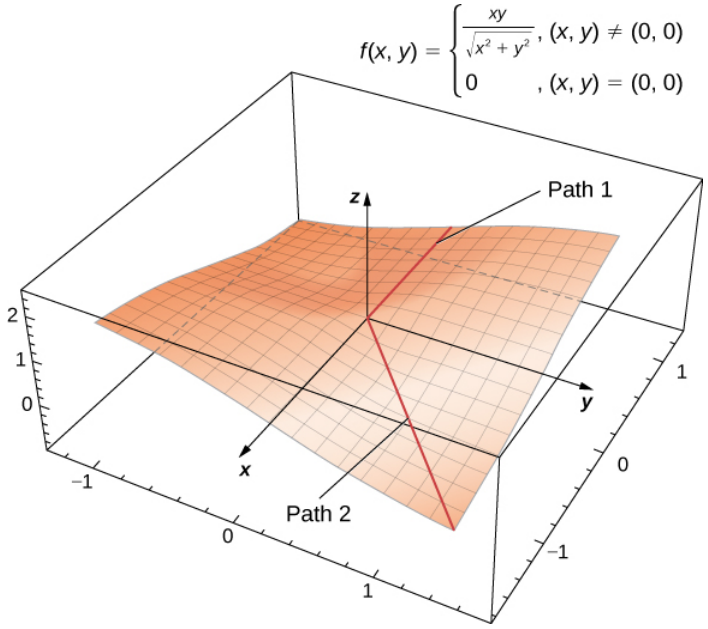


Figure 13.6.6: This function $f(x,y)$ (Equation 13.6.2) is not differentiable at the origin.

Differentiability and continuity for functions of two or more variables are connected, the same as for functions of one variable. In fact, with some adjustments of notation, the basic theorem is the same.

Theorem 13.6.1: Differentiability Implies Continuity

Let $z = f(x,y)$ be a function of two variables with (x_0,y_0) in the domain of f. If $f(x,y)$ is differentiable at (x_0,y_0), then $f(x,y)$ is continuous at (x_0,y_0).

Theorem 13.6.1 shows that if a function is differentiable at a point, then it is continuous there. However, if a function is continuous at a point, then it is not necessarily differentiable at that point. For example, the function discussed above (Equation 13.6.2)

$$f(x,y)=\begin{cases}\dfrac{xy}{\sqrt{x^2+y^2}}, & (x,y)\neq(0,0)\\ 0, & (x,y)=(0,0)\end{cases}$$

is *continuous* at the origin, but it is *not differentiable* at the origin. This observation is also similar to the situation in single-variable calculus.

We can further explore the connection between continuity and differentiability at a point. This next theorem says that if the function and its partial derivatives are continuous at a point, the function is differentiable.

Theorem 13.6.2: Continuity of First Partials Implies Differentiability

Let $z=f(x,y)$ be a function of two variables with (x_0,y_0) in the domain of f. If $f(x,y)$, $f_x(x,y)$, and $f_y(x,y)$ all exist in a neighborhood of (x_0,y_0) and are continuous at (x_0,y_0), then $f(x,y)$ is differentiable there.

Recall that earlier we showed that the function in Equation 13.6.2 was not differentiable at the origin. Let's calculate the partial derivatives f_x and f_y:

$$\frac{\partial f}{\partial x}=\frac{y^3}{(x^2+y^2)^{3/2}}$$

and

$$\frac{\partial f}{\partial y}=\frac{x^3}{(x^2+y^2)^{3/2}}.$$

The contrapositive of the preceding theorem states that if a function is not differentiable, then at least one of the hypotheses must be false. Let's explore the condition that $f_x(0,0)$ must be continuous. For this to be true, it must be true that

$$\lim_{(x,y)\to(0,0)} f_x(x,y)=f_x(0,0)$$

therefor

$$\lim_{(x,y)\to(0,0)} f_x(x,y)=\lim_{(x,y)\to(0,0)}\frac{y^3}{(x^2+y^2)^{3/2}}.$$

Let $x=ky$. Then

$$\begin{aligned}\lim_{(x,y)\to(0,0)}\frac{y^3}{(x^2+y^2)^{3/2}} &=\lim_{y\to0}\frac{y^3}{((ky)^2+y^2)^{3/2}}\\ &=\lim_{y\to0}\frac{y^3}{(k^2y^2+y^2)^{3/2}}\\ &=\lim_{y\to0}\frac{y^3}{|y|^3(k^2+1)^{3/2}}\\ &=\frac{1}{(k^2+1)^{3/2}}\lim_{y\to0}\frac{|y|}{y}.\end{aligned}$$

If $y>0$, then this expression equals $1/(k^2+1)^{3/2}$; if $y<0$, then it equals $-(1/(k^2+1)^{3/2})$. In either case, the value depends on k, so the limit fails to exist.

Differentials

In Linear Approximations and Differentials we first studied the concept of differentials. The differential of y, written dy, is defined as $f'(x)dx$. The differential is used to approximate $\Delta y=f(x+\Delta x)-f(x)$, where $\Delta x=dx$. Extending this idea to the linear approximation of a function of two variables at the point (x_0,y_0) yields the formula for the total differential for a function of two variables.

Definition: Total Differential

Let $z = f(x, y)$ be a function of two variables with (x_0, y_0) in the domain of f, and let Δx and Δy be chosen so that $(x_0 + \Delta x, y_0 + \Delta y)$ is also in the domain of f. If f is differentiable at the point (x_0, y_0), then the differentials dx and dy are defined as

$$dx = \Delta x$$

and

$$dy = \Delta y.$$

The differential dz, also called the ***total differential*** of $z = f(x, y)$ at (x_0, y_0), is defined as

$$dz = f_x(x_0, y_0)dx + f_y(x_0, y_0)dy. \tag{13.6.6}$$

Notice that the symbol ∂ is not used to denote the total differential; rather, d appears in front of z. Now, let's define $\Delta z = f(x + \Delta x, y + \Delta y) - f(x, y)$. We use dz to approximate Δz, so

$$\Delta z \approx dz = f_x(x_0, y_0)dx + f_y(x_0, y_0)dy.$$

Therefore, the differential is used to approximate the change in the function $z = f(x_0, y_0)$ at the point (x_0, y_0) for given values of Δx and Δy. Since $\Delta z = f(x + \Delta x, y + \Delta y) - f(x, y)$, this can be used further to approximate $f(x + \Delta x, y + \Delta y)$:

$$f(x + \Delta x, y + \Delta y) = f(x, y) + \Delta z \approx f(x, y) + f_x(x_0, y_0)\Delta x + f_y(x_0, y_0)\Delta y.$$

See the following figure.

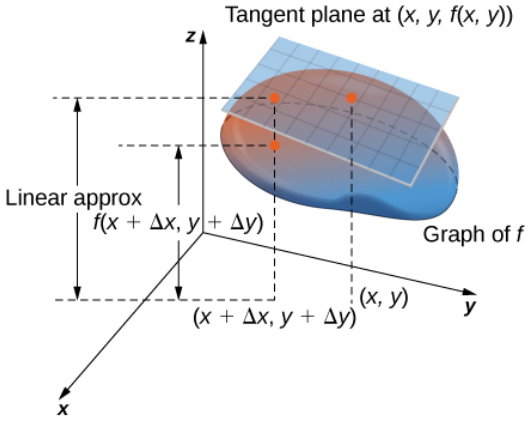


Figure 13.6.7: The linear approximation is calculated via the formula $f(x + \Delta x, y + \Delta y) \approx f(x, y) + f_x(x_0, y_0)\Delta x + f_y(x_0, y_0)\Delta y$.

One such application of this idea is to determine error propagation. For example, if we are manufacturing a gadget and are off by a certain amount in measuring a given quantity, the differential can be used to estimate the error in the total volume of the gadget.

Example 13.6.5: Approximation by Differentials

Find the differential dz of the function $f(x, y) = 3x^2 - 2xy + y^2$ and use it to approximate Δz at point $(2, -3)$. Use $\Delta x = 0.1$ and $\Delta y = -0.05$. What is the exact value of Δz?

Solution

First, we must calculate $f(x_0, y_0)$, $f_x(x_0, y_0)$, and $f_y(x_0, y_0)$ using $x_0 = 2$ and $y_0 = -3$:

$$f(x_0,y_0) = f(2,-3) = 3(2)^2 - 2(2)(-3) + (-3)^2 = 12+12+9 = 33$$
$$f_x(x,y) = 6x - 2y$$
$$f_y(x,y) = -2x + 2y$$
$$f_x(x_0,y_0) = fx(2,-3) = 6(2) - 2(-3) = 12+6 = 18$$
$$f_y(x_0,y_0) = f_y(2,-3) = -2(2) + 2(-3) = -4-6 = -10.$$

Then, we substitute these quantities into Equation 13.6.6:

$$dz = f_x(x_0,y_0)dx + f_y(x_0,y_0)dy$$
$$dz = 18(0.1) - 10(-0.05) = 1.8 + 0.5 = 2.3.$$

This is the approximation to $\Delta z = f(x_0+\Delta x, y_0+\Delta y) - f(x_0,y_0)$. The exact value of Δz is given by

$$\begin{aligned}\Delta z &= f(x_0+\Delta x, y_0+\Delta y) - f(x_0,y_0)\\ &= f(2+0.1,-3-0.05) - f(2,-3)\\ &= f(2.1,-3.05) - f(2,-3)\\ &= 2.3425.\end{aligned}$$

? Exercise 13.6.4

Find the differential dz of the function $f(x,y) = 4y^2 + x^2y - 2xy$ and use it to approximate Δz at point $(1,-1)$. Use $\Delta x = 0.03$ and $\Delta y = -0.02$. What is the exact value of Δz?

Hint

First, calculate $f_x(x_0,y_0)$ and $f_y(x_0,y_0)$ using $x_0 = 1$ and $y_0 = -1$, then use Equation 13.6.6

Answer

$dz = 0.18$

$\Delta z = f(1.03,-1.02) - f(1,-1) = 0.180682$

Differentiability of a Function of Three Variables

All of the preceding results for differentiability of functions of two variables can be generalized to functions of three variables. First, the definition:

Definition: Differentiability at a Point

A function $f(x,y,z)$ is differentiable at a point $P(x_0,y_0,z_0)$ if for all points (x,y,z) in a δ disk around P we can write

$$f(x,y) = f(x_0,y_0,z_0) + f_x(x_0,y_0,z_0)(x-x_0) + f_y(x_0,y_0,z_0)(y-y_0) + f_z(x_0,y_0,z_0)(z-z_0) + E(x,y,z),$$

where the error term E satisfies

$$\lim_{(x,y,z)\to(x_0,y_0,z_0)} \frac{E(x,y,z)}{\sqrt{(x-x_0)^2+(y-y_0)^2+(z-z_0)^2}} = 0.$$

If a function of three variables is differentiable at a point (x_0,y_0,z_0), then it is continuous there. Furthermore, continuity of first partial derivatives at that point guarantees differentiability.

Key Concepts

- The analog of a tangent line to a curve is a tangent plane to a surface for functions of two variables.
- Tangent planes can be used to approximate values of functions near known values.
- A function is differentiable at a point if it is ”smooth” at that point (i.e., no corners or discontinuities exist at that point).
- The total differential can be used to approximate the change in a function $z = f(x_0, y_0)$ at the point (x_0, y_0) for given values of Δx and Δy.

Key Equations

- **Tangent plane**

  $$z = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0)$$

- **Linear approximation**

  $$L(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0)$$

- **Total differential**

  $$dz = f_x(x_0, y_0)dx + f_y(x_0, y_0)dy\ .$$

- **Differentiability (two variables)**

  $$f(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) + E(x, y),$$

  where the error term E satisfies

  $$\lim_{(x,y)\to(x_0,y_0)} \frac{E(x, y)}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} = 0\,.$$

- **Differentiability (three variables)**

  $$f(x, y) = f(x_0, y_0, z_0) + f_x(x_0, y_0, z_0)(x - x_0) + f_y(x_0, y_0, z_0)(y - y_0) + f_z(x_0, y_0, z_0)(z - z_0) + E(x, y, z),$$

  where the error term E satisfies

  $$\lim_{(x,y,z)\to(x_0,y_0,z_0)} \frac{E(x, y, z)}{\sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}} = 0\,.$$

Glossary

differentiable

a function $f(x, y)$ is differentiable at (x_0, y_0) if $f(x, y)$ can be expressed in the form $f(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) + E(x, y)$,

where the error term $E(x, y)$ satisfies $\lim_{(x,y)\to(x_0,y_0)} \dfrac{E(x, y)}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} = 0$

linear approximation

given a function $f(x, y)$ and a tangent plane to the function at a point (x_0, y_0), we can approximate $f(x, y)$ for points near (x_0, y_0) using the tangent plane formula

tangent plane

given a function $f(x, y)$ that is differentiable at a point (x_0, y_0), the equation of the tangent plane to the surface $z = f(x, y)$ is given by $z = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0)$

total differential

the total differential of the function $f(x, y)$ at (x_0, y_0) is given by the formula $dz = f_x(x_0, y_0)dx + fy(x_0, y_0)dy$

13.6: Tangent Planes and Differentials is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **14.4: Tangent Planes and Linear Approximations** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

13.7: Extreme Values and Saddle Points

Learning Objectives

- Use partial derivatives to locate critical points for a function of two variables.
- Apply a second derivative test to identify a critical point as a local maximum, local minimum, or saddle point for a function of two variables.
- Examine critical points and boundary points to find absolute maximum and minimum values for a function of two variables.

One of the most useful applications for derivatives of a function of one variable is the determination of maximum and/or minimum values. This application is also important for functions of two or more variables, but as we have seen in earlier sections of this chapter, the introduction of more independent variables leads to more possible outcomes for the calculations. The main ideas of finding critical points and using derivative tests are still valid, but new wrinkles appear when assessing the results.

Critical Points

For functions of a single variable, we defined critical points as the values of the variable at which the function's derivative equals zero or does not exist. For functions of two or more variables, the concept is essentially the same, except for the fact that we are now working with partial derivatives.

Definition: Critical Points

Let $z = f(x, y)$ be a function of two variables that is differentiable on an open set containing the point (x_0, y_0). The point (x_0, y_0) is called a *critical point* of a function of two variables f if one of the two following conditions holds:

1. $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$
2. Either $f_x(x_0, y_0)$ or $f_y(x_0, y_0)$ does not exist.

Example 13.7.1: Finding Critical Points

Find the critical points of each of the following functions:

a. $f(x, y) = \sqrt{4y^2 - 9x^2 + 24y + 36x + 36}$
b. $g(x, y) = x^2 + 2xy - 4y^2 + 4x - 6y + 4$

Solution

a. First, we calculate $f_x(x, y)$ and $f_y(x, y)$:

$$\begin{aligned} f_x(x, y) &= \frac{1}{2}(-18x + 36)(4y^2 - 9x^2 + 24y + 36x + 36)^{-1/2} \\ &= \frac{-9x + 18}{\sqrt{4y^2 - 9x^2 + 24y + 36x + 36}} \\ f_y(x, y) &= \frac{1}{2}(8y + 24)(4y^2 - 9x^2 + 24y + 36x + 36)^{-1/2} \\ &= \frac{4y + 12}{\sqrt{4y^2 - 9x^2 + 24y + 36x + 36}} \end{aligned}.$$

Next, we set each of these expressions equal to zero:

$$\begin{aligned} \frac{-9x + 18}{\sqrt{4y^2 - 9x^2 + 24y + 36x + 36}} &= 0 \\ \frac{4y + 12}{\sqrt{4y^2 - 9x^2 + 24y + 36x + 36}} &= 0. \end{aligned}$$

Then, multiply each equation by its common denominator:

$$-9x+18 = 0$$
$$4y+12 = 0.$$

Therefore, $x=2$ and $y=-3$, so $(2,-3)$ is a critical point of f.

We must also check for the possibility that the denominator of each partial derivative can equal zero, thus causing the partial derivative not to exist. Since the denominator is the same in each partial derivative, we need only do this once:

$$4y^2-9x^2+24y+36x+36=0. \tag{13.7.1}$$

Equation 13.7.1 represents a hyperbola. We should also note that the domain of f consists of points satisfying the inequality

$$4y^2-9x^2+24y+36x+36\geq 0.$$

Therefore, any points on the hyperbola are not only critical points, they are also on the boundary of the domain. To put the hyperbola in standard form, we use the method of completing the square:

$$\begin{aligned} 4y^2-9x^2+24y+36x+36 &= 0 \\ 4y^2-9x^2+24y+36x &= -36 \\ 4y^2+24y-9x^2+36x &= -36 \\ 4(y^2+6y)-9(x^2-4x) &= -36 \\ 4(y^2+6y+9)-9(x^2-4x+4) &= -36-36+36 \\ 4(y+3)^2-9(x-2)^2 &= -36. \end{aligned}$$

Dividing both sides by -36 puts the equation in standard form:

$$\begin{aligned} \frac{4(y+3)^2}{-36}-\frac{9(x-2)^2}{-36} &= 1 \\ \frac{(x-2)^2}{4}-\frac{(y+3)^2}{9} &= 1. \end{aligned}$$

Notice that point $(2,-3)$ is the center of the hyperbola.

Thus, the critical points of the function f are $(2,-3)$ and all points on the hyperbola, $\frac{(x-2)^2}{4}-\frac{(y+3)^2}{9}=1$.

b. First, we calculate $g_x(x,y)$ and $g_y(x,y)$:

$$\begin{aligned} g_x(x,y) &= 2x+2y+4 \\ g_y(x,y) &= 2x-8y-6. \end{aligned}$$

Next, we set each of these expressions equal to zero, which gives a system of equations in x and y:

$$\begin{aligned} 2x+2y+4 &= 0 \\ 2x-8y-6 &= 0. \end{aligned}$$

Subtracting the second equation from the first gives $10y+10=0$, so $y=-1$. Substituting this into the first equation gives $2x+2(-1)+4=0$, so $x=-1$.

Therefore $(-1,-1)$ is a critical point of g. There are no points in $\mathbb{R}^2$ that make either partial derivative not exist.

Figure 13.7.1 shows the behavior of the surface at the critical point.

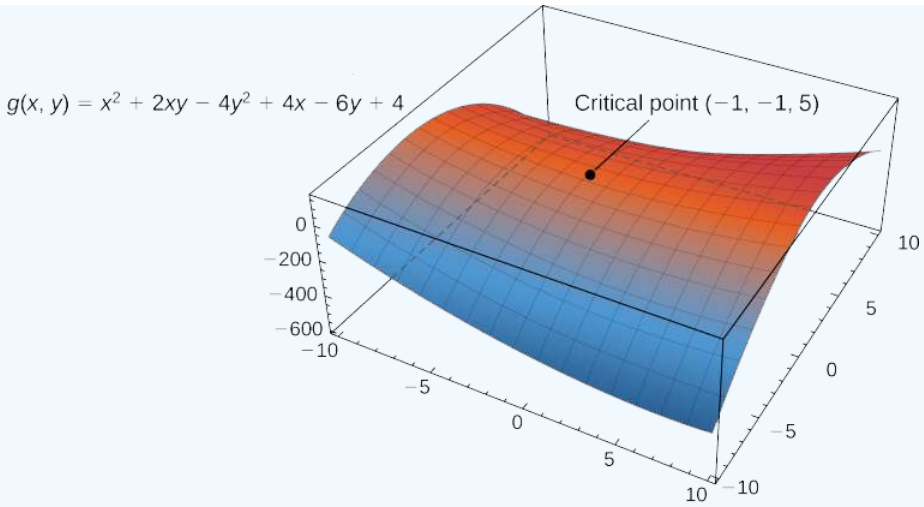


Figure 13.7.1: The function $g(x,y)$ has a critical point at $(-1,-1,5)$.

? Exercise 13.7.1

Find the critical point of the function $f(x,y)=x^3+2xy-2x-4y$.

Hint

Calculate $f_x(x,y)$ and $f_y(x,y)$, then set them equal to zero.

Answer

The only critical point of f is $(2,-5)$.

The main purpose for determining critical points is to locate relative maxima and minima, as in single-variable calculus. When working with a function of one variable, the definition of a local extremum involves finding an interval around the critical point such that the function value is either greater than or less than all the other function values in that interval. When working with a function of two or more variables, we work with an open disk around the point.

Definition: Global and Local Extrema

Let $z=f(x,y)$ be a function of two variables that is defined and continuous on an open set containing the point (x_0,y_0). Then f has a **local maximum** at (x_0,y_0) if

$$f(x_0,y_0)\geq f(x,y)$$

for all points (x,y) within some disk centered at (x_0,y_0). The number $f(x_0,y_0)$ is called a local maximum value. If the preceding inequality holds for every point (x,y) in the domain of f, then f has a **global maximum** (also called an absolute maximum) at (x_0,y_0).

The function f has a local minimum at (x_0,y_0) if

$$f(x_0,y_0)\leq f(x,y)$$

for all points (x,y) within some disk centered at (x_0,y_0). The number $f(x_0,y_0)$ is called a local minimum value. If the preceding inequality holds for every point (x,y) in the domain of f, then f has a **global minimum** (also called an absolute minimum) at (x_0,y_0).

If $f(x_0,y_0)$ is either a local maximum or local minimum value, then it is called a **local extremum** (see the following figure).

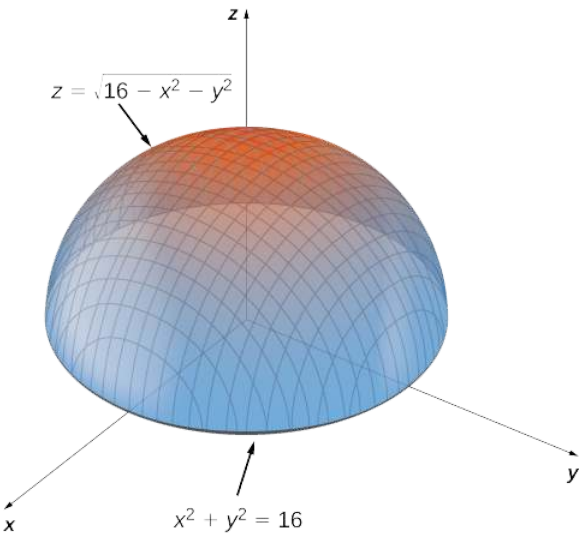


Figure 13.7.2: The graph of $z = \sqrt{16 - x^2 - y^2}$ has a maximum value when $(x, y) = (0, 0)$. It attains its minimum value at the boundary of its domain, which is the circle $x^2 + y^2 = 16$.

In Calculus 1, we showed that extrema of functions of one variable occur at critical points. The same is true for functions of more than one variable, as stated in the following theorem.

Fermat's Theorem for Functions of Two Variables

Let $z = f(x, y)$ be a function of two variables that is defined and continuous on an open set containing the point (x_0, y_0). Suppose f_x and f_y each exist at (x_0, y_0). If f has a local extremum at (x_0, y_0), then (x_0, y_0) is a critical point of f.

Second Derivative Test

Consider the function $f(x) = x^3$. This function has a critical point at $x = 0$, since $f'(0) = 3(0)^2 = 0$. However, f does not have an extreme value at $x = 0$. Therefore, the existence of a critical value at $x = x_0$ does not guarantee a local extremum at $x = x_0$. The same is true for a function of two or more variables. One way this can happen is at a **saddle point**. An example of a saddle point appears in the following figure.

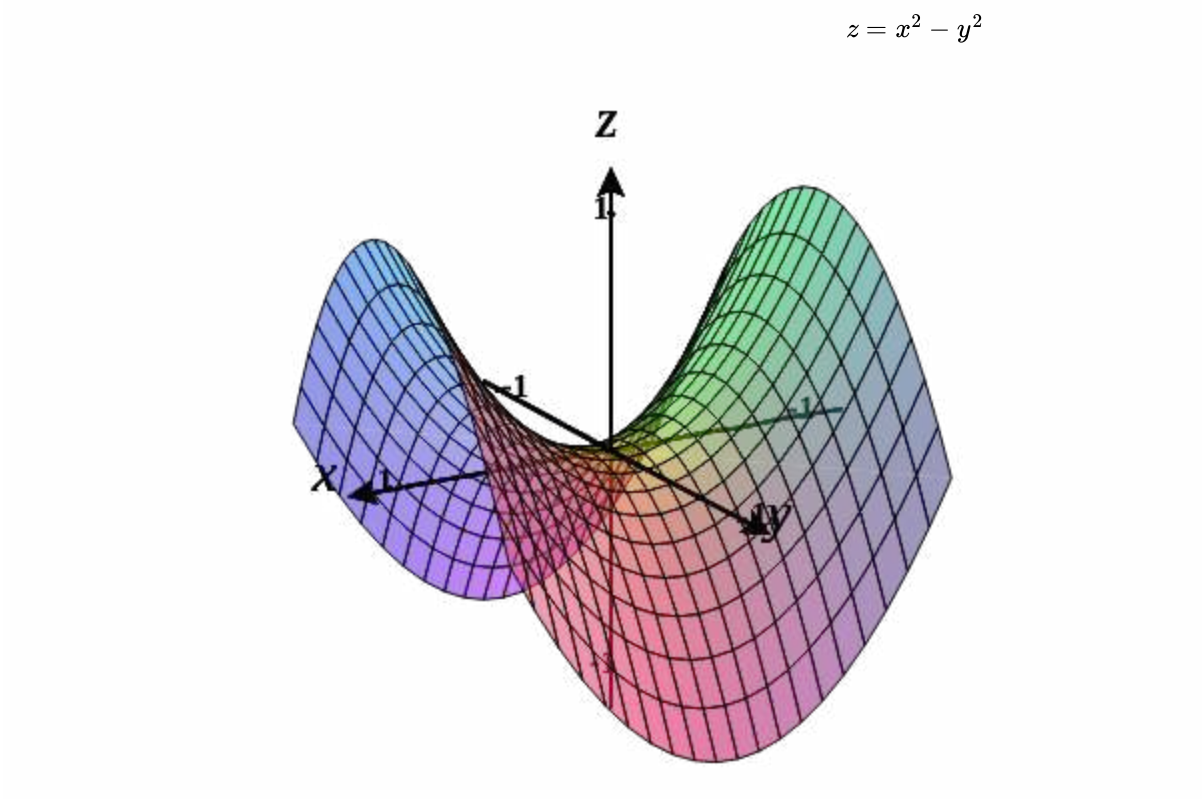


Figure 13.7.3: Graph of the function $z = x^2 - y^2$. This graph has a saddle point at the origin.

In this graph, the origin is a saddle point. This is because the first partial derivatives of f$(x, y) = x^2 - y^2$ are both equal to zero at this point, but it is neither a maximum nor a minimum for the function. Furthermore the vertical trace corresponding to $y = 0$ is $z = x^2$ (a parabola opening upward), but the vertical trace corresponding to $x = 0$ is $z = -y^2$ (a parabola opening downward). Therefore, it is both a global maximum for one trace and a global minimum for another.

Definition: Saddle Point

Given the function $z = f(x, y)$, the point $\big(x_0, y_0, f(x_0, y_0)\big)$ is a saddle point if both $f_x(x_0, y_0) = 0$ and $f_y(x_0, y_0) = 0$, but f does not have a local extremum at (x_0, y_0).

The second derivative test for a function of one variable provides a method for determining whether an extremum occurs at a critical point of a function. When extending this result to a function of two variables, an issue arises related to the fact that there are, in fact, four different second-order partial derivatives, although equality of mixed partials reduces this to three. The second derivative test for a function of two variables, stated in the following theorem, uses a **discriminant** D that replaces $f''(x_0)$ in the second derivative test for a function of one variable.

Second Derivative Test

Let $z = f(x, y)$ be a function of two variables for which the first- and second-order partial derivatives are continuous on some disk containing the point (x_0, y_0). Suppose $f_x(x_0, y_0) = 0$ and $f_y(x_0, y_0) = 0$. Define the quantity

$$D = f_{xx}(x_0, y_0) f_{yy}(x_0, y_0) - \big(f_{xy}(x_0, y_0)\big)^2.$$

Then:

i. If $D > 0$ and $f_{xx}(x_0, y_0) > 0$, then f has a local minimum at (x_0, y_0).
ii. If $D > 0$ and $f_{xx}(x_0, y_0) < 0$, then f has a local maximum at (x_0, y_0).
iii. If $D < 0$, then f has a saddle point at (x_0, y_0).
iv. If $D = 0$, then the test is inconclusive.

See Figure 13.7.4

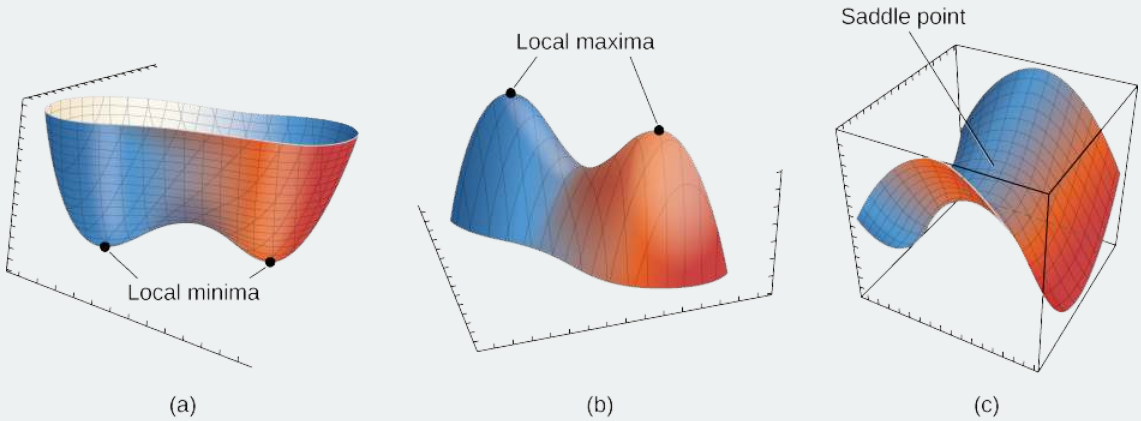


Figure 13.7.4: The second derivative test can often determine whether a function of two variables has local minima (a), local maxima (b), or a saddle point (c).

To apply the second derivative test, it is necessary that we first find the critical points of the function. There are several steps involved in the entire procedure, which are outlined in a problem-solving strategy.

Problem-Solving Strategy: Using the Second Derivative Test for Functions of Two Variables

Let $z = f(x, y)$ be a function of two variables for which the first- and second-order partial derivatives are continuous on some disk containing the point (x_0, y_0). To apply the second derivative test to find local extrema, use the following steps:

1. Determine the critical points (x_0, y_0) of the function f where $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$. Discard any points where at least one of the partial derivatives does not exist.
2. Calculate the discriminant $D = f_{xx}(x_0, y_0) f_{yy}(x_0, y_0) - \big(f_{xy}(x_0, y_0)\big)^2$ for each critical point of f.
3. Apply the four cases of the test to determine whether each critical point is a local maximum, local minimum, or saddle point, or whether the theorem is inconclusive.

Example 13.7.2: Using the Second Derivative Test

Find the critical points for each of the following functions, and use the second derivative test to find the local extrema:

a. $f(x, y) = 4x^2 + 9y^2 + 8x - 36y + 24$

b. $g(x, y) = \frac{1}{3}x^3 + y^2 + 2xy - 6x - 3y + 4$

Solution

a. **Step 1** of the problem-solving strategy involves finding the critical points of f. To do this, we first calculate $f_x(x, y)$ and $f_y(x, y)$, then set each of them equal to zero:

$$\begin{aligned} f_x(x, y) &= 8x + 8 \\ f_y(x, y) &= 18y - 36. \end{aligned}$$

Setting them equal to zero yields the system of equations

$$\begin{aligned} 8x + 8 &= 0 \\ 18y - 36 &= 0. \end{aligned}$$

The solution to this system is $x = -1$ and $y = 2$. Therefore $(-1, 2)$ is a critical point of f.

Step 2 of the problem-solving strategy involves calculating D. To do this, we first calculate the second partial derivatives of f:

$$f_{xx}(x,y) = 8$$
$$f_{xy}(x,y) = 0$$
$$f_{yy}(x,y) = 18.$$

Therefore, $D = f_{xx}(-1,2)f_{yy}(-1,2) - \left(f_{xy}(-1,2)\right)^2 = (8)(18) - (0)^2 = 144.$

Step 3 states to apply the four cases of the test to classify the function's behavior at this critical point.

Since $D > 0$ and $f_{xx}(-1,2) > 0$, this corresponds to case 1. Therefore, f has a local minimum at $(-1,2)$ as shown in the following figure.

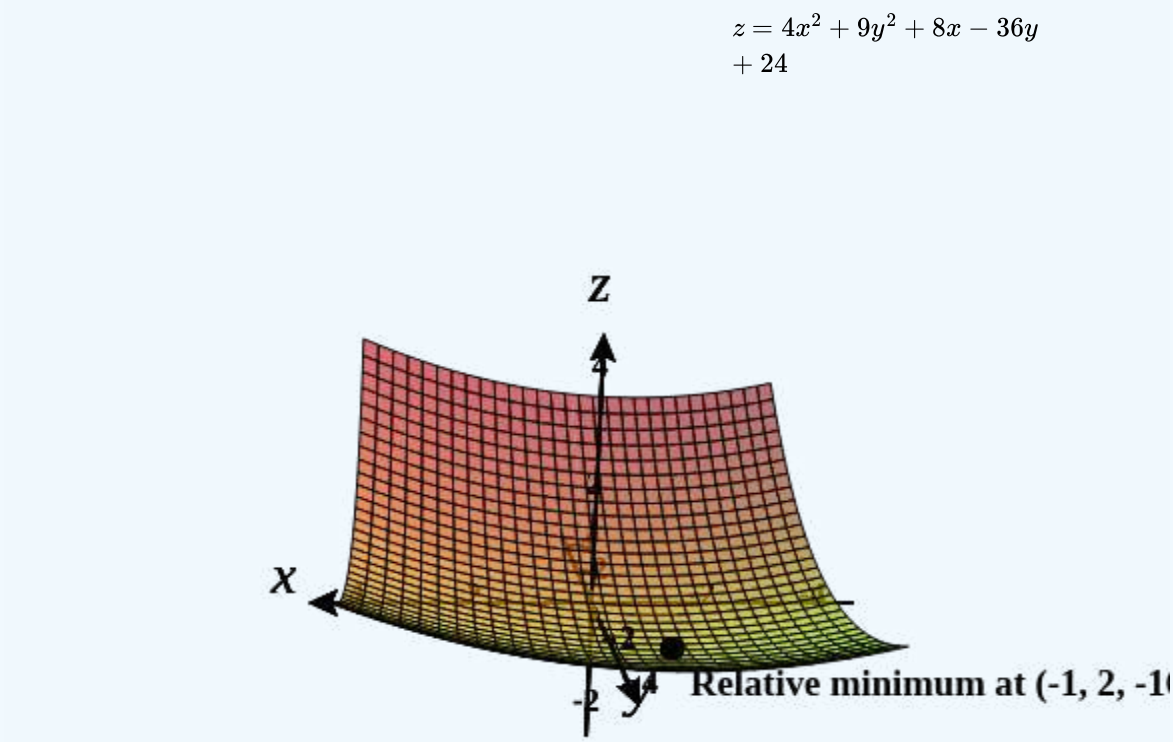


Figure 13.7.5 : The function $f(x,y)$ has a local minimum at $(-1,2,-16)$. Note the scale on the y-axis in this plot is in thousands.

b. For **step 1**, we first calculate $g_x(x,y)$ and $g_y(x,y)$, then set each of them equal to zero:

$$g_x(x,y) = x^2 + 2y - 6$$
$$g_y(x,y) = 2y + 2x - 3.$$

Setting them equal to zero yields the system of equations

$$x^2 + 2y - 6 = 0$$
$$2y + 2x - 3 = 0.$$

To solve this system, first solve the second equation for y. This gives $y = \dfrac{3-2x}{2}$. Substituting this into the first equation gives

$$x^2 + 3 - 2x - 6 = 0$$
$$x^2 - 2x - 3 = 0$$
$$(x-3)(x+1) = 0.$$

Therefore, $x=-1$ or $x=3$. Substituting these values into the equation $y=\dfrac{3-2x}{2}$ yields the critical points $\left(-1,\frac{5}{2}\right)$ and $\left(3,-\frac{3}{2}\right)$.

Step 2 involves calculating the second partial derivatives of g:

$$\begin{aligned} g_{xx}(x,y) &= 2x \\ g_{xy}(x,y) &= 2 \\ g_{yy}(x,y) &= 2. \end{aligned}$$

Then, we find a general formula for D:

$$\begin{aligned} D(x_0,y_0) &= g_{xx}(x_0,y_0)g_{yy}(x_0,y_0) - \left(g_{xy}(x_0,y_0)\right)^2 \\ &= (2x_0)(2) - 2^2 \\ &= 4x_0 - 4. \end{aligned}$$

Next, we substitute each critical point into this formula:

$$\begin{aligned} D\left(-1,\tfrac{5}{2}\right) &= (2(-1))(2) - (2)^2 = -4-4 = -8 \\ D\left(3,-\tfrac{3}{2}\right) &= (2(3))(2) - (2)^2 = 12-4 = 8. \end{aligned}$$

In step 3, we note that, applying Note to point $\left(-1,\frac{5}{2}\right)$ leads to case 3, which means that $\left(-1,\frac{5}{2}\right)$ is a saddle point. Applying the theorem to point $\left(3,-\frac{3}{2}\right)$ leads to case 1, which means that $\left(3,-\frac{3}{2}\right)$ corresponds to a local minimum as shown in the following figure.

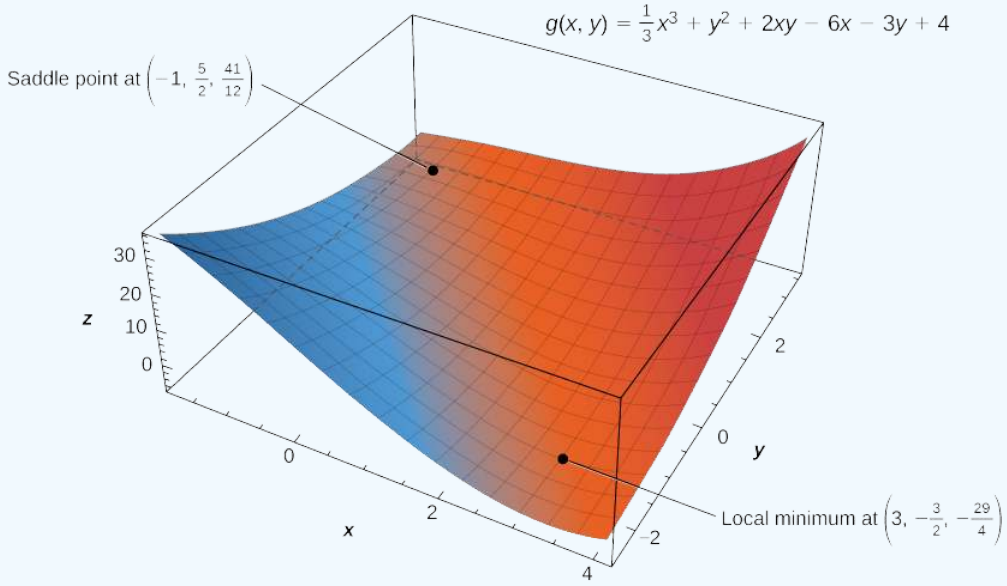


Figure 13.7.6: The function $g(x,y)$ has a local minimum and a saddle point.

? Exercise 13.7.2

Use the second derivative test to find the local extrema of the function

$$f(x,y) = x^3 + 2xy - 6x - 4y^2.$$

Hint

Follow the problem-solving strategy for applying the second derivative test.

Answer

$\left(\frac{4}{3}, \frac{1}{3}\right)$ is a saddle point, $\left(-\frac{3}{2}, -\frac{3}{8}\right)$ is a local maximum.

Absolute Maxima and Minima

When finding global extrema of functions of one variable on a closed interval, we start by checking the critical values over that interval and then evaluate the function at the endpoints of the interval. When working with a function of two variables, the closed interval is replaced by a closed, bounded set. A set is bounded if all the points in that set can be contained within a ball (or disk) of finite radius. First, we need to find the critical points inside the set and calculate the corresponding critical values. Then, it is necessary to find the maximum and minimum value of the function on the boundary of the set. When we have all these values, the largest function value corresponds to the global maximum and the smallest function value corresponds to the absolute minimum. First, however, we need to be assured that such values exist. The following theorem does this.

Extreme Value Theorem

A continuous function $f(x, y)$ on a closed and bounded set D in the plane attains an absolute maximum value at some point of D and an absolute minimum value at some point of D.

Now that we know any continuous function f defined on a closed, bounded set attains its extreme values, we need to know how to find them.

Finding Extreme Values of a Function of Two Variables

Assume $z = f(x, y)$ is a differentiable function of two variables defined on a closed, bounded set D. Then f will attain the absolute maximum value and the absolute minimum value, which are, respectively, the largest and smallest values found among the following:

1. The values of f at the critical points of f in D.
2. The values of f on the boundary of D.

The proof of this theorem is a direct consequence of the extreme value theorem and Fermat's theorem. In particular, if either extremum is not located on the boundary of D, then it is located at an interior point of D. But an interior point (x_0, y_0) of D that's an absolute extremum is also a local extremum; hence, (x_0, y_0) is a critical point of f by Fermat's theorem. Therefore the only possible values for the global extrema of f on D are the extreme values of f on the interior or boundary of D.

Problem-Solving Strategy: Finding Absolute Maximum and Minimum Values

Let $z = f(x, y)$ be a continuous function of two variables defined on a closed, bounded set D, and assume f is differentiable on D. To find the absolute maximum and minimum values of f on D, do the following:

1. Determine the critical points of f in D.
2. Calculate f at each of these critical points.
3. Determine the maximum and minimum values of f on the boundary of its domain.
4. The maximum and minimum values of f will occur at one of the values obtained in steps 2 and 3.

Finding the maximum and minimum values of f on the boundary of D can be challenging. If the boundary is a rectangle or set of straight lines, then it is possible to parameterize the line segments and determine the maxima on each of these segments, as seen in Example 13.7.3. The same approach can be used for other shapes such as circles and ellipses.

If the boundary of the set D is a more complicated curve defined by a function $g(x, y) = c$ for some constant c, and the first-order partial derivatives of g exist, then the method of Lagrange multipliers can prove useful for determining the extrema of f on the boundary which is introduced in Lagrange Multipliers.

Example 13.7.3: Finding Absolute Extrema

Use the problem-solving strategy for finding absolute extrema of a function to determine the absolute extrema of each of the following functions:

a. $f(x,y)=x^2-2xy+4y^2-4x-2y+24$ on the domain defined by $0\le x\le 4$ and $0\le y\le 2$
b. $g(x,y)=x^2+y^2+4x-6y$ on the domain defined by $x^2+y^2\le 16$

Solution

a. Using the problem-solving strategy, step 1 involves finding the critical points of f on its domain. Therefore, we first calculate $f_x(x,y)$ and $f_y(x,y)$, then set them each equal to zero:

$$\begin{aligned} f_x(x,y) &= 2x-2y-4 \\ f_y(x,y) &= -2x+8y-2. \end{aligned}$$

Setting them equal to zero yields the system of equations

$$\begin{aligned} 2x-2y-4 &= 0 \\ -2x+8y-2 &= 0. \end{aligned}$$

The solution to this system is $x=3$ and $y=1$. Therefore $(3,1)$ is a critical point of f. Calculating $f(3,1)$ gives $f(3,1)=17$.

The next step involves finding the extrema of f on the boundary of its domain. The boundary of its domain consists of four line segments as shown in the following graph:

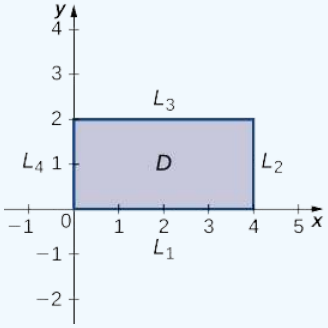


Figure 13.7.7: Graph of the domain of the function $f(x,y)=x^2-2xy+4y^2-4x-2y+24$.

L_1 is the line segment connecting $(0,0)$ and $(4,0)$, and it can be parameterized by the equations $x(t)=t, y(t)=0$ for $0\le t\le 4$. Define $g(t)=f\big(x(t),y(t)\big)$. This gives $g(t)=t^2-4t+24$. Differentiating g leads to $g'(t)=2t-4$. Therefore, g has a critical value at $t=2$, which corresponds to the point $(2,0)$. Calculating $f(2,0)$ gives the z-value 20.

L_2 is the line segment connecting $(4,0)$ and $(4,2)$, and it can be parameterized by the equations $x(t)=4, y(t)=t$ for $0\le t\le 2$. Again, define $g(t)=f\big(x(t),y(t)\big)$. This gives $g(t)=4t^2-10t+24$. Then, $g'(t)=8t-10$. g has a critical value at $t=\frac{5}{4}$, which corresponds to the point $\left(4,\frac{5}{4}\right)$. Calculating $f\left(4,\frac{5}{4}\right)$ gives the z-value 17.75.

L_3 is the line segment connecting $(0,2)$ and $(4,2)$, and it can be parameterized by the equations $x(t)=t, y(t)=2$ for $0\le t\le 4$. Again, define $g(t)=f\big(x(t),y(t)\big)$. This gives $g(t)=t^2-8t+36$. The critical value corresponds to the point $(4,2)$. So, calculating $f(4,2)$ gives the z-value 20.

L_4 is the line segment connecting $(0,0)$ and $(0,2)$, and it can be parameterized by the equations $x(t)=0, y(t)=t$ for $0\le t\le 2$. This time, $g(t)=4t^2-2t+24$ and the critical value $t=\frac{1}{4}$ correspond to the point $\left(0,\frac{1}{4}\right)$. Calculating $f\left(0,\frac{1}{4}\right)$ gives the z-value 23.75.

We also need to find the values of $f(x,y)$ at the corners of its domain. These corners are located at $(0,0),(4,0),(4,2)$ and $(0,2)$:

$$\begin{aligned} f(0,0) &= (0)^2 - 2(0)(0) + 4(0)^2 - 4(0) - 2(0) + 24 = 24 \\ f(4,0) &= (4)^2 - 2(4)(0) + 4(0)^2 - 4(4) - 2(0) + 24 = 24 \\ f(4,2) &= (4)^2 - 2(4)(2) + 4(2)^2 - 4(4) - 2(2) + 24 = 20 \\ f(0,2) &= (0)^2 - 2(0)(2) + 4(2)^2 - 4(0) - 2(2) + 24 = 36. \end{aligned}$$

The absolute maximum value is 36, which occurs at $(0,2)$, and the global minimum value is 20, which occurs at both $(4,2)$ and $(2,0)$ as shown in the following figure.

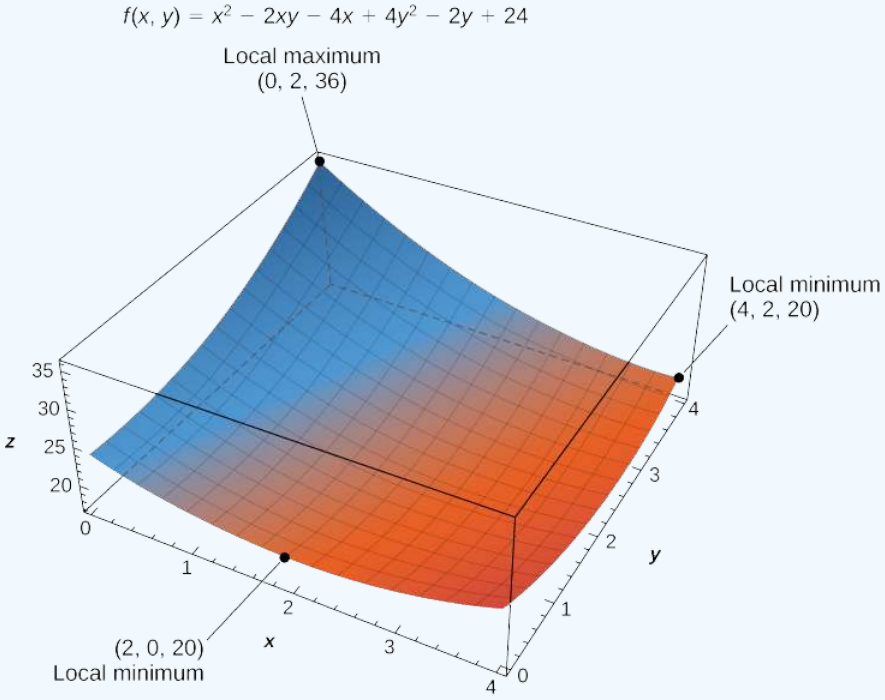


Figure 13.7.8: The function $f(x,y)$ has two global minima and one global maximum over its domain.

b. Using the problem-solving strategy, step 1 involves finding the critical points of g on its domain. Therefore, we first calculate $g_x(x,y)$ and $g_y(x,y)$, then set them each equal to zero:

$$\begin{aligned} g_x(x,y) &= 2x + 4 \\ g_y(x,y) &= 2y - 6. \end{aligned}$$

Setting them equal to zero yields the system of equations

$$\begin{aligned} 2x + 4 &= 0 \\ 2y - 6 &= 0. \end{aligned}$$

The solution to this system is $x = -2$ and $y = 3$. Therefore, $(-2,3)$ is a critical point of g. Calculating $g(-2,3)$, we get

$$g(-2,3) = (-2)^2 + 3^2 + 4(-2) - 6(3) = 4 + 9 - 8 - 18 = -13.$$

The next step involves finding the extrema of g on the boundary of its domain. The boundary of its domain consists of a circle of radius 4 centered at the origin as shown in the following graph.

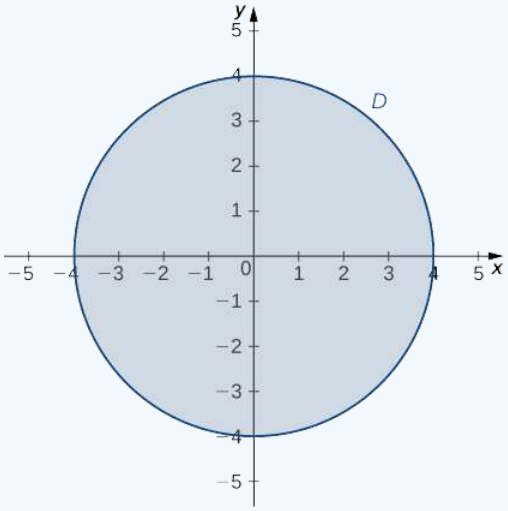


Figure 13.7.9: Graph of the restricted domain of the function $g(x,y)=x^2+y^2+4x-6y$.

The boundary of the domain of g can be parameterized using the functions $x(t)=4\cos t$, $y(t)=4\sin t$ for $0\le t\le 2\pi$. Define $h(t)=g\big(x(t),y(t)\big)$:

$$\begin{aligned} h(t) &= g\big(x(t),y(t)\big) \\ &= (4\cos t)^2+(4\sin t)^2+4(4\cos t)-6(4\sin t) \\ &= 16\cos^2 t+16\sin^2 t+16\cos t-24\sin t \\ &= 16+16\cos t-24\sin t. \end{aligned}$$

Setting $h'(t)=0$ leads to

$$\begin{aligned} -16\sin t-24\cos t &= 0 \\ -16\sin t &= 24\cos t \\ \frac{-16\sin t}{-16\cos t} &= \frac{24\cos t}{-16\cos t} \\ \tan t &= -\frac{3}{2}. \end{aligned}$$

This equation has two solutions over the interval $0\le t\le 2\pi$. One is $t=\pi-\arctan(\frac{3}{2})$ and the other is $t=2\pi-\arctan(\frac{3}{2})$. For the first angle,

$$\begin{aligned} \sin t &= \sin(\pi-\arctan(\tfrac{3}{2}))=\sin(\arctan(\tfrac{3}{2}))=\frac{3\sqrt{13}}{13} \\ \cos t &= \cos(\pi-\arctan(\tfrac{3}{2}))=-\cos(\arctan(\tfrac{3}{2}))=-\frac{2\sqrt{13}}{13}. \end{aligned}$$

Therefore, $x(t)=4\cos t=-\frac{8\sqrt{13}}{13}$ and $y(t)=4\sin t=\frac{12\sqrt{13}}{13}$, so $\left(-\frac{8\sqrt{13}}{13},\frac{12\sqrt{13}}{13}\right)$ is a critical point on the boundary and

$$\begin{aligned} g\left(-\tfrac{8\sqrt{13}}{13},\tfrac{12\sqrt{13}}{13}\right) &= \left(-\tfrac{8\sqrt{13}}{13}\right)^2+\left(\tfrac{12\sqrt{13}}{13}\right)^2+4\left(-\tfrac{8\sqrt{13}}{13}\right)-6\left(\tfrac{12\sqrt{13}}{13}\right) \\ &= \frac{144}{13}+\frac{64}{13}-\frac{32\sqrt{13}}{13}-\frac{72\sqrt{13}}{13} \\ &= \frac{208-104\sqrt{13}}{13}\approx -12.844. \end{aligned}$$

For the second angle,

$$\sin t = \sin(2\pi - \arctan(\tfrac{3}{2})) = -\sin(\arctan(\tfrac{3}{2})) = -\frac{3\sqrt{13}}{13}$$

$$\cos t = \cos(2\pi - \arctan(\tfrac{3}{2})) = \cos(\arctan(\tfrac{3}{2})) = \frac{2\sqrt{13}}{13}.$$

Therefore, $x(t) = 4\cos t = \frac{8\sqrt{13}}{13}$ and $y(t) = 4\sin t = -\frac{12\sqrt{13}}{13}$, so $\left(\frac{8\sqrt{13}}{13}, -\frac{12\sqrt{13}}{13}\right)$ is a critical point on the boundary and

$$\begin{aligned} g\left(\tfrac{8\sqrt{13}}{13}, -\tfrac{12\sqrt{13}}{13}\right) &= \left(\tfrac{8\sqrt{13}}{13}\right)^2 + \left(-\tfrac{12\sqrt{13}}{13}\right)^2 + 4\left(\tfrac{8\sqrt{13}}{13}\right) - 6\left(-\tfrac{12\sqrt{13}}{13}\right) \\ &= \frac{144}{13} + \frac{64}{13} + \frac{32\sqrt{13}}{13} + \frac{72\sqrt{13}}{13} \\ &= \frac{208 + 104\sqrt{13}}{13} \approx 44.844. \end{aligned}$$

The absolute minimum of g is -13, which is attained at the point $(-2, 3)$, which is an interior point of D. The absolute maximum of g is approximately equal to 44.844, which is attained at the boundary point $\left(\frac{8\sqrt{13}}{13}, -\frac{12\sqrt{13}}{13}\right)$. These are the absolute extrema of g on D as shown in the following figure.

$$f(x, y) = x^2 + 4x + y^2 - 6y$$

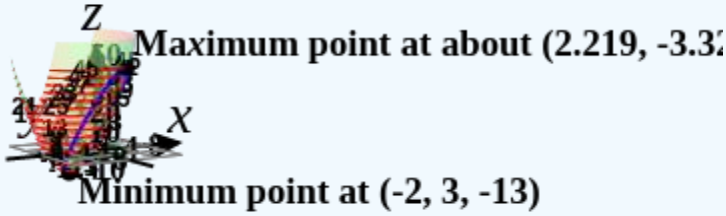


Figure 13.7.10: The function $f(x, y)$ has a local minimum and a local maximum.

Exercise 13.7.3:

Use the problem-solving strategy for finding absolute extrema of a function to find the absolute extrema of the function

$$f(x, y) = 4x^2 - 2xy + 6y^2 - 8x + 2y + 3$$

on the domain defined by $0 \le x \le 2$ and $-1 \le y \le 3$.

Hint

Calculate $f_x(x,y)$ and $f_y(x,y)$, and set them equal to zero. Then, calculate f for each critical point and find the extrema of f on the boundary of D.

Answer

The absolute minimum occurs at $(1,0) : f(1,0) = -1$.

The absolute maximum occurs at $(0,3) : f(0,3) = 63$.

Example 13.7.4: Profitable Golf Balls

Pro-T company has developed a profit model that depends on the number x of golf balls sold per month (measured in thousands), and the number of hours per month of advertising y, according to the function

$$z = f(x,y) = 48x + 96y - x^2 - 2xy - 9y^2,$$

where z is measured in thousands of dollars. The maximum number of golf balls that can be produced and sold is $50,000$, and the maximum number of hours of advertising that can be purchased is 25. Find the values of x and y that maximize profit, and find the maximum profit.



Figure 13.7.11: (credit: modification of work by oatsy40, Flickr)

Solution

Using the problem-solving strategy, step 1 involves finding the critical points of f on its domain. Therefore, we first calculate $f_x(x,y)$ and $f_y(x,y)$, then set them each equal to zero:

$$\begin{aligned} f_x(x,y) &= 48 - 2x - 2y \\ f_y(x,y) &= 96 - 2x - 18y. \end{aligned}$$

Setting them equal to zero yields the system of equations

$$\begin{aligned} 48 - 2x - 2y &= 0 \\ 96 - 2x - 18y &= 0. \end{aligned}$$

The solution to this system is $x = 21$ and $y = 3$. Therefore $(21,3)$ is a critical point of f. Calculating $f(21,3)$ gives $f(21,3) = 48(21) + 96(3) - 21^2 - 2(21)(3) - 9(3)^2 = 648$.

The domain of this function is $0 \le x \le 50$ and $0 \le y \le 25$ as shown in the following graph.

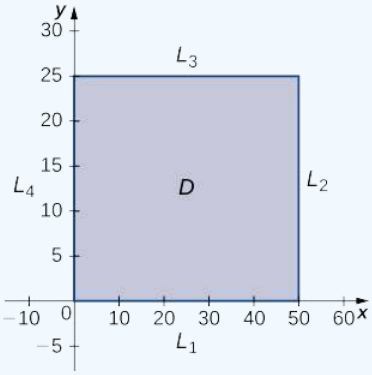


Figure 13.7.12: Graph of the domain of the function $f(x,y) = 48x + 96y - x^2 - 2xy - 9y^2$.

L_1 is the line segment connecting $(0,0)$ and $(50,0)$, and it can be parameterized by the equations $x(t) = t, y(t) = 0$ for $0 \le t \le 50$. We then define $g(t) = f(x(t), y(t))$:

$$\begin{aligned} g(t) &= f(x(t), y(t)) \\ &= f(t, 0) \\ &= 48t + 96(0) - y^2 - 2(t)(0) - 9(0)^2 \\ &= 48t - t^2. \end{aligned}$$

Setting $g'(t) = 0$ yields the critical point $t = 24$, which corresponds to the point $(24,0)$ in the domain of f. Calculating $f(24,0)$ gives 576.

L_2 is the line segment connecting $(50,0)$ and $(50,25)$, and it can be parameterized by the equations $x(t) = 50, y(t) = t$ for $0 \le t \le 25$. Once again, we define $g(t) = f\big(x(t), y(t)\big)$:

$$\begin{aligned} g(t) &= f\big(x(t), y(t)\big) \\ &= f(50, t) \\ &= 48(50) + 96t - 50^2 - 2(50)t - 9t^2 \\ &= -9t^2 - 4t - 100. \end{aligned}$$

This function has a critical point at $t = -\frac{2}{9}$, which corresponds to the point $(50, -29)$. This point is not in the domain of f.

L_3 is the line segment connecting $(0,25)$ and $(50,25)$, and it can be parameterized by the equations $x(t) = t, y(t) = 25$ for $0 \le t \le 50$. We define $g(t) = f\big(x(t), y(t)\big)$:

$$\begin{aligned} g(t) &= f\big(x(t), y(t)\big) \\ &= f(t, 25) \\ &= 48t + 96(25) - t^2 - 2t(25) - 9(25^2) \\ &= -t^2 - 2t - 3225. \end{aligned}$$

This function has a critical point at $t = -1$, which corresponds to the point $(-1, 25)$, which is not in the domain.

L_4 is the line segment connecting $(0,0)$ to $(0,25)$, and it can be parameterized by the equations $x(t) = 0, y(t) = t$ for $0 \le t \le 25$. We define $g(t) = f\big(x(t), y(t)\big)$:

$$\begin{aligned} g(t) &= f\big(x(t), y(t)\big) \\ &= f(0, t) \\ &= 48(0) + 96t - (0)^2 - 2(0)t - 9t^2 \\ &= 96t - 9t^2. \end{aligned}$$

This function has a critical point at $t = \frac{16}{3}$, which corresponds to the point $\left(0, \frac{16}{3}\right)$, which is on the boundary of the domain. Calculating $f\left(0, \frac{16}{3}\right)$ gives 256.

We also need to find the values of $f(x, y)$ at the corners of its domain. These corners are located at $(0, 0)$, $(50, 0)$, $(50, 25)$ and $(0, 25)$:

$$\begin{aligned} f(0,0) &= 48(0) + 96(0) - (0)^2 - 2(0)(0) - 9(0)^2 = 0 \\ f(50,0) &= 48(50) + 96(0) - (50)^2 - 2(50)(0) - 9(0)^2 = -100 \\ f(50,25) &= 48(50) + 96(25) - (50)^2 - 2(50)(25) - 9(25)^2 = -5825 \\ f(0,25) &= 48(0) + 96(25) - (0)^2 - 2(0)(25) - 9(25)^2 = -3225. \end{aligned}$$

The maximum value is 648, which occurs at $(21, 3)$. Therefore, a maximum profit of $\$648,000$ is realized when $21,000$ golf balls are sold and 3 hours of advertising are purchased per month as shown in the following figure.

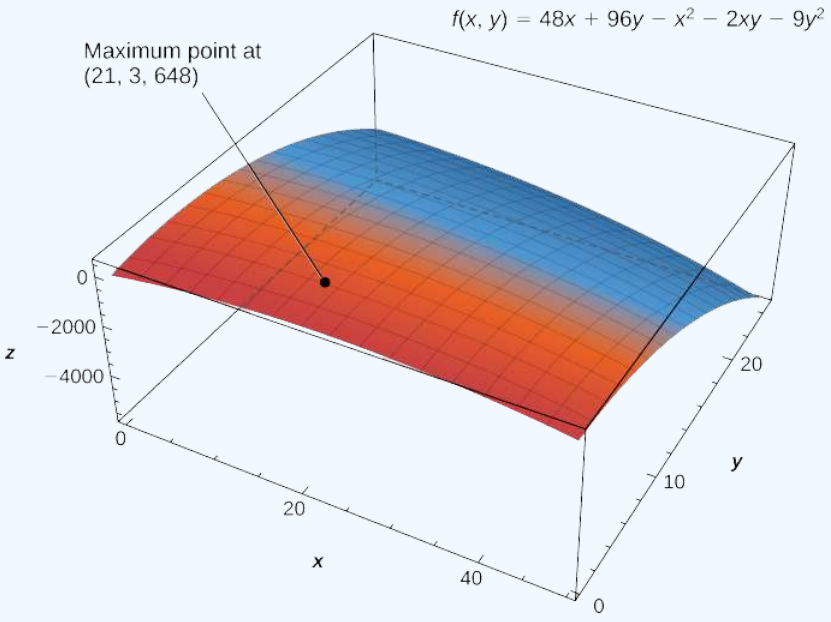


Figure 13.7.13: The profit function $f(x, y)$ has a maximum at $(21, 3, 648)$.

Key Concepts

- A critical point of the function $f(x, y)$ is any point (x_0, y_0) where either $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$, or at least one of $f_x(x_0, y_0)$ and $f_y(x_0, y_0)$ do not exist.
- A saddle point is a point (x_0, y_0) where $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$, but $f(x_0, y_0)$ is neither a maximum nor a minimum at that point.
- To find extrema of functions of two variables, first find the critical points, then calculate the discriminant and apply the second derivative test.

Key Equations

- **Discriminant**

$$D = f_{xx}(x_0, y_0) f_{yy}(x_0, y_0) - (f_{xy}(x_0, y_0))^2$$

Glossary

critical point of a function of two variables

the point (x_0, y_0) is called a critical point of $f(x, y)$ if one of the two following conditions holds:

1. $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$
2. At least one of $f_x(x_0, y_0)$ and $f_y(x_0, y_0)$ do not exist

discriminant

the discriminant of the function $f(x,y)$ is given by the formula $D = f_{xx}(x_0,y_0)f_{yy}(x_0,y_0) - (f_{xy}(x_0,y_0))^2$

saddle point

given the function $z = f(x,y)$, the point $(x_0, y_0, f(x_0,y_0))$ is a saddle point if both $f_x(x_0,y_0) = 0$ and $f_y(x_0,y_0) = 0$, but f does not have a local extremum at (x_0,y_0)

13.7: Extreme Values and Saddle Points is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **14.7: Maxima/Minima Problems** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

13.8: Lagrange Multipliers

Learning Objectives

- Use the method of Lagrange multipliers to solve optimization problems with one constraint.
- Use the method of Lagrange multipliers to solve optimization problems with two constraints.

Solving optimization problems for functions of two or more variables can be similar to solving such problems in single-variable calculus. However, techniques for dealing with multiple variables allow us to solve more varied optimization problems for which we need to deal with additional conditions or constraints. In this section, we examine one of the more common and useful methods for solving optimization problems with constraints.

Lagrange Multipliers

In the previous section, an applied situation was explored involving maximizing a profit function, subject to certain **constraints.** In that example, the constraints involved a maximum number of golf balls that could be produced and sold in 1 month (x), and a maximum number of advertising hours that could be purchased per month (y). Suppose these were combined into a single budgetary constraint, such as $20x+4y \leq 216$, that took into account both the cost of producing the golf balls and the number of advertising hours purchased per month. The goal is still to maximize profit, but now there is a different type of constraint on the values of x and y. This constraint and the corresponding profit function

$$f(x,y)=48x+96y-x^2-2xy-9y^2$$

is an example of an **optimization problem**, and the function $f(x,y)$ is called the **objective function**. A graph of various level curves of the function $f(x,y)$ follows.

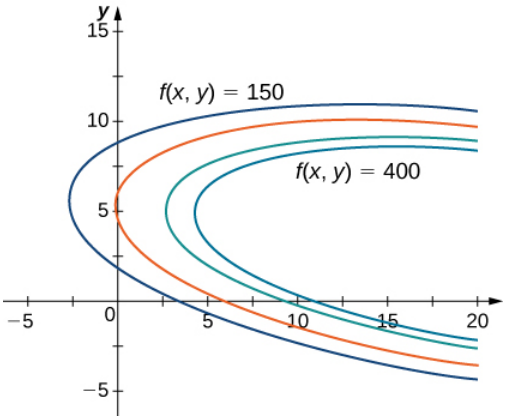


Figure 13.8.1: Graph showing level curves of the function $f(x,y)=48x+96y-x^2-2xy-9y^2$ corresponding to $c=150,250,350$, and 400.

In Figure 13.8.1, the value c represents different profit levels (i.e., values of the function f). As the value of c increases, the curve shifts to the right. Since our goal is to maximize profit, we want to choose a curve as far to the right as possible. If there were no restrictions on the number of golf balls the company could produce or the number of units of advertising available, then we could produce as many golf balls as we want, and advertise as much as we want, and there would be not be a maximum profit for the company. Unfortunately, we have a budgetary constraint that is modeled by the inequality $20x+4y \leq 216$. To see how this constraint interacts with the profit function, Figure 13.8.2 shows the graph of the line $20x+4y=216$ superimposed on the previous graph.

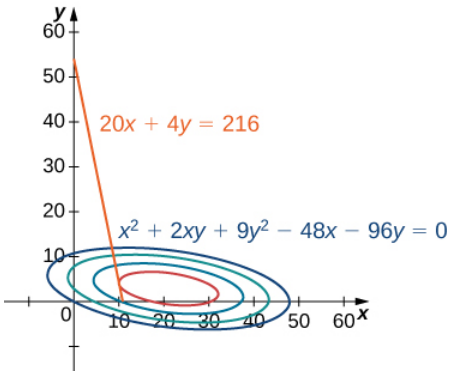


Figure 13.8.2: Graph of level curves of the function $f(x,y) = 48x + 96y - x^2 - 2xy - 9y^2$ corresponding to $c = 150, 250, 350,$ and 395. The red graph is the constraint function.

As mentioned previously, the maximum profit occurs when the level curve is as far to the right as possible. However, the level of production corresponding to this maximum profit must also satisfy the budgetary constraint, so the point at which this profit occurs must also lie on (or to the left of) the red line in Figure 13.8.2 Inspection of this graph reveals that this point exists where the line is tangent to the level curve of f. Trial and error reveals that this profit level seems to be around 395, when x and y are both just less than 5. We return to the solution of this problem later in this section. From a theoretical standpoint, at the point where the profit curve is tangent to the constraint line, the gradient of both of the functions evaluated at that point must point in the same (or opposite) direction. Recall that the gradient of a function of more than one variable is a vector. If two vectors point in the same (or opposite) directions, then one must be a constant multiple of the other. This idea is the basis of the **method of Lagrange multipliers.**

Theorem 13.8.1: Method of Lagrange Multipliers with One Constraint

Let f and g be functions of two variables with continuous partial derivatives at every point of some open set containing the smooth curve $g(x,y) = 0$. Suppose that f, when restricted to points on the curve $g(x,y) = 0$, has a local extremum at the point (x_0, y_0) and that $\vec{\nabla} g(x_0, y_0) \neq 0$. Then there is a number λ called a **Lagrange multiplier,** for which

$$\vec{\nabla} f(x_0, y_0) = \lambda \vec{\nabla} g(x_0, y_0).$$

Proof

Assume that a constrained extremum occurs at the point (x_0, y_0). Furthermore, we assume that the equation $g(x,y) = 0$ can be smoothly parameterized as

$$x = x(s) \text{ and } y = y(s)$$

where s is an arc length parameter with reference point (x_0, y_0) at $s = 0$. Therefore, the quantity $z = f(x(s), y(s))$ has a relative maximum or relative minimum at $s = 0$, and this implies that $\dfrac{dz}{ds} = 0$ at that point. From the chain rule,

$$\begin{aligned} \frac{dz}{ds} &= \frac{\partial f}{\partial x} \cdot \frac{\partial x}{\partial s} + \frac{\partial f}{\partial y} \cdot \frac{\partial y}{\partial s} \\ &= \left(\frac{\partial f}{\partial x} \hat{\mathbf{i}} + \frac{\partial f}{\partial y} \hat{\mathbf{j}} \right) \cdot \left(\frac{\partial x}{\partial s} \hat{\mathbf{i}} + \frac{\partial y}{\partial s} \hat{\mathbf{j}} \right) \\ &= 0, \end{aligned}$$

where the derivatives are all evaluated at $s = 0$. However, the first factor in the dot product is the gradient of f, and the second factor is the unit tangent vector $\vec{\mathbf{T}}(0)$ to the constraint curve. Since the point (x_0, y_0) corresponds to $s = 0$, it follows from this equation that

$$\vec{\nabla} f(x_0, y_0) \cdot \vec{\mathbf{T}}(0) = 0,$$

which implies that the gradient is either the zero vector $\vec{\mathbf{0}}$ or it is normal to the constraint curve at a constrained relative extremum. However, the constraint curve $g(x,y)=0$ is a level curve for the function $g(x,y)$ so that if $\vec{\nabla} g(x_0,y_0)\neq 0$ then $\vec{\nabla} g(x_0,y_0)$ is normal to this curve at (x_0,y_0) It follows, then, that there is some scalar λ such that

$$\vec{\nabla} f(x_0,y_0)=\lambda\vec{\nabla} g(x_0,y_0)$$

□

To apply Theorem 13.8.1 to an optimization problem similar to that for the golf ball manufacturer, we need a problem-solving strategy.

Problem-Solving Strategy: Steps for Using Lagrange Multipliers

1. Determine the objective function $f(x,y)$ and the constraint function $g(x,y)$. Does the optimization problem involve maximizing or minimizing the objective function?
2. Set up a system of equations using the following template:

$$\vec{\nabla} f(x_0,y_0)=\lambda\vec{\nabla} g(x_0,y_0) \quad (13.8.1)$$
$$g(x_0,y_0)=0 \quad (13.8.2)$$

3. Solve for x_0 and y_0.
4. The largest of the values of f at the solutions found in step 3 maximizes f; the smallest of those values minimizes f.

Example 13.8.1: Using Lagrange Multipliers

Use the method of Lagrange multipliers to find the minimum value of $f(x,y)=x^2+4y^2-2x+8y$ subject to the constraint $x+2y=7$.

Solution

Let's follow the problem-solving strategy:

1. The objective function is $f(x,y)=x^2+4y^2-2x+8y$. To determine the constraint function, we must first subtract 7 from both sides of the constraint. This gives $x+2y-7=0$. The constraint function is equal to the left-hand side, so $g(x,y)=x+2y-7$. The problem asks us to solve for the minimum value of f, subject to the constraint (Figure 13.8.3).

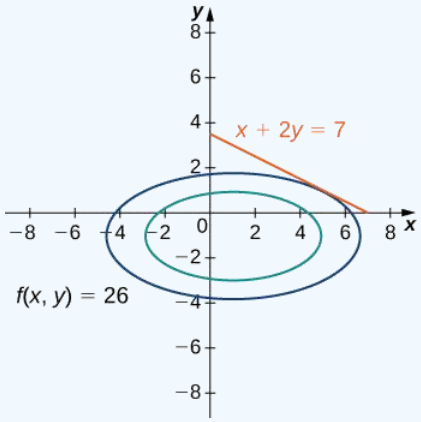


Figure 13.8.3: Graph of level curves of the function $f(x,y)=x^2+4y^2-2x+8y$ corresponding to $c=10$ and 26. The red graph is the constraint function.

2. We then must calculate the gradients of both f and g:

$$\vec{\nabla} f(x,y) = (2x-2)\,\hat{\mathbf{i}}+(8y+8)\,\hat{\mathbf{j}}$$
$$\vec{\nabla} g(x,y) = \hat{\mathbf{i}}+2\hat{\mathbf{j}}.$$

The equation $\vec{\nabla} f(x_0,y_0)=\lambda\vec{\nabla} g(x_0,y_0)$ becomes

$$(2x_0 - 2)\,\hat{\mathbf{i}} + (8y_0 + 8)\,\hat{\mathbf{j}} = \lambda\left(\hat{\mathbf{i}} + 2\hat{\mathbf{j}}\right),$$

which can be rewritten as

$$(2x_0 - 2)\,\hat{\mathbf{i}} + (8y_0 + 8)\,\hat{\mathbf{j}} = \lambda\hat{\mathbf{i}} + 2\lambda\hat{\mathbf{j}}.$$

Next, we set the coefficients of $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ equal to each other:

$$\begin{aligned} 2x_0 - 2 &= \lambda \\ 8y_0 + 8 &= 2\lambda. \end{aligned}$$

The equation $g(x_0, y_0) = 0$ becomes $x_0 + 2y_0 - 7 = 0$. Therefore, the system of equations that needs to be solved is

$$\begin{aligned} 2x_0 - 2 &= \lambda \\ 8y_0 + 8 &= 2\lambda \\ x_0 + 2y_0 - 7 &= 0. \end{aligned}$$

3. This is a linear system of three equations in three variables. We start by solving the second equation for λ and substituting it into the first equation. This gives $\lambda = 4y_0 + 4$, so substituting this into the first equation gives

$$2x_0 - 2 = 4y_0 + 4.$$

Solving this equation for x_0 gives $x_0 = 2y_0 + 3$. We then substitute this into the third equation:

$$\begin{aligned} (2y_0 + 3) + 2y_0 - 7 &= 0 \\ 4y_0 - 4 &= 0 \\ y_0 &= 1. \end{aligned}$$

Since $x_0 = 2y_0 + 3,$ this gives $x_0 = 5.$

4. Next, we evaluate $f(x, y) = x^2 + 4y^2 - 2x + 8y$ at the point $(5, 1)$,

$$f(5, 1) = 5^2 + 4(1)^2 - 2(5) + 8(1) = 27.$$

To ensure this corresponds to a minimum value on the constraint function, let's try some other points on the constraint from either side of the point $(5, 1)$, such as the intercepts of $g(x, y) = 0$, Which are $(7, 0)$ and $(0, 3.5)$.

We get $f(7, 0) = 35 > 27$ and $f(0, 3.5) = 77 > 27.$

So it appears that f has a relative minimum of 27 at $(5, 1)$, subject to the given constraint.

? Exercise 13.8.1

Use the method of Lagrange multipliers to find the maximum value of

$$f(x, y) = 9x^2 + 36xy - 4y^2 - 18x - 8y$$

subject to the constraint $3x + 4y = 32.$

Hint

Use the problem-solving strategy for the method of Lagrange multipliers.

Answer

Subject to the given constraint, f has a maximum value of 976 at the point $(8, 2)$.

Let's now return to the problem posed at the beginning of the section.

Example 13.8.2: Golf Balls and Lagrange Multipliers

The golf ball manufacturer, Pro-T, has developed a profit model that depends on the number x of golf balls sold per month (measured in thousands), and the number of hours per month of advertising y, according to the function

$$z = f(x,y) = 48x + 96y - x^2 - 2xy - 9y^2,$$

where z is measured in thousands of dollars. The budgetary constraint function relating the cost of the production of thousands golf balls and advertising units is given by $20x + 4y = 216$. Find the values of x and y that maximize profit, and find the maximum profit.

Solution:

Again, we follow the problem-solving strategy:

1. The objective function is $f(x,y) = 48x + 96y - x^2 - 2xy - 9y^2$. To determine the constraint function, we first subtract 216 from both sides of the constraint, then divide both sides by 4, which gives $5x + y - 54 = 0$. The constraint function is equal to the left-hand side, so $g(x,y) = 5x + y - 54$. The problem asks us to solve for the maximum value of f, subject to this constraint.
2. So, we calculate the gradients of both f and g:

$$\begin{aligned} \vec{\nabla} f(x,y) &= (48 - 2x - 2y)\hat{\mathbf{i}} + (96 - 2x - 18y)\hat{\mathbf{j}} \\ \vec{\nabla} g(x,y) &= 5\hat{\mathbf{i}} + \hat{\mathbf{j}}. \end{aligned}$$

   The equation $\vec{\nabla} f(x_0,y_0) = \lambda \vec{\nabla} g(x_0,y_0)$ becomes

$$(48 - 2x_0 - 2y_0)\hat{\mathbf{i}} + (96 - 2x_0 - 18y_0)\hat{\mathbf{j}} = \lambda(5\hat{\mathbf{i}} + \hat{\mathbf{j}}),$$

   which can be rewritten as

$$(48 - 2x_0 - 2y_0)\hat{\mathbf{i}} + (96 - 2x_0 - 18y_0)\hat{\mathbf{j}} = \lambda 5\hat{\mathbf{i}} + \lambda\hat{\mathbf{j}}.$$

   We then set the coefficients of $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ equal to each other:

$$\begin{aligned} 48 - 2x_0 - 2y_0 &= 5\lambda \\ 96 - 2x_0 - 18y_0 &= \lambda. \end{aligned}$$

   The equation $g(x_0,y_0) = 0$ becomes $5x_0 + y_0 - 54 = 0$. Therefore, the system of equations that needs to be solved is

$$\begin{aligned} 48 - 2x_0 - 2y_0 &= 5\lambda \\ 96 - 2x_0 - 18y_0 &= \lambda \\ 5x_0 + y_0 - 54 &= 0. \end{aligned}$$

3. We use the left-hand side of the second equation to replace λ in the first equation:

$$\begin{aligned} 48 - 2x_0 - 2y_0 &= 5(96 - 2x_0 - 18y_0) \\ 48 - 2x_0 - 2y_0 &= 480 - 10x_0 - 90y_0 \\ 8x_0 &= 432 - 88y_0 \\ x_0 &= 54 - 11y_0. \end{aligned}$$

   Then we substitute this into the third equation:

$$\begin{aligned} 5(54 - 11y_0) + y_0 - 54 &= 0 \\ 270 - 55y_0 + y_0 - 54 &= 0 \\ 216 - 54y_0 &= 0 \\ y_0 &= 4. \end{aligned}$$

   Since $x_0 = 54 - 11y_0$, this gives $x_0 = 10$.

4. We then substitute $(10,4)$ into $f(x,y)=48x+96y-x^2-2xy-9y^2,$ which gives

$$\begin{aligned} f(10,4) &= 48(10)+96(4)-(10)^2-2(10)(4)-9(4)^2 \\ &= 480+384-100-80-144 \\ &= 540. \end{aligned}$$

Therefore the maximum profit that can be attained, subject to budgetary constraints, is $\$540,000$ with a production level of $10,000$ golf balls and 4 hours of advertising bought per month. Let's check to make sure this truly is a maximum. The endpoints of the line that defines the constraint are $(10.8,0)$ and $(0,54)$ Let's evaluate f at both of these points:

$$\begin{aligned} f(10.8,0) &= 48(10.8)+96(0)-10.8^2-2(10.8)(0)-9(0^2) \\ &= 401.76 \\ f(0,54) &= 48(0)+96(54)-0^2-2(0)(54)-9(54^2) \\ &= -21,060. \end{aligned}$$

The second value represents a loss, since no golf balls are produced. Neither of these values exceed 540, so it seems that our extremum is a maximum value of f, subject to the given constraint.

? Exercise 13.8.2: Optimizing the Cobb-Douglas function

A company has determined that its production level is given by the Cobb-Douglas function $f(x,y)=2.5x^{0.45}y^{0.55}$ where x represents the total number of labor hours in 1 year and y represents the total capital input for the company. Suppose 1 unit of labor costs $\$40$ and 1 unit of capital costs $\$50$. Use the method of Lagrange multipliers to find the maximum value of $f(x,y)=2.5x^{0.45}y^{0.55}$ subject to a budgetary constraint of $\$500,000$ per year.

Hint

Use the problem-solving strategy for the method of Lagrange multipliers.

Answer

Subject to the given constraint, a maximum production level of 13890 occurs with 5625 labor hours and $\$5500$ of total capital input.

In the case of an objective function with three variables and a single constraint function, it is possible to use the method of Lagrange multipliers to solve an optimization problem as well. An example of an objective function with three variables could be the **Cobb-Douglas function** in Exercise 13.8.2 $f(x,y,z)=x^{0.2}y^{0.4}z^{0.4}$, where x represents the cost of labor, y represents capital input, and z represents the cost of advertising. The method is the same as for the method with a function of two variables; the equations to be solved are

$$\begin{aligned} \vec{\nabla} f(x,y,z) &= \lambda \vec{\nabla} g(x,y,z) \\ g(x,y,z) &= 0. \end{aligned}$$

✔ Example 13.8.3: Lagrange Multipliers with a Three-Variable objective function

Maximize the function $f(x,y,z)=x^2+y^2+z^2$ subject to the constraint $x+y+z=1$.

Solution

1. The objective function is $f(x,y,z)=x^2+y^2+z^2$. To determine the constraint function, we subtract 1 from each side of the constraint: $x+y+z-1=0$ which gives the constraint function as $g(x,y,z)=x+y+z-1$.

2. Next, we calculate $\vec{\nabla} f(x,y,z)$ and $\vec{\nabla} g(x,y,z)$:

$$\begin{aligned} \vec{\nabla} f(x,y,z) &= \langle 2x,2y,2z\rangle \\ \vec{\nabla} g(x,y,z) &= \langle 1,1,1\rangle. \end{aligned}$$

This leads to the equations

$$\begin{aligned}\langle 2x_0, 2y_0, 2z_0\rangle &= \lambda\langle 1,1,1\rangle \\ x_0 + y_0 + z_0 - 1 &= 0\end{aligned}$$

which can be rewritten in the following form:

$$\begin{aligned}2x_0 &= \lambda \\ 2y_0 &= \lambda \\ 2z_0 &= \lambda \\ x_0 + y_0 + z_0 - 1 &= 0.\end{aligned}$$

3. Since each of the first three equations has λ on the right-hand side, we know that $2x_0 = 2y_0 = 2z_0$ and all three variables are equal to each other. Substituting $y_0 = x_0$ and $z_0 = x_0$ into the last equation yields $3x_0 - 1 = 0,$ so $x_0 = \frac{1}{3}$ and $y_0 = \frac{1}{3}$ and $z_0 = \frac{1}{3}$ which corresponds to a critical point on the constraint curve.

4. Then, we evaluate f at the point $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$:

$$f\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) = \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 = \frac{3}{9} = \frac{1}{3}$$

Therefore, a possible extremum of the function is $\frac{1}{3}$. To verify it is a minimum, choose other points that satisfy the constraint from either side of the point we obtained above and calculate f at those points. For example,

$$\begin{aligned}f(1,0,0) &= 1^2 + 0^2 + 0^2 = 1 \\ f(0,-2,3) &= 0^2 + (-2)^2 + 3^2 = 13.\end{aligned}$$

Both of these values are greater than $\frac{1}{3}$, leading us to believe the extremum is a minimum, subject to the given constraint.

? Exercise 13.8.3

Use the method of Lagrange multipliers to find the minimum value of the function

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

subject to the constraint $x^2 + y^2 + z^2 = 1$.

Hint

Use the problem-solving strategy for the method of Lagrange multipliers with an objective function of three variables.

Answer

Evaluating f at both points we obtained, gives us,

$$f\left(\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}\right) = \frac{\sqrt{3}}{3} + \frac{\sqrt{3}}{3} + \frac{\sqrt{3}}{3} = \sqrt{3}$$

$$f\left(-\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}\right) = -\frac{\sqrt{3}}{3} - \frac{\sqrt{3}}{3} - \frac{\sqrt{3}}{3} = -\sqrt{3}$$

Since the constraint is continuous, we compare these values and conclude that f has a relative minimum of $-\sqrt{3}$ at the point $\left(-\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}\right)$, subject to the given constraint.

Problems with Two Constraints

The method of Lagrange multipliers can be applied to problems with more than one constraint. In this case the objective function, w is a function of three variables:

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

and it is subject to two constraints:

$$g(x, y, z) = 0 \text{ and } h(x, y, z) = 0.$$

There are two Lagrange multipliers, λ_1 and λ_2, and the system of equations becomes

$$\begin{aligned} \vec{\nabla} f(x_0, y_0, z_0) &= \lambda_1 \vec{\nabla} g(x_0, y_0, z_0) + \lambda_2 \vec{\nabla} h(x_0, y_0, z_0) \\ g(x_0, y_0, z_0) &= 0 \\ h(x_0, y_0, z_0) &= 0 \end{aligned}$$

Example 13.8.4: Lagrange Multipliers with Two Constraints

Find the maximum and minimum values of the function

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

subject to the constraints $z^2 = x^2 + y^2$ and $x + y - z + 1 = 0$.

Solution

Let's follow the problem-solving strategy:

1. The objective function is $f(x, y, z) = x^2 + y^2 + z^2$. To determine the constraint functions, we first subtract z^2 from both sides of the first constraint, which gives $x^2 + y^2 - z^2 = 0$, so $g(x, y, z) = x^2 + y^2 - z^2$. The second constraint function is $h(x, y, z) = x + y - z + 1$.
2. We then calculate the gradients of f, g, and h:

$$\begin{aligned} \vec{\nabla} f(x, y, z) &= 2x\,\hat{\mathbf{i}} + 2y\hat{\mathbf{j}} + 2z\hat{\mathbf{k}} \\ \vec{\nabla} g(x, y, z) &= 2x\,\hat{\mathbf{i}} + 2y\hat{\mathbf{j}} - 2z\hat{\mathbf{k}} \\ \vec{\nabla} h(x, y, z) &= \hat{\mathbf{i}} + \hat{\mathbf{j}} - \hat{\mathbf{k}}. \end{aligned}$$

The equation $\vec{\nabla} f(x_0, y_0, z_0) = \lambda_1 \vec{\nabla} g(x_0, y_0, z_0) + \lambda_2 \vec{\nabla} h(x_0, y_0, z_0)$ becomes

$$2x_0\,\hat{\mathbf{i}} + 2y_0\hat{\mathbf{j}} + 2z_0\hat{\mathbf{k}} = \lambda_1(2x_0\,\hat{\mathbf{i}} + 2y_0\hat{\mathbf{j}} - 2z_0\hat{\mathbf{k}}) + \lambda_2(\hat{\mathbf{i}} + \hat{\mathbf{j}} - \hat{\mathbf{k}}),$$

which can be rewritten as

$$2x_0\,\hat{\mathbf{i}} + 2y_0\hat{\mathbf{j}} + 2z_0\hat{\mathbf{k}} = (2\lambda_1 x_0 + \lambda_2)\hat{\mathbf{i}} + (2\lambda_1 y_0 + \lambda_2)\hat{\mathbf{j}} - (2\lambda_1 z_0 + \lambda_2)\hat{\mathbf{k}}.$$

Next, we set the coefficients of $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ equal to each other:

$$\begin{aligned} 2x_0 &= 2\lambda_1 x_0 + \lambda_2 \\ 2y_0 &= 2\lambda_1 y_0 + \lambda_2 \\ 2z_0 &= -2\lambda_1 z_0 - \lambda_2. \end{aligned}$$

The two equations that arise from the constraints are $z_0^2 = x_0^2 + y_0^2$ and $x_0 + y_0 - z_0 + 1 = 0$. Combining these equations with the previous three equations gives

$$\begin{aligned} 2x_0 &= 2\lambda_1 x_0 + \lambda_2 \\ 2y_0 &= 2\lambda_1 y_0 + \lambda_2 \\ 2z_0 &= -2\lambda_1 z_0 - \lambda_2 \\ z_0^2 &= x_0^2 + y_0^2 \\ x_0 + y_0 - z_0 + 1 &= 0. \end{aligned}$$

3. The first three equations contain the variable λ_2. Solving the third equation for λ_2 and replacing into the first and second equations reduces the number of equations to four:

$$\begin{aligned} 2x_0 &= 2\lambda_1 x_0 - 2\lambda_1 z_0 - 2z_0 \\ 2y_0 &= 2\lambda_1 y_0 - 2\lambda_1 z_0 - 2z_0 \\ z_0^2 &= x_0^2 + y_0^2 \\ x_0 + y_0 - z_0 + 1 &= 0. \end{aligned}$$

Next, we solve the first and second equation for λ_1. The first equation gives $\lambda_1 = \dfrac{x_0 + z_0}{x_0 - z_0}$, the second equation gives $\lambda_1 = \dfrac{y_0 + z_0}{y_0 - z_0}$. We set the right-hand side of each equation equal to each other and cross-multiply:

$$\begin{aligned} \frac{x_0 + z_0}{x_0 - z_0} &= \frac{y_0 + z_0}{y_0 - z_0} \\ (x_0 + z_0)(y_0 - z_0) &= (x_0 - z_0)(y_0 + z_0) \\ x_0 y_0 - x_0 z_0 + y_0 z_0 - z_0^2 &= x_0 y_0 + x_0 z_0 - y_0 z_0 - z_0^2 \\ 2y_0 z_0 - 2x_0 z_0 &= 0 \\ 2z_0(y_0 - x_0) &= 0. \end{aligned}$$

Therefore, either $z_0 = 0$ or $y_0 = x_0$. If $z_0 = 0$, then the first constraint becomes $0 = x_0^2 + y_0^2$. The only real solution to this equation is $x_0 = 0$ and $y_0 = 0$, which gives the ordered triple $(0, 0, 0)$. This point does not satisfy the second constraint, so it is not a solution. Next, we consider $y_0 = x_0$, which reduces the number of equations to three:

$$\begin{aligned} y_0 &= x_0 \\ z_0^2 &= x_0^2 + y_0^2 \\ x_0 + y_0 - z_0 + 1 &= 0. \end{aligned}$$

We substitute the first equation into the second and third equations:

$$\begin{aligned} z_0^2 &= x_0^2 + x_0^2 \\ &= x_0 + x_0 - z_0 + 1 \qquad = 0. \end{aligned}$$

Then, we solve the second equation for z_0, which gives $z_0 = 2x_0 + 1$. We then substitute this into the first equation,

$$\begin{aligned} z_0^2 &= 2x_0^2 \\ (2x_0 + 1)^2 &= 2x_0^2 \\ 4x_0^2 + 4x_0 + 1 &= 2x_0^2 \\ 2x_0^2 + 4x_0 + 1 &= 0, \end{aligned}$$

and use the quadratic formula to solve for x_0:

$$x_0 = \frac{-4 \pm \sqrt{4^2 - 4(2)(1)}}{2(2)} = \frac{-4 \pm \sqrt{8}}{4} = \frac{-4 \pm 2\sqrt{2}}{4} = -1 \pm \frac{\sqrt{2}}{2}.$$

Recall $y_0 = x_0$, so this solves for y_0 as well. Then, $z_0 = 2x_0 + 1$, so

$$z_0 = 2x_0 + 1 = 2\left(-1 \pm \frac{\sqrt{2}}{2}\right) + 1 = -2 + 1 \pm \sqrt{2} = -1 \pm \sqrt{2}.$$

Therefore, there are two ordered triplet solutions:

$$\left(-1 + \frac{\sqrt{2}}{2}, -1 + \frac{\sqrt{2}}{2}, -1 + \sqrt{2}\right) \text{ and } \left(-1 - \frac{\sqrt{2}}{2}, -1 - \frac{\sqrt{2}}{2}, -1 - \sqrt{2}\right).$$

4. We substitute $\left(-1 + \dfrac{\sqrt{2}}{2}, -1 + \dfrac{\sqrt{2}}{2}, -1 + \sqrt{2}\right)$ into $f(x, y, z) = x^2 + y^2 + z^2$, which gives

$$f\left(-1+\frac{\sqrt{2}}{2},-1+\frac{\sqrt{2}}{2},-1+\sqrt{2}\right) = \left(-1+\frac{\sqrt{2}}{2}\right)^2+\left(-1+\frac{\sqrt{2}}{2}\right)^2+(-1+\sqrt{2})^2$$
$$= \left(1-\sqrt{2}+\frac{1}{2}\right)+\left(1-\sqrt{2}+\frac{1}{2}\right)+(1-2\sqrt{2}+2)$$
$$= 6-4\sqrt{2}.$$

Then, we substitute $\left(-1-\frac{\sqrt{2}}{2},-1+\frac{\sqrt{2}}{2},-1+\sqrt{2}\right)$ into $f(x,y,z)=x^2+y^2+z^2$, which gives

$$f\left(-1-\frac{\sqrt{2}}{2},-1+\frac{\sqrt{2}}{2},-1+\sqrt{2}\right) = \left(-1-\frac{\sqrt{2}}{2}\right)^2+\left(-1-\frac{\sqrt{2}}{2}\right)^2+(-1-\sqrt{2})^2$$
$$= \left(1+\sqrt{2}+\frac{1}{2}\right)+\left(1+\sqrt{2}+\frac{1}{2}\right)+(1+2\sqrt{2}+2)$$
$$= 6+4\sqrt{2}.$$

$6+4\sqrt{2}$ is the maximum value and $6-4\sqrt{2}$ is the minimum value of $f(x,y,z)$, subject to the given constraints.

? Exercise 13.8.4

Use the method of Lagrange multipliers to find the minimum value of the function

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

subject to the constraints $2x+y+2z=9$ and $5x+5y+7z=29$.

Hint

Use the problem-solving strategy for the method of Lagrange multipliers with two constraints.

Answer

$f(2,1,2)=9$ is a minimum value of f, subject to the given constraints.

Key Concepts

- An objective function combined with one or more constraints is an example of an optimization problem.
- To solve optimization problems, we apply the method of Lagrange multipliers using a four-step problem-solving strategy.

Key Equations

- **Method of Lagrange multipliers, one constraint**

  $\vec{\nabla} f(x_0,y_0)=\lambda\vec{\nabla} g(x_0,y_0)$

  $g(x_0,y_0)=0$

- **Method of Lagrange multipliers, two constraints**

  $\vec{\nabla} f(x_0,y_0,z_0)=\lambda_1\vec{\nabla} g(x_0,y_0,z_0)+\lambda_2\vec{\nabla} h(x_0,y_0,z_0)$

  $g(x_0,y_0,z_0)=0$

  $h(x_0,y_0,z_0)=0$

Glossary

constraint

an inequality or equation involving one or more variables that is used in an optimization problem; the constraint enforces a limit on the possible solutions for the problem

Lagrange multiplier

the constant (or constants) used in the method of Lagrange multipliers; in the case of one constant, it is represented by the variable λ

method of Lagrange multipliers

a method of solving an optimization problem subject to one or more constraints

objective function

the function that is to be maximized or minimized in an optimization problem

optimization problem

calculation of a maximum or minimum value of a function of several variables, often using Lagrange multipliers

---

13.8: Lagrange Multipliers is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **14.8: Lagrange Multipliers** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

CHAPTER OVERVIEW

14: Multiple Integrals

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

14.1: Double and Iterated Integrals over Rectangles
14.2: Double Integrals over General Regions
14.3: Area by Double Integration
14.4: Double Integrals in Polar Form
14.5: Triple Integrals in Rectangular Coordinates
14.6: Moments and Centers of Mass
14.7: Triple Integrals in Cylindrical and Spherical Coordinates
14.8: Substitutions in Multiple Integrals

14: Multiple Integrals is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

14.1: Double and Iterated Integrals over Rectangles

Learning Objectives

- Recognize when a function of two variables is integrable over a rectangular region.
- Recognize and use some of the properties of double integrals.
- Evaluate a double integral over a rectangular region by writing it as an iterated integral.
- Use a double integral to calculate the area of a region, volume under a surface, or average value of a function over a plane region.

In this section we investigate double integrals and show how we can use them to find the volume of a solid over a rectangular region in the xy-plane. Many of the properties of double integrals are similar to those we have already discussed for single integrals.

Volumes and Double Integrals

We begin by considering the space above a rectangular region R. Consider a continuous function $f(x, y) \geq 0$ of two variables defined on the closed rectangle R:

$$R = [a, b] \times [c, d] = \{(x, y) \in \mathbb{R}^2 \,|\, a \leq x \leq b,\ c \leq y \leq d\}$$

Here $[a, b] \times [c, d]$ denotes the Cartesian product of the two closed intervals $[a, b]$ and $[c, d]$. It consists of rectangular pairs (x, y) such that $a \leq x \leq b$ and $c \leq y \leq d$. The graph of f represents a surface above the xy-plane with equation $z = f(x, y)$ where z is the height of the surface at the point (x, y). Let S be the solid that lies above R and under the graph of f (Figure 14.1.1). The base of the solid is the rectangle R in the xy-plane. We want to find the volume V of the solid S.

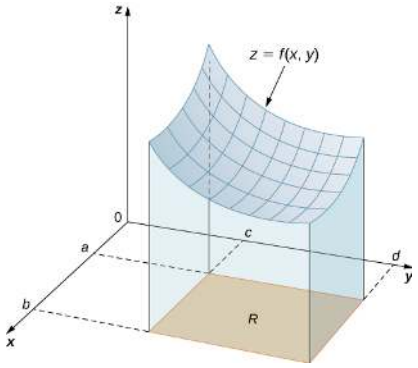


Figure 14.1.1: The graph of $f(x, y)$ *over the rectangle R in the* xy*-plane is a curved surface.*

We divide the region R into small rectangles R_{ij}, each with area ΔA and with sides Δx and Δy (Figure 14.1.2). We do this by dividing the interval $[a, b]$ into m subintervals and dividing the interval $[c, d]$ into n subintervals. Hence $\Delta x = \frac{b-a}{m}$, $\Delta y = \frac{d-c}{n}$, and $\Delta A = \Delta x \Delta y$.

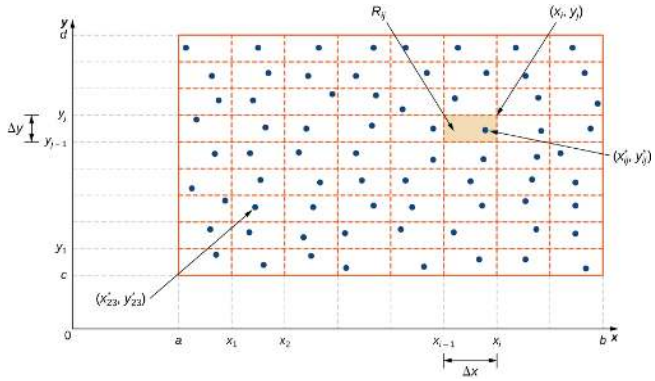


Figure 14.1.2: Rectangle R is divided into small rectangles R_{ij} each with area ΔA.

The volume of a thin rectangular box above R_{ij} is $f(x_{ij}^*, y_{ij}^*)\,\Delta A$, where (x_{ij}^*, y_{ij}^*) is an arbitrary sample point in each R_{ij} as shown in the following figure, $f(x_{ij}^*, y_{ij}^*)$ is the height of the corresponding thin rectangular box, and ΔA is the area of each rectangle R_{ij}.

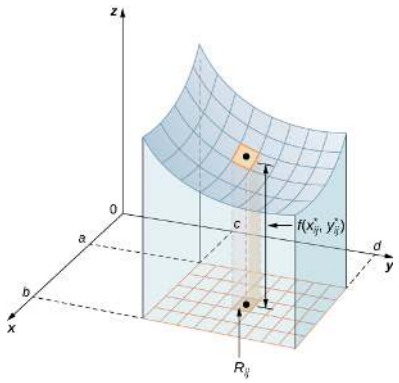


Figure 14.1.3: A thin rectangular box above R_{ij} with height $f(x_{ij}^*, y_{ij}^*)$.

Using the same idea for all the subrectangles, we obtain an approximate volume of the solid S as

$$V \approx \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_{ij}^*, y_{ij}^*) \Delta A.$$

This sum is known as a **double Riemann sum** and can be used to approximate the value of the volume of the solid. Here the double sum means that for each subrectangle we evaluate the function at the chosen point, multiply by the area of each rectangle, and then add all the results.

As we have seen in the single-variable case, we obtain a better approximation to the actual volume if m and n become larger.

$$V = \lim_{m,n \to \infty} \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_{ij}^*, y_{ij}^*) \Delta A$$

or

$$V = \lim_{\Delta x,\, \Delta y \to 0} \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_{ij}^*, y_{ij}^*) \Delta A.$$

Note that the sum approaches a limit in either case and the limit is the volume of the solid with the base R. Now we are ready to define the double integral.

Definition: Double Integral over a Rectangular Region R

The double integral of the function $f(x, y)$ over the rectangular region R in the xy-plane is defined as

$$\iint_R f(x, y) dA = \lim_{m,n \to \infty} \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_{ij}^*, y_{ij}^*) \Delta A.$$

If $f(x, y) \geq 0$, then the volume V of the solid S, which lies above R in the xy-plane and under the graph of f, is the double integral of the function $f(x, y)$ over the rectangle R. If the function is ever negative, then the double integral can be considered a "signed" volume in a manner similar to the way we defined net signed area in The Definite Integral.

Example 14.1.1: Setting up a Double Integral and Approximating It by Double Sums

Consider the function $z = f(x, y) = 3x^2 - y$ over the rectangular region $R = [0, 2] \times [0, 2]$ (Figure 14.1.4).

a. Set up a double integral for finding the value of the signed volume of the solid S that lies above R and "under" the graph of f.
b. Divide R into four squares with $m = n = 2$, and choose the sample point as the upper right corner point of each square (1,1),(2,1),(1,2), and (2,2) (Figure 14.1.4) to approximate the signed volume of the solid S that lies above R and "under" the graph of f.
c. Divide R into four squares with $m = n = 2$, and choose the sample point as the midpoint of each square: (1/2, 1/2), (3/2, 1/2), (1/2,3/2), and (3/2, 3/2) to approximate the signed volume.

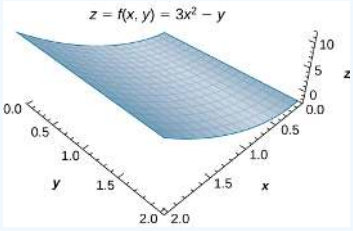


Figure 14.1.4: The function $z = f(x, y)$ graphed over the rectangular region $R = [0, 2] \times [0, 2]$.

Solution

a. As we can see, the function $z = f(x, y) = 3x^2 - y$ is above the plane. To find the signed volume of S, we need to divide the region R into small rectangles R_{ij}, each with area ΔA and with sides Δx and Δy, and choose (x_{ij}^*, y_{ij}^*) as sample points in each R_{ij}. Hence, a double integral is set up as

$$V = \iint_R (3x^2 - y)dA = \lim_{m,n\to\infty} \sum_{i=1}^{m}\sum_{j=1}^{n} [3(x_{ij}^*)^2 - y_{ij}^*]\Delta A.$$

b. Approximating the signed volume using a Riemann sum with $m = n = 2$ we have $\Delta A = \Delta x \Delta y = 1 \times 1 = 1$. Also, the sample points are (1, 1), (2, 1), (1, 2), and (2, 2) as shown in the following figure.

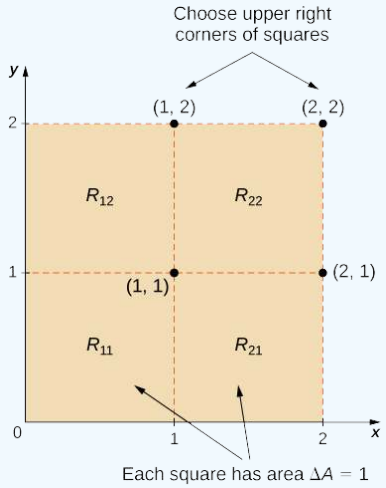


Figure 14.1.5: Subrectangles for the rectangular region $R = [0, 2] \times [0, 2]$.

Hence,

$$\begin{aligned} V &\approx \sum_{i=1}^{2}\sum_{j=1}^{2} f(x_{ij}^*, y_{ij}^*)\Delta A \\ &= \sum_{i=1}^{2} (f(x_{i1}^*, y_{i1}^*) + f(x_{i2}^*, y_{i2}^*))\Delta A \\ &= f(x_{11}^*, y_{11}^*)\Delta A + f(x_{21}^*, y_{21}^*)\Delta A + f(x_{12}^*, y_{12}^*)\Delta A + f(x_{22}^*, y_{22}^*)\Delta A \\ &= f(1,1)(1) + f(2,1)(1) + f(1,2)(1) + f(2,2)(1) \\ &= (3-1)(1) + (12-1)(1) + (3-2)(1) + (12-2)(1) \\ &= 2 + 11 + 1 + 10 = 24. \end{aligned}$$

c. Approximating the signed volume using a Riemann sum with $m = n = 2$ we have$\Delta A = \Delta x \Delta y = 1 \times 1 = 1$. In this case the sample points are (1/2, 1/2), (3/2, 1/2), (1/2, 3/2), and (3/2, 3/2).
Hence,

$$\begin{aligned} V &\approx \sum_{i=1}^{2}\sum_{j=1}^{2} f(x_{ij}^*, y_{ij}^*)\Delta A \\ &= f(x_{11}^*, y_{11}^*)\Delta A + f(x_{21}^*, y_{21}^*)\Delta A + f(x_{12}^*, y_{12}^*)\Delta A + f(x_{22}^*, y_{22}^*)\Delta A \\ &= f(1/2, 1/2)(1) + f(3/2, 1/2)(1) + f(1/2, 3/2)(1) + f(3/2, 3/2)(1) \\ &= \left(\frac{3}{4} - \frac{1}{4}\right)(1) + \left(\frac{27}{4} - \frac{1}{2}\right)(1) + \left(\frac{3}{4} - \frac{3}{2}\right)(1) + \left(\frac{27}{4} - \frac{3}{2}\right)(1) \\ &= \frac{2}{4} + \frac{25}{4} + \left(-\frac{3}{4}\right) + \frac{21}{4} = \frac{45}{4} = 11. \end{aligned}$$

Analysis

Notice that the approximate answers differ due to the choices of the sample points. In either case, we are introducing some error because we are using only a few sample points. Thus, we need to investigate how we can achieve an accurate answer.

? Exercise 14.1.1

Use the same function$z = f(x, y) = 3x^2 - y$ over the rectangular region $R = [0, 2] \times [0, 2]$.

Divide R into the same four squares with $m = n = 2$, and choose the sample points as the upper left corner point of each square (0,1), (1,1), (0,2), and (1,2) (Figure 14.1.5) to approximate the signed volume of the solid S that lies above R and "under" the graph of f.

Hint

Follow the steps of the previous example.

Answer

$$V \approx \sum_{i=1}^{2} \sum_{j=1}^{2} f(x_{ij}^*, y_{ij}^*)\, \Delta A = 0$$

Note that we developed the concept of double integral using a rectangular region R. This concept can be extended to any general region. However, when a region is not rectangular, the subrectangles may not all fit perfectly into R, particularly if the base area is curved. We examine this situation in more detail in the next section, where we study regions that are not always rectangular and subrectangles may not fit perfectly in the region R. Also, the heights may not be exact if the surface $z = f(x, y)$ is curved. However, the errors on the sides and the height where the pieces may not fit perfectly within the solid S approach 0 as m and n approach infinity. Also, the double integral of the function $z = f(x, y)$ exists provided that the function f is not too discontinuous. If the function is bounded and continuous over R except on a finite number of smooth curves, then the double integral exists and we say that ff is integrable over R.

Since $\Delta A = \Delta x \Delta y = \Delta y \Delta x$, we can express dA as $dx\, dy$ or $dy\, dx$. This means that, when we are using rectangular coordinates, the double integral over a region R denoted by

$$\iint_R f(x, y)\, dA$$

can be written as

$$\iint_R f(x, y)\, dx\, dy$$

or

$$\iint_R f(x, y)\, dy\, dx.$$

Now let's list some of the properties that can be helpful to compute double integrals.

Properties of Double Integrals

The properties of double integrals are very helpful when computing them or otherwise working with them. We list here six properties of double integrals. Properties 1 and 2 are referred to as the linearity of the integral, property 3 is the additivity of the integral, property 4 is the monotonicity of the integral, and property 5 is used to find the bounds of the integral. Property 6 is used if $f(x, y)$ is a product of two functions $g(x)$ and $h(y)$.

Theorem: Properties of Double Integrals

Assume that the functions $f(x, y)$ and $g(x, y)$ are integrable over the rectangular region R; S and T are subregions of R; and assume that m and M are real numbers.

i. The sum $f(x, y) + g(x, y)$ is integrable and

$$\iint_R [f(x, y) + g(x, y)]\, dA = \iint_R f(x, y)\, dA + \iint_R g(x, y)\, dA.$$

ii. If c is a constant, then $cf(x, y)$ is integrable and

$$\iint_R cf(x, y)\, dA = c \iint_R f(x, y)\, dA.$$

iii. If $R = S \cup T$ and $S \cap T = \emptyset$ except an overlap on the boundaries, then

$$\iint_R f(x, y)\, dA = \iint_S f(x, y)\, dA + \iint_T f(x, y)\, dA.$$

iv. If $f(x, y) \geq g(x, y)$ for (x, y) in R, then

$$\iint_R f(x, y)\, dA \geq \iint_R g(x, y)\, dA.$$

v. If $m \leq f(x, y) \leq M$ and $A(R) =$ the area of R, then

$$m \cdot A(R) \leq \iint_R f(x, y)\, dA \leq M \cdot A(R).$$

vi. In the case where $f(x, y)$ can be factored as a product of a function $g(x)$ of x only and a function $h(y)$ of y only, then over the region $R = \{(x, y) \mid a \leq x \leq b,\ c \leq y \leq d\}$, the double integral can be written as

$$\iint_R f(x, y)\, dA = \left(\int_a^b g(x)\, dx \right) \left(\int_c^d h(y)\, dy \right).$$

These properties are used in the evaluation of double integrals, as we will see later. We will become skilled in using these properties once we become familiar with the computational tools of double integrals. So let's get to that now.

Iterated Integrals

So far, we have seen how to set up a double integral and how to obtain an approximate value for it. We can also imagine that evaluating double integrals by using the definition can be a very lengthy process if we choose larger values for m and n.Therefore, we need a practical and convenient technique for computing double integrals. In other words, we need to learn how to compute double integrals without employing the definition that uses limits and double sums.

The basic idea is that the evaluation becomes easier if we can break a double integral into single integrals by integrating first with respect to one variable and then with respect to the other. The key tool we need is called an iterated integral.

Definitions: Iterated Integrals

Assume a, b, c, and d are real numbers. We define an *iterated integral* for a function $f(x,y)$ over the rectangular region $R=[a,b]\times[c,d]$ as

$$\int_a^b \int_c^d f(x,y)\,dy\,dx = \int_a^b \left[\int_c^d f(x,y)\,dy\right] dx$$

or

$$\int_c^d \int_a^b f(x,y)\,dx\,dy = \int_c^d \left[\int_a^b f(x,y)\,dx\right] dy.$$

The notation $\int_a^b \left[\int_c^d f(x,y)\,dy\right] dx$ means that we integrate $f(x,y)$ with respect to y while holding x constant. Similarly, the notation $\int_c^d \left[\int_a^b f(x,y)\,dx\right] dy$ means that we integrate $f(x,y)$ with respect to x while holding y constant. The fact that double integrals can be split into iterated integrals is expressed in Fubini's theorem. Think of this theorem as an essential tool for evaluating double integrals.

Theorem: Fubini's Theorem

Suppose that $f(x,y)$ is a function of two variables that is continuous over a rectangular region $R=\{(x,y)\in\mathbb{R}^2 \mid a\le x\le b,\ c\le y\le d\}$. Then we see from Figure 14.1.6 that the double integral of f over the region equals an iterated integral,

$$\iint_R f(x,y)\,dA = \iint_R f(x,y)\,dx\,dy = \int_a^b \int_c^d f(x,y)\,dy\,dx = \int_c^d \int_a^b f(x,y)\,dx\,dy.$$

More generally, Fubini's theorem is true if f is bounded on R and f is discontinuous only on a finite number of continuous curves. In other words, f has to be integrable over R.

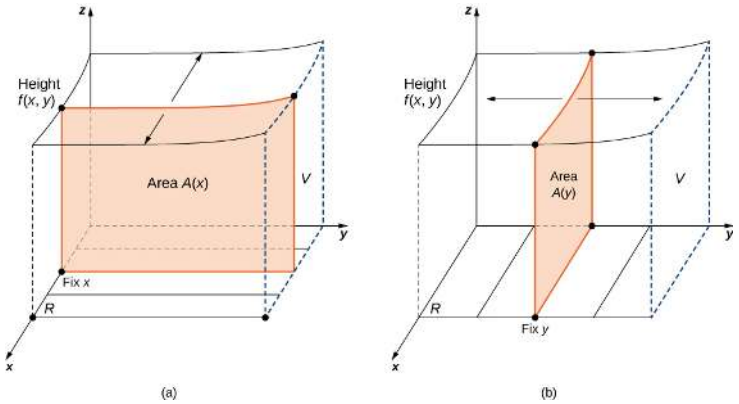


Figure 14.1.6: (a) Integrating first with respect to y and then with respect to x to find the area $A(x)$ and then the volume V; (b) integrating first with respect to x and then with respect to y to find the area $A(y)$ and then the volume V.

Example 14.1.2: Using Fubini's Theorem

Use Fubini's theorem to compute the double integral $\iint_R f(x,y)\,dA$ where $f(x,y)=x$ and $R=[0,2]\times[0,1]$.

Solution

Fubini's theorem offers an easier way to evaluate the double integral by the use of an iterated integral. Note how the boundary values of the region R become the upper and lower limits of integration.

$$\begin{aligned}\iint_R f(x,y)\,dA &= \iint_R f(x,y)\,dx\,dy \\ &= \int_{y=0}^{y=1}\int_{x=0}^{x=2} x\,dx\,dy \\ &= \int_{y=0}^{y=1}\left[\frac{x^2}{2}\Big|_{x=0}^{x=2}\right] dy \\ &= \int_{y=0}^{y=1} 2\,dy = 2y\Big|_{y=0}^{y=1} = 2\end{aligned}$$

The double integration in this example is simple enough to use Fubini's theorem directly, allowing us to convert a double integral into an iterated integral. Consequently, we are now ready to convert all double integrals to iterated integrals and demonstrate how the properties listed earlier can help us evaluate double integrals when the function $f(x,y)$ is more complex. Note that the order of integration can be changed (see Example 7).

Example 14.1.3: Illustrating Properties i and ii

Evaluate the double integral

$$\iint_R (xy - 3xy^2)\,dA, \text{ where } R = \{(x,y) \mid 0 \le x \le 2,\ 1 \le y \le 2\}.$$

Solution

This function has two pieces: one piece is xy and the other is $3xy^2$. Also, the second piece has a constant 3. Notice how we use properties i and ii to help evaluate the double integral.

$$\begin{aligned}\iint_R (xy-3xy^2)\,dA &= \iint_R xy\,dA + \iint_R (-3xy^2)\,dA && \text{Property i: Integral of a sum is the sum of the integrals.} \\ &= \int_{y=1}^{y=2}\int_{x=0}^{x=2} xy\,dx\,dy - \int_{y=1}^{y=2}\int_{x=0}^{x=2} 3xy^2\,dx\,dy && \text{Convert double integrals to iterated integrals.} \\ &= \int_{y=1}^{y=2}\left(\frac{x^2}{2}y\right)\Big|_{x=0}^{x=2} dy - 3\int_{y=1}^{y=2}\left(\frac{x^2}{2}y^2\right)\Big|_{x=0}^{x=2} dy && \text{Integrate with respect to } x\text{, holding } y \text{ constant.} \\ &= \int_{y=1}^{y=2} 2y\,dy - \int_{y=1}^{y=2} 6y^2 dy && \text{Property ii: Placing the constant before the integral.} \\ &= 2\int_1^2 y\,dy - 6\int_1^2 y^2\,dy && \text{Integrate with respect to y.} \\ &= 2\frac{y^2}{2}\Big|_1^2 - 6\frac{y^3}{3}\Big|_1^2 \\ &= y^2\Big|_1^2 - 2y^3\Big|_1^2 \\ &= (4-1) - 2(8-1) = 3 - 2(7) = 3 - 14 = -11.\end{aligned}$$

Example 14.1.4: Illustrating Property v.

Over the region $R = \{(x,y) \mid 1 \le x \le 3,\ 1 \le y \le 2\}$, we have $2 \le x^2 + y^2 \le 13$. Find a lower and an upper bound for the integral $\iint_R (x^2+y^2)\,dA$.

Solution

For a lower bound, integrate the constant function 2 over the region R. For an upper bound, integrate the constant function 13 over the region R.

$$\begin{aligned}\int_1^2\int_1^3 2\,dx\,dy &= \int_1^2 [2x\Big|_1^3]\,dy = \int_1^2 2(2)dy = 4y\Big|_1^2 = 4(2-1) = 4 \\ \int_1^2\int_1^3 13dx\,dy &= \int_1^2 [13x\Big|_1^3]\,dy = \int_1^2 13(2)\,dy = 26y\Big|_1^2 = 26(2-1) = 26.\end{aligned}$$

Hence, we obtain $4 \le \iint_R (x^2+y^2)\,dA \le 26$.

Example 14.1.5: Illustrating Property vi

Evaluate the integral $\iint_R e^y \cos x\,dA$ over the region $R = \{(x,y) \mid 0 \le x \le \frac{\pi}{2},\ 0 \le y \le 1\}$.

Solution

This is a great example for property vi because the function $f(x,y)$ is clearly the product of two single-variable functions e^y and $\cos x$. Thus we can split the integral into two parts and then integrate each one as a single-variable integration problem.

$$\begin{aligned}\iint_R e^y \cos x \, dA &= \int_0^1 \int_0^{\pi/2} e^y \cos x \, dx \, dy \\ &= \left(\int_0^1 e^y dy\right)\left(\int_0^{\pi/2} \cos x \, dx\right) \\ &= (e^y\Big|_0^1)(\sin x\Big|_0^{\pi/2}) \\ &= e-1.\end{aligned}$$

? Exercise 14.1.2

a. Use the properties of the double integral and Fubini's theorem to evaluate the integral

$$\int_0^1 \int_{-1}^3 (3-x+4y)\, dy\, dx.$$

b. Show that $0 \le \iint_R \sin \pi x \, \cos \pi y \, dA \le \frac{1}{32}$ where $R = \left(0, \frac{1}{4}\right)\left(\frac{1}{4}, \frac{1}{2}\right)$.

Hint

Use properties i. and ii. and evaluate the iterated integral, and then use property v.

Answer

a. 26

b. Answers may vary.

As we mentioned before, when we are using rectangular coordinates, the double integral over a region R denoted by $\iint_R f(x,y)\, dA$ can be written as $\iint_R f(x,y)\, dx\, dy$ or $\iint_R f(x,y)\, dy\, dx$. The next example shows that the results are the same regardless of which order of integration we choose.

✓ Example 14.1.6: Evaluating an Iterated Integral in Two Ways

Let's return to the function $f(x,y) = 3x^2 - y$ from Example 1, this time over the rectangular region $R = [0,2] \times [0,3]$. Use Fubini's theorem to evaluate $\iint_R f(x,y)\, dA$ in two different ways:

a. First integrate with respect to y and then with respect to x;
b. First integrate with respect to x and then with respect to y.

Solution

Figure 14.1.6 shows how the calculation works in two different ways.

a. First integrate with respect to y and then integrate with respect to x:

$$\begin{aligned}\iint_R f(x,y)\, dA &= \int_{x=0}^{x=2} \int_{y=0}^{y=3} (3x^2 - y)\, dy\, dx \\ &= \int_{x=0}^{x=2} \left(\int_{y=0}^{y=3} (3x^2 - y)\, dy\right) dx = \int_{x=0}^{x=2} \left[3x^2 y - \frac{y^2}{2}\Big|_{y=0}^{y=3}\right] dx \\ &= \int_{x=0}^{x=2} \left(9x^2 - \frac{9}{2}\right) dx = 3x^3 - \frac{9}{2}x\Big|_{x=0}^{x=2} = 15.\end{aligned}$$

b. First integrate with respect to x and then integrate with respect to y:

$$\begin{aligned}\iint_R f(x,y)\, dA &= \int_{y=0}^{y=3} \int_{x=0}^{x=2} (3x^2 - y)\, dx\, dy \\ &= \int_{y=0}^{y=3} \left(\int_{x=0}^{x=2} (3x^2 - y)\, dx\right) dy \\ &= \int_{y=0}^{y=3} \left[x^3 - xy\Big|_{x=0}^{x=2}\right] dy \\ &= \int_{y=0}^{y=3} (8 - 2y)\, dy = 8y - y^2\Big|_{y=0}^{y=3} = 15.\end{aligned}$$

Analysis

With either order of integration, the double integral gives us an answer of 15. We might wish to interpret this answer as a volume in cubic units of the solid S below the function $f(x,y)=3x^2-y$ over the region $R=[0,2]\times[0,3]$. However, remember that the interpretation of a double integral as a (non-signed) volume works only when the integrand f is a nonnegative function over the base region R.

? Exercise 14.1.3

Evaluate

$$\int_{y=-3}^{y=2}\int_{x=3}^{x=5}(2-3x^2+y^2)\,dx\,dy.$$

Hint

Use Fubini's theorem.

Answer

$-\frac{1340}{3}$

In the next example we see that it can actually be beneficial to switch the order of integration to make the computation easier. We will come back to this idea several times in this chapter.

✓ Example 14.1.7: Switching the Order of Integration

Consider the double integral $\iint_R x\,\sin(xy)\,dA$ over the region $R=\{(x,y)\mid 0\le x\le\pi,\ 1\le y\le 2\}$ (Figure 14.1.7).

a. Express the double integral in two different ways.
b. Analyze whether evaluating the double integral in one way is easier than the other and why.
c. Evaluate the integral.

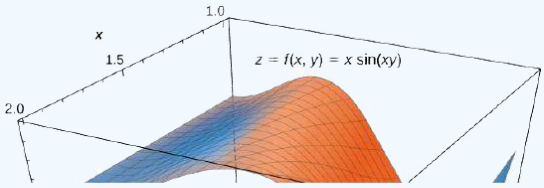


Figure 14.1.7: The function $z=f(x,y)=x\,\sin(xy)$ over the rectangular region $R=[0,\pi]\times[1,2]$.

a. We can express $\iint_R x\,\sin(xy)\,dA$ in the following two ways: first by integrating with respect to y and then with respect to x; second by integrating with respect to x and then with respect to y.

$$\iint_R x\,\sin(xy)\,dA=\int_{x=0}^{x=\pi}\int_{y=1}^{y=2}x\,\sin(xy)\,dy\,dx$$

Integrate first with respect to y.

$$=\int_{y=1}^{y=2}\int_{x=0}^{x=\pi}x\,\sin(xy)\,dx\,dy$$

Integrate first with respect to x.

b. If we want to integrate with respect to y first and then integrate with respect to x, we see that we can use the substitution $u=xy$, which gives $du=x\,dy$. Hence the inner integral is simply $\int\sin u\,du$ and we can change the limits to be functions of x,

$$\iint_R x\,\sin(xy)\,dA=\int_{x=0}^{x=\pi}\int_{y=1}^{y=2}x\,\sin(xy)\,dy\,dx=\int_{x=0}^{x=\pi}\left[\int_{u=x}^{u=2x}\sin(u)\,du\right]dx.$$

However, integrating with respect to x first and then integrating with respect to y requires integration by parts for the inner integral, with $u=x$ and $dv=\sin(xy)dx$

Then $du = dx$ and $v = -\frac{\cos(xy)}{y}$, so

$$\iint_R x\sin(xy)\,dA = \int_{y=1}^{y=2}\int_{x=0}^{x=\pi} x\sin(xy)\,dx\,dy = \int_{y=1}^{y=2}\left[-\frac{x\,\cos(xy)}{y}\Big|_{x=0}^{x=\pi} + \frac{1}{y}\int_{x=0}^{x=\pi}\cos(xy)\,dx\right]dy.$$

Since the evaluation is getting complicated, we will only do the computation that is easier to do, which is clearly the first method.

c. Evaluate the double integral using the easier way.

$$\begin{aligned}\iint_R x\,\sin(xy)\,dA &= \int_{x=0}^{x=\pi}\int_{y=1}^{y=2} x\,\sin(xy)\,dy\,dx \\ &= \int_{x=0}^{x=\pi}\left[\int_{u=x}^{u=2x}\sin(u)\,du\right]dx = \int_{x=0}^{x=\pi}\left[-\cos u\Big|_{u=x}^{u=2x}\right]dx \\ &= \int_{x=0}^{x=\pi}(-\cos 2x + \cos x)\,dx \\ &= \left(-\frac{1}{2}\sin 2x + \sin x\right)\Big|_{x=0}^{x=\pi} = 0.\end{aligned}$$

? Exercise 14.1.4

Evaluate the integral $\iint_R xe^{xy}\,dA$ where $R = [0,1]\times[0,\ln 5]$.

Hint

Integrate with respect to y first.

Answer

$\frac{4-\ln 5}{\ln 5}$

Applications of Double Integrals

Double integrals are very useful for finding the area of a region bounded by curves of functions. We describe this situation in more detail in the next section. However, if the region is a rectangular shape, we can find its area by integrating the constant function $f(x,y) = 1$ over the region R.

Definition: Area of a Region R

The area of the region R is given by

$$A(R) = \iint_R 1\,dA.$$

This definition makes sense because using $f(x,y) = 1$ and evaluating the integral make it a product of length and width. Let's check this formula with an example and see how this works.

Example 14.1.8: Finding Area Using a Double Integral

Find the area of the region $R = \{(x,y)\,|\,0 \le x \le 3,\ 0 \le y \le 2\}$ by using a double integral, that is, by integrating 1 over the region R.

Solution

The region is rectangular with length 3 and width 2, so we know that the area is 6. We get the same answer when we use a double integral:

$$A(R) = \int_0^2\int_0^3 1\,dx\,dy = \int_0^2 [x]_0^3\,dy = \int_0^2 3dy = 3\int_0^2 dy = 3y\Big|_0^2 = 3(2) = 6\text{ units}^2.$$

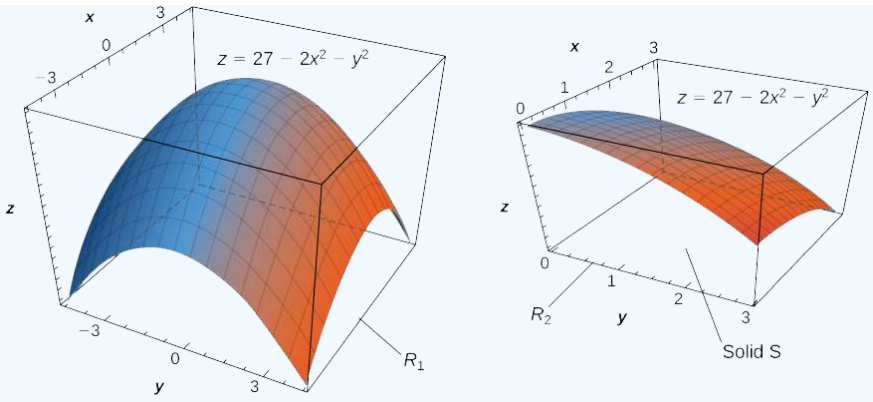
We have already seen how double integrals can be used to find the volume of a solid bounded above by a function $f(x,y) \ge 0$ over a region R provided $f(x,y) \ge 0$ for all (x,y) in R. Here is another example to illustrate this concept.

Example 14.1.9: Volume of an Elliptic Paraboloid

Find the volume V of the solid S that is bounded by the elliptic paraboloid $2x^2 + y^2 + z = 27$, the planes $x = 3$ and $y = 3$, and the three coordinate planes.

Solution

First notice the graph of the surface $z = 27 - 2x^2 - y^2$ in Figure 14.1.8(a) and above the square region $R_1 = [-3,3]\times[-3,3]$. However, we need the volume of the solid bounded by the elliptic paraboloid $2x^2 + y^2 + z = 27$, the planes $x = 3$ and $y = 3$, and the three coordinate planes.



(a) (b)

Figure 14.1.8: (a) The surface $z = 27 - 2x^2 - y^2$ above the square region $R1 = [-3,3] \times [-3,3]$. (b) The solid S lies under the surface $z = 27 - 2x^2 - y^2$ above the square region $R1 = [0,3] \times [0,3]$.

Now let's look at the graph of the surface in Figure 14.1.8(b). We determine the volume V by evaluating the double integral over R_2:

$$\begin{aligned} V &= \iint_R z\,dA = \iint_R (27 - 2x^2 - y^2)\,dA \\ &= \int_{y=0}^{y=3} \int_{x=0}^{x=3} (27 - 2x^2 - y^2)\,dx\,dy && \text{Convert to literal integral.} \\ &= \int_{y=0}^{y=3} \left[27x - \frac{2}{3}x^3 - y^2 x\right]\Big|_{x=0}^{x=3} dy && \text{Integrate with respect to } x. \\ &= \int_{y=0}^{y=3} (63 - 3y^2)dy = 63y - y^3\Big|_{y=0}^{y=3} = 162. \end{aligned}$$

? Exercise 14.1.5

Find the volume of the solid bounded above by the graph of $f(x,y) = xy\sin(x^2 y)$ and below by the xy-plane on the rectangular region $R = [0,1] \times [0,\pi]$.

Hint

Graph the function, set up the integral, and use an iterated integral.

Answer

$\frac{\pi}{2}$

Recall that we defined the average value of a function of one variable on an interval $[a,b]$ as

$$f_{ave} = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

Similarly, we can define the average value of a function of two variables over a region R. The main difference is that we divide by an area instead of the width of an interval.

Definition: Average Value of a Function

The **average value of a function** of two variables over a region R is

$$F_{ave} = \frac{1}{\text{Area of } R}\iint_R f(x,y)\,dx\,dy.$$

In the next example we find the average value of a function over a rectangular region. This is a good example of obtaining useful information for an integration by making individual measurements over a grid, instead of trying to find an algebraic expression for a function.

Example 14.1.10: Calculating Average Storm Rainfall

The weather map in Figure 14.1.9 shows an unusually moist storm system associated with the remnants of Hurricane Karl, which dumped 4–8 inches (100–200 mm) of rain in some parts of the Midwest on September 22–23, 2010. The area of rainfall measured 300 miles east to west and 250 miles north to south. Estimate the average rainfall over the entire area in those two days.

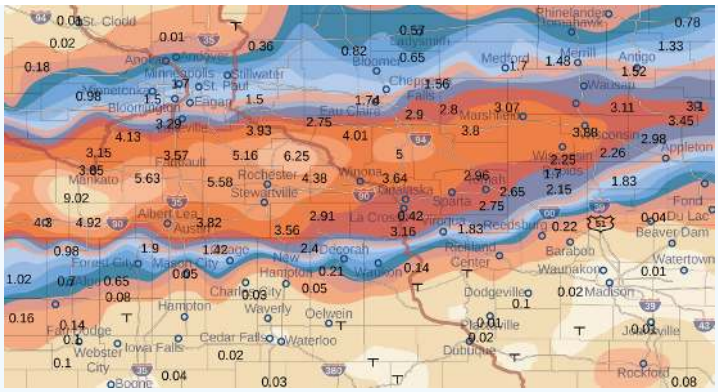


Figure 14.1.9: Effects of Hurricane Karl, which dumped 4–8 inches (100–200 mm) of rain in some parts of southwest Wisconsin, southern Minnesota, and southeast South Dakota over a span of 300 miles east to west and 250 miles north to south.

Solution

Place the origin at the southwest corner of the map so that all the values can be considered as being in the first quadrant and hence all are positive. Now divide the entire map into six rectangles ($m = 2$ and $n = 3$), as shown in Figure 14.1.9. Assume $f(x, y)$ denotes the storm rainfall in inches at a point approximately x miles to the east of the origin and y miles to the north of the origin. Let R represent the entire area of $250 \times 300 = 75000$ square miles. Then the area of each subrectangle is

$$\Delta A = \frac{1}{6}(75000) = 12500.$$

Assume $(x_{ij}*, y_{ij}*)$ are approximately the midpoints of each subrectangle R_{ij}. Note the color-coded region at each of these points, and estimate the rainfall. The rainfall at each of these points can be estimated as:

- At (x_{11}, y_{11}), the rainfall is 0.08.
- At (x_{12}, y_{12}), the rainfall is 0.08.
- At (x_{13}, y_{13}), the rainfall is 0.01.
- At (x_{21}, y_{21}), the rainfall is 1.70.
- At (x_{22}, y_{22}), the rainfall is 1.74.
- At (x_{23}, y_{23}), the rainfall is 3.00.

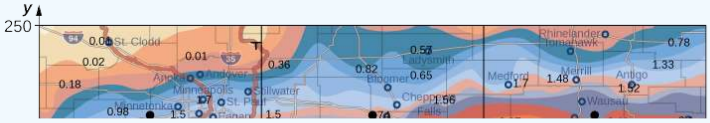


Figure 14.1.10: Storm rainfall with rectangular axes and showing the midpoints of each subrectangle.

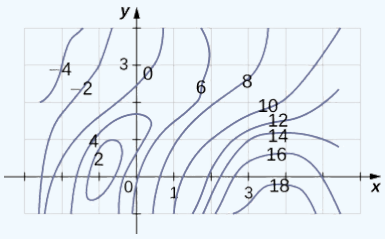
According to our definition, the average storm rainfall in the entire area during those two days was

$$
\begin{aligned}
f_{ave} = \frac{1}{Area\, R}\iint_R f(x,y)\,dx\,dy &= \frac{1}{75000}\iint_R f(x,y)\,dx\,dy \\
&\approx \frac{1}{75000}\sum_{i=1}^{3}\sum_{j=1}^{2} f(x_{ij}^*, y_{ij}^*)\Delta A \\
&= \frac{1}{75000}\left[f(x_{11}^*, y_{11}^*)\Delta A + f(x_{12}^*, y_{12}^*)\Delta A + f(x_{13}^*, y_{13}^*)\Delta A + f(x_{21}^*, y_{21}^*)\Delta A + f(x_{22}^*, y_{22}^*)\Delta A + f(x_{23}^*, y_{23}^*)\Delta A\right] \\
&\approx \frac{1}{75000}\Big[0.08 + 0.08 + 0.01 + 1.70 + 1.74 + 3.00\Big]\Delta A \\
&= \frac{1}{75000}\Big[0.08 + 0.08 + 0.01 + 1.70 + 1.74 + 3.00\Big]12500 \\
&= \frac{1}{6}\Big[0.08 + 0.08 + 0.01 + 1.70 + 1.74 + 3.00\Big] \\
&\approx 1.10 \text{ in.}
\end{aligned}
$$

During September 22–23, 2010 this area had an average storm rainfall of approximately 1.10 inches.

? Exercise 14.1.6

A contour map is shown for a function $f(x,y)$ on the rectangle $R = [-3,6]\times[-1,4]$.



a. Use the midpoint rule with $m = 3$ and $n = 2$ to estimate the value of $\iint_R f(x,y)\,dA$.

b. Estimate the average value of the function $f(x,y)$.

Hint

Divide the region into six rectangles, and use the contour lines to estimate the values for $f(x,y)$.

Answer

Answers to both parts a. and b. may vary.

Key Concepts

- We can use a double Riemann sum to approximate the volume of a solid bounded above by a function of two variables over a rectangular region. By taking the limit, this becomes a double integral representing the volume of the solid.
- Properties of double integral are useful to simplify computation and find bounds on their values.
- We can use Fubini's theorem to write and evaluate a double integral as an iterated integral.
- Double integrals are used to calculate the area of a region, the volume under a surface, and the average value of a function of two variables over a rectangular region.

Key Equations

- $$\iint_R f(x,y)\,dA = \lim_{m,n\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n} f(x_i j*, y_i j*)\,\Delta A$$
- $$\int_a^b\int_c^d f(x,y)\,dx\,dy = \int_a^b\left[\int_c^d f(x,y)\,dy\right]dx$$

  or

  $$\int_c^d\int_a^b f(x,y)\,dx\,dy = \int_c^d\left[\int_a^b f(x,y)\,dx\right]dy$$
- $$f_{ave} = \frac{1}{\text{Area of } R}\iint_R f(x,y)\,dx\,dy$$

Glossary

double integral

of the function $f(x,y)$ over the region R in the xy-plane is defined as the limit of a double Riemann sum,

$$\iint_R f(x,y)\,dA = \lim_{m,n\to\infty} \sum_{i=1}^{m}\sum_{j=1}^{n} f(x_{ij}^*, y_{ij}^*)\,\Delta A.$$

double Riemann sum

of the function $f(x,y)$ over a rectangular region R is

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f(x_{ij}^*, y_{ij}^*)\,\Delta A,$$

where R is divided into smaller subrectangles R_{ij} and (x_{ij}^*, y_{ij}^*) is an arbitrary point in R_{ij}

Fubini's theorem

if $f(x,y)$ is a function of two variables that is continuous over a rectangular region $R = \{(x,y) \in \mathbb{R}^2 \mid a \le x \le b,\ c \le y \le d\}$, then the double integral of f over the region equals an iterated integral,

$$\iint_R f(x,y)\,dA = \int_a^b \int_c^d f(x,y)\,dx\,dy = \int_c^d \int_a^b f(x,y)\,dx\,dy$$

iterated integral

for a function $f(x,y)$ over the region R is

a. $\displaystyle\int_a^b \int_c^d f(x,y)\,dx\,dy = \int_a^b \left[\int_c^d f(x,y)\,dy\right] dx,$

b. $\displaystyle\int_c^d \int_a^b f(x,y)\,dx\,dy = \int_c^d \left[\int_a^b f(x,y)\,dx\right] dy,$

where a, b, c, and d are any real numbers and $R = [a,b] \times [c,d]$

14.1: Double and Iterated Integrals over Rectangles is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **15.1: Double Integrals over Rectangular Regions** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

14.2: Double Integrals over General Regions

Learning Objectives

- Recognize when a function of two variables is integrable over a general region.
- Evaluate a double integral by computing an iterated integral over a region bounded by two vertical lines and two functions of x, or two horizontal lines and two functions of y.
- Simplify the calculation of an iterated integral by changing the order of integration.
- Use double integrals to calculate the volume of a region between two surfaces or the area of a plane region.
- Solve problems involving double improper integrals.

Previously, we studied the concept of double integrals and examined the tools needed to compute them. We learned techniques and properties to integrate functions of two variables over rectangular regions. We also discussed several applications, such as finding the volume bounded above by a function over a rectangular region, finding area by integration, and calculating the average value of a function of two variables.

In this section we consider double integrals of functions defined over a general bounded region D on the plane. Most of the previous results hold in this situation as well, but some techniques need to be extended to cover this more general case.

General Regions of Integration

An example of a general bounded region D on a plane is shown in Figure 14.2.1. Since D is bounded on the plane, there must exist a rectangular region R on the same plane that encloses the region D that is, a rectangular region R exists such that D is a subset of $R (D \subseteq R)$.

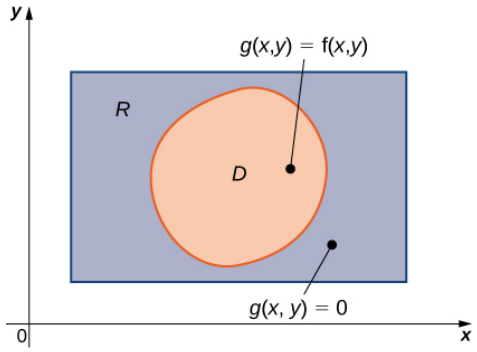


Figure 14.2.1: For a region D that is a subset of R, we can define a function $g(x, y)$ to equal $f(x, y)$ at every point in D and 0 at every point of R not in D.

Suppose $z = f(x, y)$ is defined on a general planar bounded region D as in Figure 14.2.1. In order to develop double integrals of f over D we extend the definition of the function to include all points on the rectangular region R and then use the concepts and tools from the preceding section. But how do we extend the definition of f to include all the points on R? We do this by defining a new function $g(x, y)$ on R as follows:

$$g(x,y) = \begin{cases} f(x,y), & \text{if } (x,y) \text{ is in } D \\ 0, & \text{if } (x,y) \text{ is in } R \text{ but not in } D \end{cases}$$

Note that we might have some technical difficulties if the boundary of D is complicated. So we assume the boundary to be a piecewise smooth and continuous simple closed curve. Also, since all the results developed in the section on Double Integrals over Rectangular Regions used an integrable function $f(x, y)$ we must be careful about $g(x, y)$ and verify that $g(x, y)$ is an integrable function over the rectangular region R. This happens as long as the region D is bounded by simple closed curves. For now we will concentrate on the descriptions of the regions rather than the function and extend our theory appropriately for integration.

We consider two types of planar bounded regions.

Definition: Type I and Type II regions

A region D in the (x, y)-plane is of Type I if it lies between two vertical lines and the graphs of two continuous functions $g_1(x)$ and $g_2(x)$. That is (Figure 14.2.2),

$$D = \{(x,y) \mid a \le x \le b,\ g_1(x) \le y \le g_2(x)\}.$$

A region D in the xy-plane is of Type II if it lies between two horizontal lines and the graphs of two continuous functions $h_1(y)$ and $h_2(y)$. That is (Figure 14.2.3),

$$D = \{(x, y) \mid c \le y \le d,\ h_1(y) \le x \le h_2(y)\}.$$

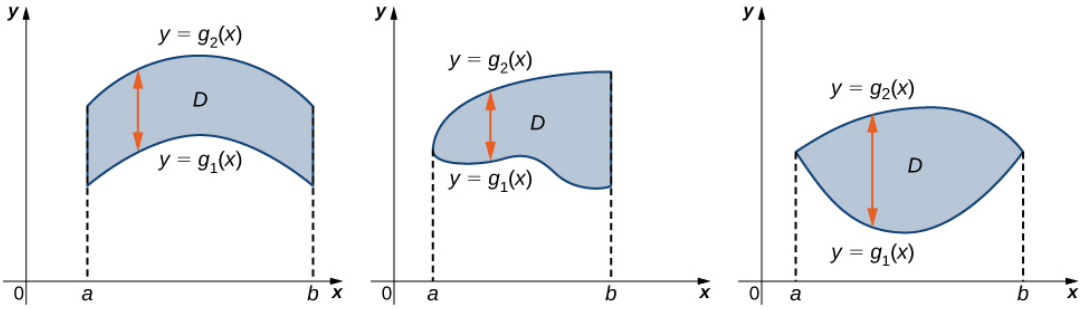


Figure 14.2.2. A Type I region lies between two vertical lines and the graphs of two functions of x.

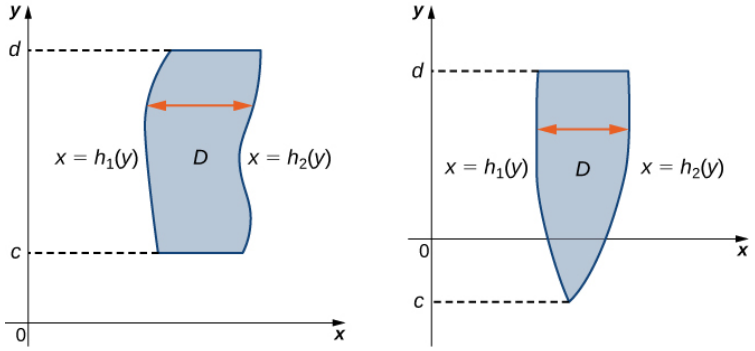


Figure 14.2.3: A Type II region lies between two horizontal lines and the graphs of two functions of y.

Example 14.2.1: Describing a Region as Type I and Also as Type II

Consider the region in the first quadrant between the functions $y = \sqrt{x}$ and $y = x^3$ (Figure 14.2.4). Describe the region first as Type I and then as Type II.

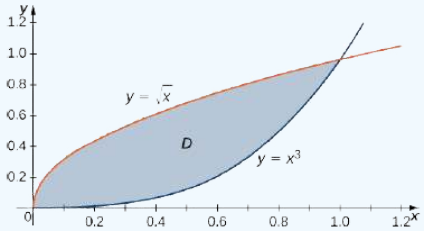


Figure 14.2.4: Region D can be described as Type I or as Type II.

When describing a region as Type I, we need to identify the function that lies above the region and the function that lies below the region. Here, region D is bounded above by $y = \sqrt{x}$ and below by $y = x^3$ in the interval for x in $[0, 1]$. Hence, as Type I, D is described as the set $\{(x, y) \mid 0 \le x \le 1,\ x^3 \le y \le \sqrt[3]{x}\}$.

However, when describing a region as Type II, we need to identify the function that lies on the left of the region and the function that lies on the right of the region. Here, the region D is bounded on the left by $x = y^2$ and on the right by $x = \sqrt[3]{y}$ in the interval for y in $[0, 1]$. Hence, as Type II, D is described as the set $\{(x, y) \mid 0 \le y \le 1,\ y^2 \le x \le \sqrt[3]{y}\}$.

Exercise 14.2.1

Consider the region in the first quadrant between the functions $y = 2x$ and $y = x^2$. Describe the region first as Type I and then as Type II.

Hint

Graph the functions, and draw vertical and horizontal lines.

Answer

Type I and Type II are expressed as $\{(x,y)\,|\,0\le x\le 2,\ x^2\le y\le 2x\}$ and $\{(x,y)|\,0\le y\le 4,\ \frac{1}{2}y\le x\le\sqrt{y}\}$, respectively.

Double Integrals over Non-rectangular Regions

To develop the concept and tools for evaluation of a double integral over a general, nonrectangular region, we need to first understand the region and be able to express it as Type I or Type II or a combination of both. Without understanding the regions, we will not be able to decide the limits of integrations in double integrals. As a first step, let us look at the following theorem.

Theorem: Double Integrals over Nonrectangular Regions

Suppose $g(x,y)$ is the extension to the rectangle R of the function $f(x,y)$ defined on the regions D and R as shown in Figure 14.2.1 inside R. Then $g(x,y)$ is integrable and we define the double integral of $f(x,y)$ over D by

$$\iint_D f(x,y)\,dA=\iint_R g(x,y)\,dA.$$

The right-hand side of this equation is what we have seen before, so this theorem is reasonable because R is a rectangle and $\iint_R g(x,y)dA$ has been discussed in the preceding section. Also, the equality works because the values of $g(x,y)$ are 0 for any point (x,y) that lies outside D and hence these points do not add anything to the integral. However, it is important that the rectangle R contains the region D.

As a matter of fact, if the region D is bounded by smooth curves on a plane and we are able to describe it as Type I or Type II or a mix of both, then we can use the following theorem and not have to find a rectangle R containing the region.

Theorem: Fubini's Theorem (Strong Form)

For a function $f(x,y)$ that is continuous on a region D of Type I, we have

$$\iint_D f(x,y)\,dA=\iint_D f(x,y)\,dy\,dx=\int_a^b\left[\int_{g_1(x)}^{g_2(x)} f(x,y)\,dy\right]dx.$$

Similarly, for a function $f(x,y)$ that is continuous on a region D of Type II, we have

$$\iint_D f(x,y)\,dA=\iint_D f(x,y)\,dx\,dy=\int_c^d\left[\int_{h_1(y)}^{h_2(y)} f(x,y)\,dx\right]dy.$$

The integral in each of these expressions is an iterated integral, similar to those we have seen before. Notice that, in the inner integral in the first expression, we integrate $f(x,y)$ with x being held constant and the limits of integration being $g_1(x)$ and $g_2(x)$. In the inner integral in the second expression, we integrate $f(x,y)$ with y being held constant and the limits of integration are $h_1(x)$ and $h_2(x)$.

Example 14.2.2: Evaluating an Iterated Integral over a Type I Region

Evaluate the integral $\iint_D x^2e^{xy}\,dA$ where D is shown in Figure 14.2.5.

Solution

First construct the region as a Type I region (Figure 14.2.5). Here $D=\{(x,y)\,|\,0\le x\le 2,\ \frac{1}{2}x\le y\le 1\}$. Then we have

$$\iint_D x^2e^{xy}\,dA=\int_{x=0}^{x=2}\int_{y=1/2x}^{y=1} x^2e^{xy}\,dy\,dx.$$

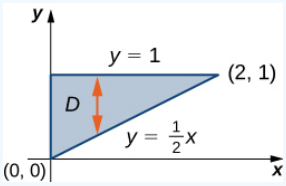


Figure 14.2.5: We can express region D as a Type I region and integrate from $y=\frac{1}{2}x$ to $y=1$ between the lines $x=0$ and $x=2$.

Therefore, we have

$$\int_{x=0}^{x=2}\int_{y=\frac{1}{2}x}^{y=1} x^2e^{xy}\,dy\,dx = \int_{x=0}^{x=2}\left[\int_{y=\frac{1}{2}x}^{y=1} x^2e^{xy}\,dy\right]dx \qquad \text{Iterated integral for a Type I region.}$$

$$= \int_{x=0}^{x=2}\left[x^2\frac{e^{xy}}{x}\right]\Big|_{y=1/2x}^{y=1}dx \qquad \text{Integrate with respect to } y$$

$$= \int_{x=0}^{x=2}\left[xe^x - xe^{x^2/2}\right]dx \qquad \text{Integrate with respect to } x$$

$$= \left[xe^x - e^x - e^{\frac{1}{2}x^2}\right]\Big|_{x=0}^{x=2} = 2.$$

In Example 14.2.2, we could have looked at the region in another way, such as $D = \{(x,y)\,|\,0 \le y \le 1,\ 0 \le x \le 2y\}$ (Figure 14.2.6).

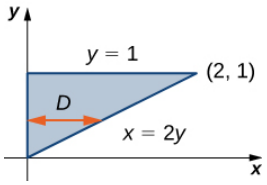


Figure 14.2.6.

This is a Type II region and the integral would then look like

$$\iint_D x^2e^{xy}\,dA = \int_{y=0}^{y=1}\int_{x=0}^{x=2y} x^2e^{xy}\,dx\,dy.$$

However, if we integrate first with respect to x this integral is lengthy to compute because we have to use integration by parts twice.

Example 14.2.3: Evaluating an Iterated Integral over a Type II Region

Evaluate the integral

$$\iint_D (3x^2 + y^2)\,dA$$

where $D = \{(x,y)\,|\,-2 \le y \le 3,\ y^2 - 3 \le x \le y + 3\}$.

Solution

Notice that D can be seen as either a Type I or a Type II region, as shown in Figure 14.2.7. However, in this case describing D as Type I is more complicated than describing it as Type II. Therefore, we use D as a Type II region for the integration.

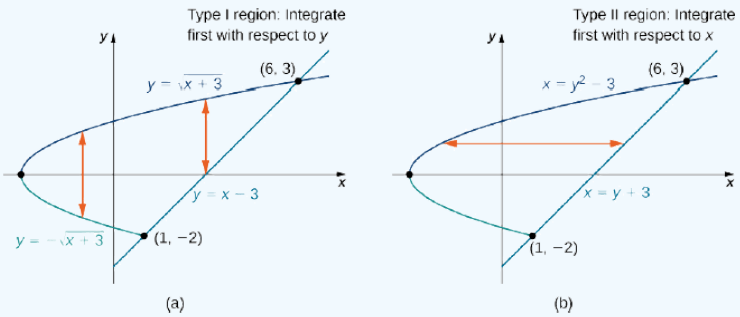


Figure 14.2.7: The region D in this example can be either (a) Type I or (b) Type II.

Choosing this order of integration, we have

$$\begin{aligned}
\iint_D (3x^2+y^2)\,dA &= \int_{y=-2}^{y=3}\int_{x=y^2-3}^{x=y+3}(3x^2+y^2)\,dx\,dy \\
&= \int_{y=-2}^{y=3}(x^3+xy^2)\Big|_{y^2-3}^{y+3}\,dy && \text{Iterated integral, Type II region} \\
&= \int_{y=-2}^{y=3}\left((y+3)^3+(y+3)y^2-(y^2-3)^3-(y^2-3)y^2\right)\,dy \\
&= \int_{-2}^{3}(54+27y-12y^2+2y^3+8y^4-y^6)\,dy && \text{Integrate with respect to } x. \\
&= \left[54y+\frac{27y^2}{2}-4y^3+\frac{y^4}{2}+\frac{8y^5}{5}-\frac{y^7}{7}\right]_{-2}^{3} \\
&= \frac{2375}{7}.
\end{aligned}$$

? Exercise 14.2.2

Sketch the region D and evaluate the iterated integral

$$\iint_D xy\,dy\,dx$$

where D is the region bounded by the curves $y=\cos\ x$ and $y=\sin\ x$ in the interval $[-3\pi/4,\ \pi/4]$

Hint

Express D as a Type I region, and integrate with respect to y first.

Answer

$\frac{\pi}{4}$

Recall from Double Integrals over Rectangular Regions the properties of double integrals. As we have seen from the examples here, all these properties are also valid for a function defined on a non-rectangular bounded region on a plane. In particular, property 3 states:

If $R=S\cup T$ and $S\cap T=0$ except at their boundaries, then

$$\iint_R f(x,y)\,dA=\iint_S f(x,y)\,dA+\iint_T f(x,y)\,dA.$$

Similarly, we have the following property of double integrals over a non-rectangular bounded region on a plane.

Theorem: Decomposing Regions into Smaller Regions

Suppose the region D can be expressed as $D=D_1\cup D_2$ where D_1 and D_2 do not overlap except at their boundaries. Then

$$\iint_D f(x,y)\,dA=\iint_{D_1} f(x,y)\,dA+\iint_{D_2} f(x,y)\,dA.$$

This theorem is particularly useful for non-rectangular regions because it allows us to split a region into a union of regions of Type I and Type II. Then we can compute the double integral on each piece in a convenient way, as in the next example.

Example 14.2.4: Decomposing Regions

Express the region D shown in Figure 14.2.8as a union of regions of Type I or Type II, and evaluate the integral

$$\iint_D (2x+5y)\,dA.$$

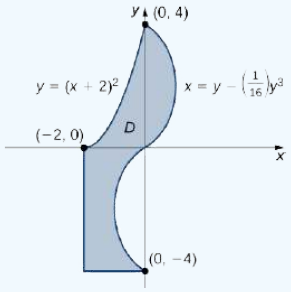


Figure 14.2.8: This region can be decomposed into a union of three regions of Type I or Type II.

Solution

The region D is not easy to decompose into any one type; it is actually a combination of different types. So we can write it as a union of three regions D_1, D_2, and D_3 where, $D_1 = \{(x,y) \mid -2 \le x \le 0,\ 0 \le y \le (x+2)^2\}$, $D_2 = \{(x,y) \mid 0 \le y \le 4,\ 0 \le x \le (y - \frac{1}{16}y^3)\}$, and $D_3 = \{(x,y) \mid -4 \le y \le 0,\ -2 \le x \le (y - \frac{1}{16}y^3)\}$. These regions are illustrated more clearly in Figure 14.2.9.

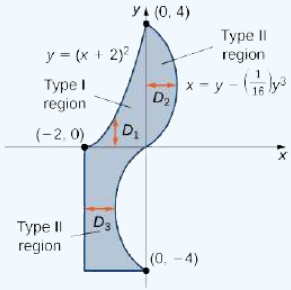


Figure 14.2.9: Breaking the region into three subregions makes it easier to set up the integration.

Here D_1 is Type I and D_2 and D_3 are both of Type II. Hence,

$$\begin{aligned}\iint_D (2x+5y)\,dA &= \iint_{D_1} (2x+5y)\,dA + \iint_{D_2} (2x+5y)\,dA + \iint_{D_3} (2x+5y)\,dA \\ &= \int_{x=-2}^{x=0}\int_{y=0}^{y=(x+2)^2} (2x+5y)\,dy\,dx + \int_{y=0}^{y=4}\int_{x=0}^{x=y-(1/16)y^3} (2x+5y)\,dx\,dy + \int_{y=-4}^{y=0}\int_{x=-2}^{x=y-(1/16)y^3} (2x+5y)\,dx\,dy \\ &= \int_{x=-2}^{x=0} \left[\frac{1}{2}(2+x)^2(20+24x+5x^2)\right] dx + \int_{y=0}^{y=4} \left[\frac{1}{256}y^6 - \frac{7}{16}y^4 + 6y^2\right] dy + \\ &\qquad \int_{y=-4}^{y=0} \left[\frac{1}{256}y^6 - \frac{7}{16}y^4 + 6y^2 + 10y - 4\right] dy \\ &= \frac{40}{3} + \frac{1664}{35} - \frac{1696}{35} = \frac{1304}{105}.\end{aligned}$$

Now we could redo this example using a union of two Type II regions (see Exercise 14.2.4below).

? Exercise 14.2.3

Consider the region bounded by the curves $y = \ln x$ and $y = e^x$ in the interval $[1,2]$. Decompose the region into smaller regions of Type II.

Hint

Sketch the region, and split it into three regions to set it up.

Answer

$$\{(x,y) \mid 0 \le y \le 1,\ 1 \le x \le e^y\} \cup \{(x,y) \mid 1 \le y \le e,\ 1 \le x \le 2\} \cup \{(x,y) \mid e \le y \le e^2,\ \ln y \le x \le 2\}$$

Exercise 14.2.4

Redo Example 14.2.4using a union of two Type II regions.

Hint

$$\{(x,y)\,|\,0\le y\le 4,\ \sqrt{y}-2\le x\le \left(y-\frac{1}{16}y^3\right)\}\cup\{(x,y)\,|\,-4\le y\le 0,\ -2\le x\le \left(y-\frac{1}{16}y^3\right)\}$$

Answer

Same as in the example shown.

Changing the Order of Integration

As we have already seen when we evaluate an iterated integral, sometimes one order of integration leads to a computation that is significantly simpler than the other order of integration. Sometimes the order of integration does not matter, but it is important to learn to recognize when a change in order will simplify our work.

Example 14.2.5: Changing the Order of Integration

Reverse the order of integration in the iterated integral

$$\int_{x=0}^{x=\sqrt{2}}\int_{y=0}^{y=2-x^2} xe^{x^2}\,dy\,dx.$$

Then evaluate the new iterated integral.

Solution

The region as presented is of Type I. To reverse the order of integration, we must first express the region as Type II. Refer to Figure 14.2.10

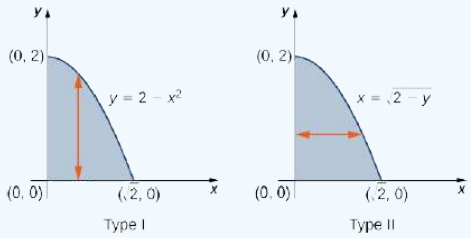


Figure 14.2.10: Converting a region from Type I to Type II.

We can see from the limits of integration that the region is bounded above by $y=2-x^2$ and below by $y=0$ where x is in the interval $[0,\sqrt{2}]$. By reversing the order, we have the region bounded on the left by $x=0$ and on the right by $x=\sqrt{2-y}$ where y is in the interval $[0,2]$. We solved $y=2-x^2$ in terms of x to obtain $x=\sqrt{2-y}$.

Hence

$$\int_0^{\sqrt{2}}\int_0^{2-x^2} xe^{x^2}\,dy\,dx = \int_0^2\int_0^{\sqrt{2-y}} xe^{x^2}\,dx\,dy \qquad \text{Reverse the order of integration then use substitution.}$$

$$= \int_0^2 \left[\frac{1}{2}e^{x^2}\Big|_0^{\sqrt{2-y}}\right]dy = \int_0^2 \frac{1}{2}(e^{2-y}-1)\,dy$$

$$= -\frac{1}{2}(e^{2-y}+y)\Big|_0^2 = \frac{1}{2}(e^2-3).$$

Example 14.2.6: Evaluating an Iterated Integral by Reversing the Order of Integration

Consider the iterated integral

$$\iint_R f(x,y)\,dx\,dy$$

where $z = f(x,y) = x - 2y$ over a triangular region R that has sides on $x = 0,\ y = 0$, and the line $x + y = 1$. Sketch the region, and then evaluate the iterated integral by

a. integrating first with respect to y and then
b. integrating first with respect to x.

Solution

A sketch of the region appears in Figure 14.2.11

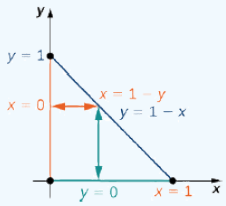


Figure 14.2.11: A triangular region R for integrating in two ways.

We can complete this integration in two different ways.

a. One way to look at it is by first integrating y from $y = 0$ to $y = 1 - x$ vertically and then integrating x from $x = 0$ to $x = 1$:

$$\iint_R f(x,y)\,dx\,dy = \int_{x=0}^{x=1}\int_{y=0}^{y=1-x}(x-2y)\,dy\,dx = \int_{x=0}^{x=1}\left(xy-y^2\right)\Big|_{y=0}^{y=1-x}dx$$

$$= \int_{x=0}^{x=1}\left[x(1-x)-(1-x)^2\right]\,dx = \int_{x=0}^{x=1}\left[-1+3x-2x^2\right]dx = \left[-x+\frac{3}{2}x^2-\frac{2}{3}x^3\right]\Big|_{x=0}^{x=1} = -\frac{1}{6}.$$

b. The other way to do this problem is by first integrating x from $x = 0$ to $x = 1 - y$ horizontally and then integrating y from $y = 0$ to $y = 1$:

$$\iint_R f(x,y)\,dx\,dy = \int_{y=0}^{y=1}\int_{x=0}^{x=1-y}(x-2y)\,dx\,dy = \int_{y=0}^{y=1}\left[\frac{1}{2}x^2-2xy\right]_{x=0}^{x=1-y}dy$$

$$= \int_{y=0}^{y=1}\left[\frac{1}{2}(1-y)^2-2y(1-y)\right]dy = \int_{y=0}^{y=1}\left[\frac{1}{2}-3y+\frac{5}{2}y^2\right]\,dy$$

$$= \left[\frac{1}{2}y-\frac{3}{2}y^2+\frac{5}{6}y^3\right]\Big|_{y=0}^{y=1} = -\frac{1}{6}.$$

? Exercise 14.2.5

Evaluate the iterated integral $\iint_D (x^2+y^2)\,dA$ over the region D in the first quadrant between the functions $y = 2x$ and $y = x^2$.

Evaluate the iterated integral by integrating first with respect to y and then integrating first with resect to x.

Hint

Sketch the region and follow Example 14.2.6

Answer

$\frac{216}{35}$

Calculating Volumes, Areas, and Average Values

We can use double integrals over general regions to compute volumes, areas, and average values. The methods are the same as those in Double Integrals over Rectangular Regions, but without the restriction to a rectangular region, we can now solve a wider variety of problems.

✓ Example 14.2.7: Finding the Volume of a Tetrahedron

Find the volume of the solid bounded by the planes $x = 0,\ y = 0,\ z = 0$, and $2x + 3y + z = 6$.

Solution

The solid is a tetrahedron with the base on the xy-plane and a height $z = 6 - 2x - 3y$. The base is the region D bounded by the lines, $x = 0$, $y = 0$ and $2x + 3y = 6$ where $z = 0$ (Figure 14.2.12). Note that we can consider the region D as Type I or as Type II, and we can integrate in both ways.

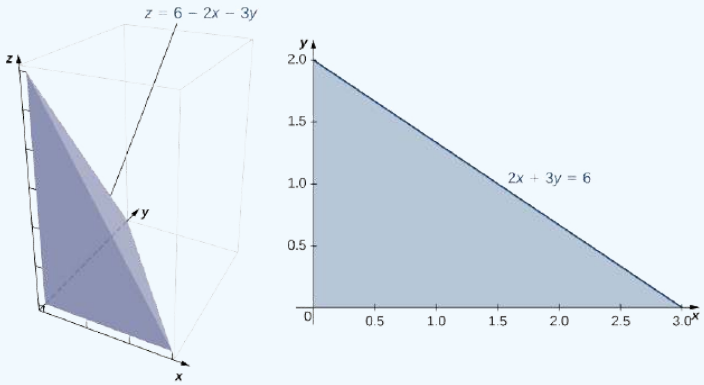


Figure 14.2.12: A tetrahedron consisting of the three coordinate planes and the plane $z = 6 - 2x - 3y$, with the base bound by $x = 0,\ y = 0$, and $2x + 3y = 6$.

First, consider D as a Type I region, and hence $D = \left\{(x, y) \mid 0 \le x \le 3,\ 0 \le y \le 2 - \frac{2}{3}x\right\}$.

Therefore, the volume is

$$V = \int_{x=0}^{x=3} \int_{y=0}^{y=2-(2x/3)} (6 - 2x - 3y)\, dy\, dx = \int_{x=0}^{x=3} \left[\left(6y - 2xy - \frac{3}{2}y^2\right)\Big|_{y=0}^{y=2-(2x/3)} \right] dx$$

$$= \int_{x=0}^{x=3} \left[\frac{2}{3}(x-3)^2\right] dx = 6.$$

Now consider D as a Type II region, so $D = \left\{(x, y) \mid 0 \le y \le 2,\ 0 \le x \le 3 - \frac{3}{2}y\right\}$. In this calculation, the volume is

$$V = \int_{y=0}^{y=2} \int_{x=0}^{x=3-(3y/2)} (6 - 2x - 3y)\, dx\, dy = \int_{y=0}^{y=2} \left[(6x - x^2 - 3xy)\Big|_{x=0}^{x=3-(3y/2)} \right] dy$$

$$= \int_{y=0}^{y=2} \left[\frac{9}{4}(y-2)^2\right] dy = 6.$$

Therefore, the volume is 6 cubic units.

? Exercise 14.2.6

Find the volume of the solid bounded above by $f(x, y) = 10 - 2x + y$ over the region enclosed by the curves $y = 0$ and $y = e^x$ where x is in the interval $[0, 1]$.

Hint

Sketch the region, and describe it as Type I.

Answer

$\frac{e^2}{4} + 10e - \frac{49}{4}$ cubic units

Finding the area of a rectangular region is easy, but finding the area of a non-rectangular region is not so easy. As we have seen, we can use double integrals to find a rectangular area. As a matter of fact, this comes in very handy for finding the area of a general non-rectangular region, as stated in the next definition.

Definition: Double Integrals

The area of a plane-bounded region D is defined as the double integral

$$\iint_D 1\, dA.$$

We have already seen how to find areas in terms of single integration. Here we are seeing another way of finding areas by using double integrals, which can be very useful, as we will see in the later sections of this chapter.

Example 14.2.8: Finding the Area of a Region

Find the area of the region bounded below by the curve $y = x^2$ and above by the line $y = 2x$ in the first quadrant (Figure 14.2.13).

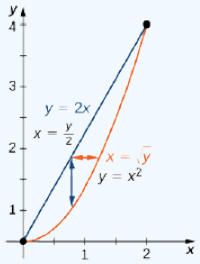


Figure 14.2.13: The region bounded by $y = x^2$ and $y = 2x$.

Solution

We just have to integrate the constant function $f(x, y) = 1$ over the region. Thus, the area A of the bounded region is $\int_{x=0}^{x=2} \int_{y=x^2}^{y=2x} dy\, dx$ or $\int_{y=0}^{y=4} \int_{x=y/2}^{x=\sqrt{y}} dx\, dy$:

$$\begin{aligned} A &= \iint_D 1\, dx\, dy \\ &= \int_{x=0}^{x=2} \int_{y=x^2}^{y=2x} 1\, dy\, dx \\ &= \int_{x=0}^{x=2} \left(y\Big|_{y=x^2}^{y=2x} \right) dx \\ &= \int_{x=0}^{x=2} (2x - x^2)\, dx \\ &= \left(x^2 - \frac{x^3}{3} \right) \Big|_0^2 = \frac{4}{3}. \end{aligned}$$

Exercise 14.2.7

Find the area of a region bounded above by the curve $y = x^3$ and below by $y = 0$ over the interval $[0, 3]$.

Hint

Sketch the region.

Answer

$\frac{81}{4}$ square units

We can also use a double integral to find the average value of a function over a general region. The definition is a direct extension of the earlier formula.

Definition: The Average Value of a Function

If $f(x, y)$ is integrable over a plane-bounded region D with positive area $A(D)$, then the average value of the function is

$$f_{ave} = \frac{1}{A(D)} \iint_D f(x, y)\, dA.$$

Note that the area is $A(D) = \iint_D 1\, dA$.

Example 14.2.9: Finding an Average Value

Find the average value of the function $f(x,y)=7xy^2$ on the region bounded by the line $x=y$ and the curve $x=\sqrt{y}$ (Figure 14.2.14).

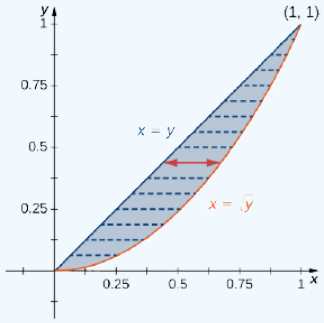


Figure 14.2.14: The region bounded by $x=y$ and $x=\sqrt{y}$.

Solution

First find the area $A(D)$ where the region D is given by the figure. We have

$$A(D)=\iint_D 1\,dA=\int_{y=0}^{y=1}\int_{x=y}^{x=\sqrt{y}} 1\,dx\,dy=\int_{y=0}^{y=1}\left[x\Big|_{x=y}^{x=\sqrt{y}}\right]dy=\int_{y=0}^{y=1}(\sqrt{y}-y)\,dy=\frac{2}{3}y^{2/3}-\frac{y^2}{2}\Big|_0^1=\frac{1}{6}$$

Then the average value of the given function over this region is

$$f_{ave}=\frac{1}{A(D)}\iint_D f(x,y)\,dA=\frac{1}{A(D)}\int_{y=0}^{y=1}\int_{x=y}^{x=\sqrt{y}}7xy^2\,dx\,dy=\frac{1}{1/6}\int_{y=0}^{y=1}\left[\frac{7}{2}x^2y^2\Big|_{x=y}^{x=\sqrt{y}}\right]dy$$

$$=6\int_{y=0}^{y=1}\left[\frac{7}{2}y^2(y-y^2)\right]dy=6\int_{y=0}^{y=1}\left[\frac{7}{2}(y^3-y^4)\right]dy=\frac{42}{2}\left(\frac{y^4}{4}-\frac{y^5}{5}\right)\Big|_0^1=\frac{42}{40}=\frac{21}{20}.$$

Exercise 14.2.8

Find the average value of the function $f(x,y)=xy$ over the triangle with vertices $(0,0),\ (1,0)$ and $(1,3)$.

Hint

Express the line joining $(0,0)$ and $(1,3)$ as a function $y=g(x)$.

Answer

$\frac{3}{4}$

Improper Double Integrals

An **improper double integral** is an integral $\iint_D f\,dA$ where either D is an unbounded region or f is an unbounded function. For example, $D=\{(x,y)\,|\,|x-y|\geq 2\}$ is an unbounded region, and the function $f(x,y)=1/(1-x^2-2y^2)$ over the ellipse $x^2+3y^2\geq 1$ is an unbounded function. Hence, both of the following integrals are improper integrals:

i. $$\iint_D xy\,dA \text{ where } D=\{(x,y)||\,x-y|\geq 2\};$$

ii. $$\iint_D \frac{1}{1-x^2-2y^2}\,dA \text{ where } D=\{(x,y)|\,x^2+3y^2\leq 1\}.$$

In this section we would like to deal with improper integrals of functions over rectangles or simple regions such that f has only finitely many discontinuities. Not all such improper integrals can be evaluated; however, a form of Fubini's theorem does apply for some types of improper integrals.

Theorem: Fubini's Theorem for Improper Integrals

If D is a bounded rectangle or simple region in the plane defined by

$\{(x,y) : a \le x \le b,\ g(x) \le y \le h(x)\}$ and also by

$\{(x,y) : c \le y \le d,\ j(y) \le x \le k(y)\}$ and f is a nonnegative function on D with finitely many discontinuities in the interior of D then

$$\iint_D f\,dA = \int_{x=a}^{x=b}\int_{y=g(x)}^{y=h(x)} f(x,y)\,dy\,dx = \int_{y=c}^{y=d}\int_{x=j(y)}^{x=k(y)} f(x,y)\,dx\,dy$$

It is very important to note that we required that the function be nonnegative on D for the theorem to work. We consider only the case where the function has finitely many discontinuities inside D.

Example 14.2.10: Evaluating a Double Improper Integral

Consider the function $f(x,y) = \frac{e^y}{y}$ over the region $D = \{(x,y) : 0 \le x \le 1,\ x \le y \le \sqrt{x}\}$.

Notice that the function is nonnegative and continuous at all points on D except $(0,0)$. Use Fubini's theorem to evaluate the improper integral.

Solution

First we plot the region D (Figure 14.2.15); then we express it in another way.

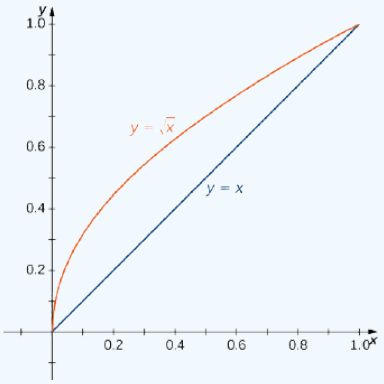


Figure 14.2.15: The function f is continuous at all points of the region D except $(0,0)$.

The other way to express the same region D is

$$D = \{(x,y) : 0 \le y \le 1,\ y^2 \le x \le y\}.$$

Thus we can use Fubini's theorem for improper integrals and evaluate the integral as

$$\int_{y=0}^{y=1}\int_{x=y^2}^{x=y} \frac{e^y}{y}\,dx\,dy.$$

Therefore, we have

$$\int_{y=0}^{y=1}\int_{x=y^2}^{x=y} \frac{e^y}{y}\,dx\,dy = \int_{y=0}^{y=1} \frac{e^y}{y}x\Big|_{x=y^2}^{x=y} dy = \int_{y=0}^{y=1} \frac{e^y}{y}(y-y^2)\,dy = \int_0^1 (e^y - ye^y)\,dy = e-2.$$

As mentioned before, we also have an improper integral if the region of integration is unbounded. Suppose now that the function f is continuous in an unbounded rectangle R.

Theorem: Improper Integrals on an Unbounded Region

If R is an unbounded rectangle such as $R = \{(x,y) : a \le x \le \infty,\ c \le y \le \infty\}$, then when the limit exists, we have

$$\iint_R f(x,y)\,dA = \lim_{(b,d)\to(\infty,\infty)} \int_a^b \left(\int_c^d f(x,y)\,dy\right)dx = \lim_{(b,d)\to(\infty,\infty)} \int_c^d \left(\int_a^b f(x,y)\,dx\right)dy.$$

The following example shows how this theorem can be used in certain cases of improper integrals.

Example 14.2.11

Evaluate the integral $\iint\limits_R xye^{-x^2-y^2}\,dA$ where R is the first quadrant of the plane.

Solution

The region R is the first quadrant of the plane, which is unbounded. So

$$\begin{aligned}\iint\limits_R xye^{-x^2-y^2}\,dA &= \lim_{(b,d)\to(\infty,\infty)}\int_{x=0}^{x=b}\left(\int_{y=0}^{y=d} xye^{-x^2-y^2}\,dy\right)dx \\ &= \lim_{(b,d)\to(\infty,\infty)}\int_{y=0}^{x=b} xye^{-x^2-y^2}\,dy \\ &= \lim_{(b,d)\to(\infty,\infty)}\frac{1}{4}\left(1-e^{-b^2}\right)\left(1-e^{-d^2}\right)=\frac{1}{4}\end{aligned}$$

Thus,

$$\iint\limits_R xye^{-x^2-y^2}\,dA$$

is convergent and the value is $\frac{1}{4}$.

Exercise 14.2.9

$$\iint\limits_D \frac{y}{\sqrt{1-x^2-y^2}}\,dA$$

where $D=\left\{(x,y)\,:\,x\geq 0,\ y\geq 0,\ x^2+y^2\leq 1\right\}$.

Hint

Notice that the integral is nonnegative and discontinuous on $x^2+y^2=1$. Express the region D as $D=\left\{(x,y)\,:\,0\leq x\leq 1,\ 0\leq y\leq\sqrt{1-x^2}\right\}$ and integrate using the method of substitution.

Answer

$\frac{\pi}{4}$

In some situations in probability theory, we can gain insight into a problem when we are able to use double integrals over general regions. Before we go over an example with a double integral, we need to set a few definitions and become familiar with some important properties.

Definition: Joint Density Function

Consider a pair of continuous random variables X and Y such as the birthdays of two people or the number of sunny and rainy days in a month. The **joint density function** f of X and Y satisfies the probability that (X,Y) lies in a certain region D:

$$P((X,Y)\in D)=\iint\limits_D f(x,y)\,dA.$$

Since the probabilities can never be negative and must lie between 0 and 1 the joint density function satisfies the following inequality and equation:

$$f(x,y)\geq 0 \text{ and } \iint\limits_R f(x,y)\,dA=1.$$

Definition: Independent Random Variables

The variables X and Y are said to be **independent random variables** if their joint density function is the product of their individual density functions:

$$f(x,y) = f_1(x)f_2(y).$$

Example 14.2.12: Application to Probability

At Sydney's Restaurant, customers must wait an average of 15 minutes for a table. From the time they are seated until they have finished their meal requires an additional 40 minutes, on average. What is the probability that a customer spends less than an hour and a half at the diner, assuming that waiting for a table and completing the meal are independent events?

Solution

Waiting times are mathematically modeled by exponential density functions, with m being the average waiting time, as

$$f(t) = \begin{cases} 0, & \text{if } t < 0 \\ \dfrac{1}{m}e^{-t/m}, & \text{if } t \geq 0. \end{cases}$$

if X and Y are random variables for 'waiting for a table' and 'completing the meal,' then the probability density functions are, respectively,

$$f_1(x) = \begin{cases} 0, & \text{if } x < 0. \\ \dfrac{1}{15}e^{-x/15}, & \text{if } x \geq 0. \end{cases} \quad \text{and} \quad f_2(y) = \begin{cases} 0, & \text{if } y < 0 \\ \dfrac{1}{40}e^{-y/40}, & \text{if } y \geq 0. \end{cases}$$

Clearly, the events are independent and hence the joint density function is the product of the individual functions

$$f(x,y) = f_1(x)f_2(y) = \begin{cases} 0, & \text{if } x < 0 \text{ or } y < 0, \\ \dfrac{1}{600}e^{-x/15}, & \text{if } x, y \geq 0 \end{cases}$$

We want to find the probability that the combined time $X+Y$ is less than 90 minutes. In terms of geometry, it means that the region D is in the first quadrant bounded by the line $x+y=90$ (Figure 14.2.16).

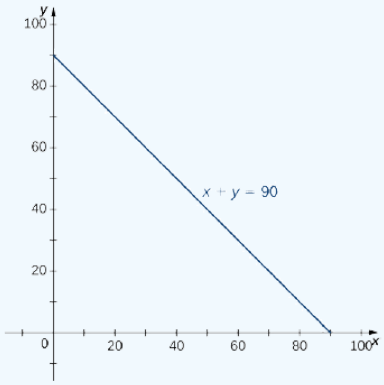


Figure 14.2.16: The region of integration for a joint probability density function.

Hence, the probability that (X,Y) is in the region D is

$$P(X+Y \leq 90) = P((X,Y) \in D) = \iint_D f(x,y)\,dA = \iint_D \frac{1}{600}e^{-x/15}e^{-y/40}\,dA.$$

Since $x+y=90$ is the same as $y=90-x$, we have a region of Type I, so

$$D = \{(x,y) \mid 0 \leq x \leq 90,\ 0 \leq y \leq 90-x\},$$

$$P(X+Y \leq 90) = \frac{1}{600}\int_{x=0}^{x=90}\int_{y=0}^{y=90-x} e^{-(/15}e^{-y/40}dx\,dy = \frac{1}{600}\int_{x=0}^{x=90}\int_{y=0}^{y=90-x} e^{-x/15}e^{-y/40}dx\,dy$$

$$= \frac{1}{600}\int_{x=0}^{x=90}\int_{y=0}^{y=90-x} e^{-(x/15+y/40)}dx\,dy = 0.8328$$

Thus, there is an 83.2% chance that a customer spends less than an hour and a half at the restaurant.

Another important application in probability that can involve improper double integrals is the calculation of expected values. First we define this concept and then show an example of a calculation.

Definition: Expected Values

In probability theory, we denote the **expected values** $E(X)$ and $E(Y)$ respectively, as the most likely outcomes of the events. The expected values $E(X)$ and $E(Y)$ are given by

$$E(X) = \iint_S x\, f(x,y)\, dA \text{ and } E(Y) = \iint_S y\, f(x,y)\, dA,$$

where S is the sample space of the random variables X and Y.

Example 14.2.13: Finding Expected Value

Find the expected time for the events 'waiting for a table' and 'completing the meal' in Example 14.2.12

Solution

Using the first quadrant of the rectangular coordinate plane as the sample space, we have improper integrals for $E(X)$ and $E(Y)$. The expected time for a table is

$$\begin{aligned}
E(X) &= \iint_S x \frac{1}{600} e^{-x/15} e^{-y/40}\, dA \\
&= \frac{1}{600} \int_{x=0}^{x=\infty} \int_{y=0}^{y=\infty} x e^{-x/15} e^{-y/40} dA \\
&= \frac{1}{600} \lim_{(a,b)\to(\infty,\infty)} \int_{x=0}^{x=a} \int_{y=0}^{y=b} x e^{-x/15} e^{-y/40} dx\, dy \\
&= \frac{1}{600} \left(\lim_{a\to\infty} \int_{x=0}^{x=a} x e^{-x/15} dx \right) \left(\lim_{b\to\infty} \int_{y=0}^{y=b} e^{-y/40} dy \right) \\
&= \frac{1}{600} \left(\left(\lim_{a\to\infty} (-15e^{-x/15}(x+15)) \right) \Big|_{x=0}^{x=a} \right) \left(\left(\lim_{b\to\infty} (-40e^{-y/40}) \right) \Big|_{y=0}^{y=b} \right) \\
&= \frac{1}{600} \left(\lim_{a\to\infty} (-15e^{-a/15}(x+15) + 225) \right) \left(\lim_{b\to\infty} (-40e^{-b/40} + 40) \right) \\
&= \frac{1}{600} (225)(40) = 15.
\end{aligned}$$

A similar calculation shows that $E(Y) = 40$. This means that the expected values of the two random events are the average waiting time and the average dining time, respectively.

Exercise 14.2.10

The joint density function for two random variables X and Y is given by

$$f(x,y) = \begin{cases} \frac{1}{600}(x^2+y^2), & \text{if } \le x \le 15,\ 0 \le y \le 10 \\ 0, & \text{otherwise} \end{cases}$$

Find the probability that X is at most 10 and Y is at least 5.

Hint

Compute the probability

$$P(X \le 10,\ Y \ge 5) = \int_{x=-\infty}^{10} \int_{y=5}^{y=10} \frac{1}{6000}(x^2+y^2) dy\, dx.$$

Answer

$\frac{55}{72} \approx 0.7638$

Key Concepts

- A general bounded region D on the plane is a region that can be enclosed inside a rectangular region. We can use this idea to define a double integral over a general bounded region.
- To evaluate an iterated integral of a function over a general nonrectangular region, we sketch the region and express it as a Type I or as a Type II region or as a union of several Type I or Type II regions that overlap only on their boundaries.
- We can use double integrals to find volumes, areas, and average values of a function over general regions, similarly to calculations over rectangular regions.
- We can use Fubini's theorem for improper integrals to evaluate some types of improper integrals.

Key Equations

- **Iterated integral over a Type I region**

$$\iint_D f(x,y)\,dA = \iint_D f(x,y)\,dy\,dx = \int_a^b \left[\int_{g_1(x)}^{g_2(x)} f(x,y)\,dy\right] dx$$

- **Iterated integral over a Type II region**

$$\iint_D f(x,y)\,dA = \iint_D (x,y)\,dx\,dy = \int_c^d \left[\int_{h_1(y)}^{h_2(y)} f(x,y)\,dx\right] dy$$

Glossary

improper double integral
a double integral over an unbounded region or of an unbounded function

Type I
a region D in the xy- plane is Type I if it lies between two vertical lines and the graphs of two continuous functions $g_1(x)$ and $g_2(x)$

Type II
a region D in the xy-plane is Type II if it lies between two horizontal lines and the graphs of two continuous functions $h_1(y)$ and $h_2(h)$

14.2: Double Integrals over General Regions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **15.2: Double Integrals over General Regions** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

14.3: Area by Double Integration

14.3: Area by Double Integration is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

14.4: Double Integrals in Polar Form

Learning Objectives

- Recognize the format of a double integral over a polar rectangular region.
- Evaluate a double integral in polar coordinates by using an iterated integral.
- Recognize the format of a double integral over a general polar region.
- Use double integrals in polar coordinates to calculate areas and volumes.

Double integrals are sometimes much easier to evaluate if we change rectangular coordinates to polar coordinates. However, before we describe how to make this change, we need to establish the concept of a double integral in a polar rectangular region.

Polar Rectangular Regions of Integration

When we defined the double integral for a continuous function in rectangular coordinates—say, g over a region R in the xy-plane—we divided R into subrectangles with sides parallel to the coordinate axes. These sides have either constant x-values and/or constant y-values. In polar coordinates, the shape we work with is a polar rectangle, whose sides have constant r-values and/or constant θ-values. This means we can describe a polar rectangle as in Figure 14.4.1*a*, with $R = \{(r, \theta) \mid a \le r \le b,\ \alpha \le \theta \le \beta\}$.

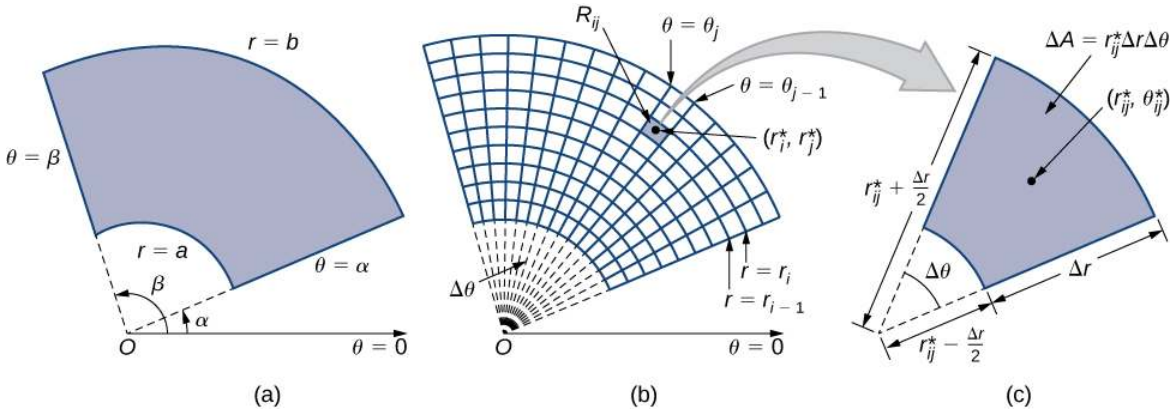


Figure 14.4.1: (a) A polar rectangle R (b) divided into subrectangles R_{ij} (c) Close-up of a subrectangle.

In this section, we are looking to integrate over polar rectangles. Consider a function $f(r, \theta)$ over a polar rectangle R. We divide the interval $[a, b]$ into m subintervals $[r_{i-1}, r_i]$ of length $\Delta r = (b-a)/m$ and divide the interval $[\alpha, \beta]$ into n subintervals $[\theta_{i-1}, \theta_i]$ of width $\Delta\theta = (\beta - \alpha)/n$. This means that the circles $r = r_i$ and rays $\theta = \theta_i$ for $1 \le i \le m$ and $1 \le j \le n$ divide the polar rectangle R into smaller polar subrectangles R_{ij} (Figure 14.4.1*b*).

As before, we need to find the area ΔA of the polar subrectangle R_{ij} and the "polar" volume of the thin box above R_{ij}. Recall that, in a circle of radius r the length s of an arc subtended by a central angle of θ radians is $s = r\theta$. Notice that the polar rectangle R_{ij} looks a lot like a trapezoid with parallel sides $r_{i-1}\Delta\theta$ and $r_i\Delta\theta$ and with a width Δr. Hence the area of the polar subrectangle R_{ij} is

$$\Delta A = \frac{1}{2}\Delta r(r_{i-1}\Delta\theta + r_i\Delta\theta).$$

Simplifying and letting

$$r_{ij}^* = \frac{1}{2}(r_{i-1} + r_i)$$

we have $\Delta A = r_{ij}^* \Delta r \Delta\theta$.

Therefore, the polar volume of the thin box above R_{ij} (Figure 14.4.2) is

$$f(r_{ij}^*, \theta_{ij}^*)\Delta A = f(r_{ij}^*, \theta_{ij}^*) r_{ij}^* \Delta r \Delta\theta.$$

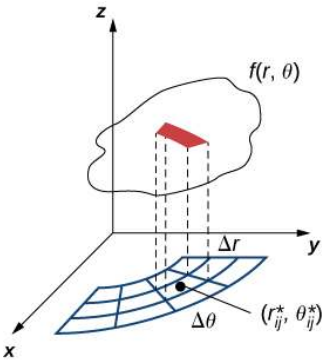


Figure 14.4.2: Finding the volume of the thin box above polar rectangle R_{ij}.

Using the same idea for all the subrectangles and summing the volumes of the rectangular boxes, we obtain a double Riemann sum as

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f(r_{ij}^*, \theta_{ij}^*) r_{ij}^* \Delta r \Delta\theta.$$

As we have seen before, we obtain a better approximation to the polar volume of the solid above the region R when we let m and n become larger. Hence, we define the polar volume as the limit of the double Riemann sum,

$$V = \lim_{m,n\to\infty} \sum_{i=1}^{m}\sum_{j=1}^{n} f(r_{ij}^*, \theta_{ij}^*) r_{ij}^* \Delta r \Delta\theta.$$

This becomes the expression for the double integral.

Definition: The double integral in polar coordinates

The double integral of the function $f(r, \theta)$ over the polar rectangular region R in the $r\theta$-plane is defined as

$$\iint_R f(r,\theta) dA = \lim_{m,n\to\infty} \sum_{i=1}^{m}\sum_{j=1}^{n} f(r_{ij}^*, \theta_{ij}^*) \Delta A \tag{14.4.1}$$

$$= \lim_{m,n\to\infty} \sum_{i=1}^{m}\sum_{j=1}^{n} f(r_{ij}^*, \theta_{ij}^*) r_{ij}^* \Delta r \Delta\theta. \tag{14.4.2}$$

Again, just as in section on Double Integrals over Rectangular Regions, the double integral over a polar rectangular region can be expressed as an iterated integral in polar coordinates. Hence,

$$\iint_R f(r,\theta)\, dA = \iint_R f(r,\theta)\, r\, dr\, d\theta = \int_{\theta=\alpha}^{\theta=\beta} \int_{r=a}^{r=b} f(r,\theta)\, r\, dr\, d\theta.$$

Notice that the expression for dA is replaced by $r\, dr\, d\theta$ when working in polar coordinates. Another way to look at the polar double integral is to change the double integral in rectangular coordinates by substitution. When the function f is given in terms of x and y using $x = r\cos\theta$, $y = r\sin\theta$, and $dA = r\, dr\, d\theta$ changes it to

$$\iint_R f(x,y)\, dA = \iint_R f(r\cos\theta,\, r\sin\theta)\, r\, dr\, d\theta.$$

Note that all the properties listed in section on Double Integrals over Rectangular Regions for the double integral in rectangular coordinates hold true for the double integral in polar coordinates as well, so we can use them without hesitation.

Example 14.4.1A: Sketching a Polar Rectangular Region

Sketch the polar rectangular region

$$R = \{(r,\theta) \mid 1 \le r \le 3, 0 \le \theta \le \pi\}.$$

Solution

As we can see from Figure 14.4.3, $r = 1$ and $r = 3$ are circles of radius 1 and 3 and $0 \le \theta \le \pi$ covers the entire top half of the plane. Hence the region R looks like a semicircular band.

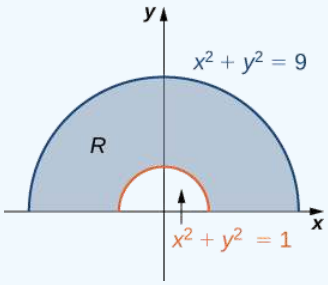


Figure 14.4.3: The polar region R lies between two semicircles.

Now that we have sketched a polar rectangular region, let us demonstrate how to evaluate a double integral over this region by using polar coordinates.

Example 14.4.1B: Evaluating a Double Integral over a Polar Rectangular Region

Evaluate the integral $\iint_R 3x\,dA$ over the region $R = \{(r, \theta) \mid 1 \le r \le 2,\ 0 \le \theta \le \pi\}$.

Solution

First we sketch a figure similar to Figure 14.4.3, but with outer radius $r = 2$. From the figure we can see that we have

$$\begin{aligned}
\iint_R 3x\,dA &= \int_{\theta=0}^{\theta=\pi} \int_{r=1}^{r=2} 3r\cos\theta\, r\,dr\,d\theta \quad \text{Use an integral with correct limits of integration.} \\
&= \int_{\theta=0}^{\theta=\pi} \cos\theta \left[r^3\Big|_{r=1}^{r=2}\right] d\theta \quad \text{Integrate first with respect to } r. \\
&= \int_{\theta=0}^{\theta=\pi} 7\cos\theta\,d\theta \\
&= 7\sin\theta\Big|_{\theta=0}^{\theta=\pi} = 0.
\end{aligned}$$

Exercise 14.4.1

Sketch the region $D = \{(r, \theta) | 1 \le r \le 2,\ -\frac{\pi}{2} \le \theta \le \frac{\pi}{2}\}$, and evaluate $\iint_R x\,dA$.

Hint

Follow the steps in Example 14.4.1A

Answer

$\frac{14}{3}$

Example 14.4.2A: Evaluating a Double Integral by Converting from Rectangular Coordinates

Evaluate the integral

$$\iint_R (1 - x^2 - y^2)\,dA$$

where R is the unit circle on the xy-plane.

Solution

The region R is a unit circle, so we can describe it as $R = \{(r, \theta) \mid 0 \le r \le 1,\ 0 \le \theta \le 2\pi\}$.

Using the conversion $x = r\cos\theta$, $y = r\sin\theta$, and $dA = r\,dr\,d\theta$, we have

$$\begin{aligned}\iint_R (1 - x^2 - y^2)\,dA &= \int_0^{2\pi}\int_0^1 (1 - r^2)\,r\,dr\,d\theta \\ &= \int_0^{2\pi}\int_0^1 (r - r^3)\,dr\,d\theta \\ &= \int_0^{2\pi}\left[\frac{r^2}{2} - \frac{r^4}{4}\right]_0^1 d\theta \\ &= \int_0^{2\pi}\frac{1}{4}\,d\theta = \frac{\pi}{2}.\end{aligned}$$

Example 14.4.2B: Evaluating a Double Integral by Converting from Rectangular Coordinates

Evaluate the integral

$$\iint_R (x + y)\,dA$$

where $R = \{(x, y) \mid 1 \le x^2 + y^2 \le 4,\ x \le 0\}$.

Solution

We can see that R is an annular region that can be converted to polar coordinates and described as $R = \{(r, \theta) \mid 1 \le r \le 2,\ \frac{\pi}{2} \le \theta \le \frac{3\pi}{2}\}$ (see the following graph).

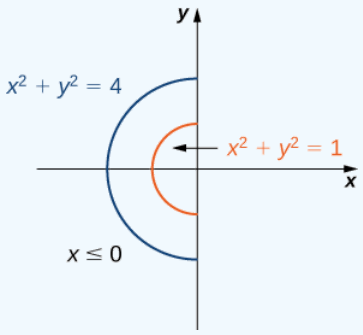


Figure 14.4.4: The annular region of integration R.

Hence, using the conversion $x = r\cos\theta$, $y = r\sin\theta$, and $dA = r\,dr\,d\theta$, we have

$$\begin{aligned}\iint_R (x + y)\,dA &= \int_{\theta=\pi/2}^{\theta=3\pi/2}\int_{r=1}^{r=2} (r\cos\theta + r\sin\theta)r\,dr\,d\theta \\ &= \left(\int_{r=1}^{r=2} r^2\,dr\right)\left(\int_{\pi/2}^{3\pi/2}(\cos\theta + \sin\theta)\,d\theta\right) \\ &= \left[\frac{r^3}{3}\right]_1^2 [\sin\theta - \cos\theta]\Big|_{\pi/2}^{3\pi/2} \\ &= -\frac{14}{3}.\end{aligned}$$

Exercise 14.4.2

Evaluate the integral

$$\iint_R (4 - x^2 - y^2)\,dA$$

where R is the circle of radius 2 on the xy-plane.

Hint

Follow the steps in the previous example.

Answer

8π

General Polar Regions of Integration

To evaluate the double integral of a continuous function by iterated integrals over general polar regions, we consider two types of regions, analogous to Type I and Type II as discussed for rectangular coordinates in section on Double Integrals over General Regions. It is more common to write polar equations as $r = f(\theta)$ than $\theta = f(r)$, so we describe a general polar region as $R = \{(r, \theta) \mid \alpha \leq \theta \leq \beta,\ h_1(\theta) \leq r \leq h_2(\theta)\}$ (Figure 14.4.5).

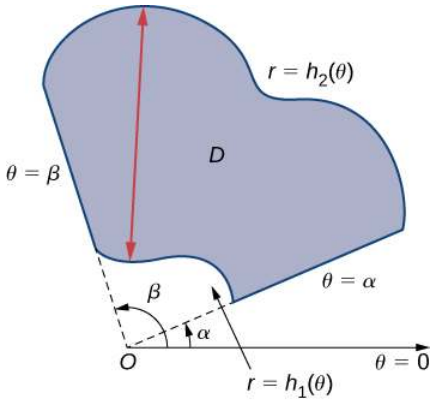


Figure 14.4.5: A general polar region between $\alpha \leq \theta \leq \beta$ and $h_1(\theta) \leq r \leq h_2(\theta)$.

Theorem: Double Integrals over General Polar Regions

If $f(r, \theta)$ is continuous on a general polar region D as described above, then

$$\iint_D f(r, \theta)\, r\, dr\, d\theta = \int_{\theta=\alpha}^{\theta=\beta} \int_{r=h_1(\theta)}^{r=h_2(\theta)} f(r, \theta)\, r\, dr\, d\theta.$$

Example 14.4.3: Evaluating a Double Integral over a General Polar Region

Evaluate the integral

$$\iint_D r^2 \sin\theta\, r\, dr\, d\theta$$

where D is the region bounded by the polar axis and the upper half of the cardioid $r = 1 + \cos\theta$.

Solution

We can describe the region D as $\{(r, \theta) \mid 0 \leq \theta \leq \pi,\ 0 \leq r \leq 1 + \cos\theta\}$ as shown in Figure 14.4.6.

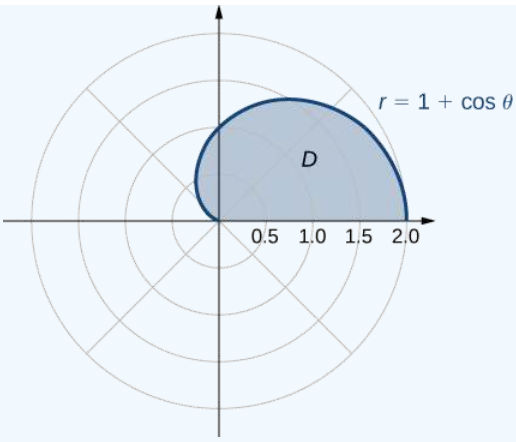


Figure 14.4.6: The region D is the top half of a cardioid.

Hence, we have

$$\begin{aligned}\iint_D r^2 \sin\theta\, r\, dr\, d\theta &= \int_{\theta=0}^{\theta=\pi}\int_{r=0}^{r=1+\cos\theta} (r^2\sin\theta)\, r\, dr\, d\theta \\ &= \frac{1}{4}\int_{\theta=0}^{\theta=\pi} \left[r^4\right]\Big|_{r=0}^{r=1+\cos\ \theta} \sin\theta\, d\theta \\ &= \frac{1}{4}\int_{\theta=0}^{\theta=\pi} (1+\cos\theta)^4 \sin\theta\, d\theta \\ &= -\frac{1}{4}\left[\frac{(1+\cos\theta)^5}{5}\right]_0^{\pi} = \frac{8}{5}.\end{aligned}$$

? Exercise 14.4.3

Evaluate the integral

$$\iint_D r^2 \sin^2 2\theta\, r\, dr\, d\theta$$

where $D = \left\{(r,\theta)\,|\,0 \le \theta \le \pi,\ 0 \le r \le 2\sqrt{\cos 2\theta}\right\}$.

Hint

Graph the region and follow the steps in the previous example.

Answer

$\frac{\pi}{8}$

Polar Areas and Volumes

As in rectangular coordinates, if a solid S is bounded by the surface $z = f(r,\theta)$, as well as by the surfaces $r = a$, $r = b$, $\theta = \alpha$, and $\theta = \beta$, we can find the volume V of S by double integration, as

$$V = \iint_R f(r,\theta)\, r\, dr\, d\theta = \int_{\theta=\alpha}^{\theta=\beta}\int_{r=a}^{r=b} f(r,\theta)\, r\, dr\, d\theta.$$

If the base of the solid can be described as $D = \{(r,\theta)|\alpha \le \theta \le \beta,\ h_1(\theta) \le r \le h_2(\theta)\}$, then the double integral for the volume becomes

$$V = \iint_D f(r,\theta)\, r\, dr\, d\theta = \int_{\theta=\alpha}^{\theta=\beta}\int_{r=h_1(\theta)}^{r=h_2(\theta)} f(r,\theta)\, r\, dr\, d\theta.$$

We illustrate this idea with some examples.

Example 14.4.4A: Finding a Volume Using a Double Integral

Find the volume of the solid that lies under the paraboloid $z = 1 - x^2 - y^2$ and above the unit circle on the xy-plane (Figure 14.4.7).

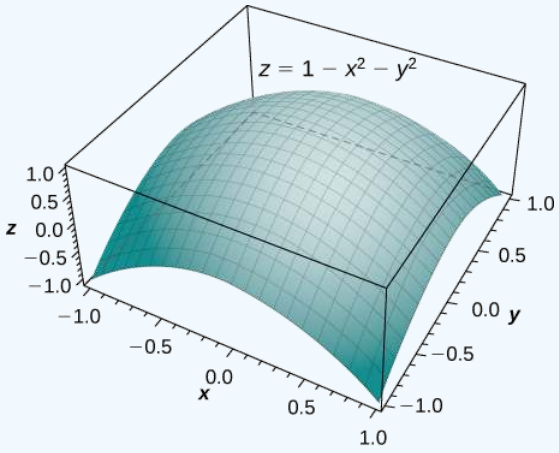


Figure 14.4.7: Finding the volume of a solid under a paraboloid and above the unit circle.

Solution

By the method of double integration, we can see that the volume is the iterated integral of the form

$$\iint_R (1 - x^2 - y^2)\, dA$$

where $R = \{(r, \theta) \mid 0 \le r \le 1,\ 0 \le \theta \le 2\pi\}$.

This integration was shown before in Example 14.4.2A, so the volume is $\frac{\pi}{2}$ cubic units.

Example 14.4.4B: Finding a Volume Using Double Integration

Find the volume of the solid that lies under the paraboloid $z = 4 - x^2 - y^2$ and above the disk $(x-1)^2 + y^2 = 1$ on the xy-plane. See the paraboloid in Figure 14.4.8intersecting the cylinder $(x-1)^2 + y^2 = 1$ above the xy-plane.

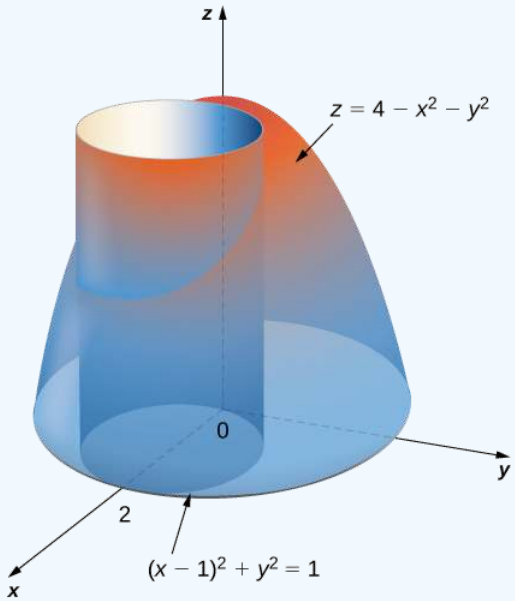


Figure 14.4.8: Finding the volume of a solid with a paraboloid cap and a circular base.

Solution

First change the disk $(x-1)^2+y^2=1$ to polar coordinates. Expanding the square term, we have $x^2-2x+1+y^2=1$. Then simplify to get $x^2+y^2=2x$, which in polar coordinates becomes $r^2=2r\cos\theta$ and then either $r=0$ or $r=2\cos\theta$. Similarly, the equation of the paraboloid changes to $z=4-r^2$. Therefore we can describe the disk $(x-1)^2+y^2=1$ on the xy-plane as the region

$$D=\{(r,\theta)\,|\,0\le\theta\le\pi,\ 0\le r\le 2\cos\theta\}.$$

Hence the volume of the solid bounded above by the paraboloid $z=4-x^2-y^2$ and below by $r=2\cos\theta$ is

$$\begin{aligned}V&=\iint_D f(r,\theta)\,r\,dr\,d\theta\\&=\int_{\theta=0}^{\theta=\pi}\int_{r=0}^{r=2\cos\theta}(4-r^2)\,r\,dr\,d\theta\\&=\int_{\theta=0}^{\theta=\pi}\left[4\frac{r^2}{2}-\frac{r^4}{4}\Big|_0^{2\cos\theta}\right]d\theta\\&=\int_0^{\pi}[8\cos^2\theta-4\cos^4\theta]\,d\theta\\&=\left[\frac{5}{2}\theta+\frac{5}{2}\sin\theta\cos\theta-\sin\theta\cos^3\theta\right]_0^{\pi}=\frac{5}{2}\pi\text{ units}^3.\end{aligned}$$

Notice in the next example that integration is not always easy with polar coordinates. Complexity of integration depends on the function and also on the region over which we need to perform the integration. If the region has a more natural expression in polar coordinates or if f has a simpler antiderivative in polar coordinates, then the change in polar coordinates is appropriate; otherwise, use rectangular coordinates.

Example 14.4.5A: Finding a Volume Using a Double Integral

Find the volume of the region that lies under the paraboloid $z=x^2+y^2$ and above the triangle enclosed by the lines $y=x$, $x=0$, and $x+y=2$ in the xy-plane.

Solution

First examine the region over which we need to set up the double integral and the accompanying paraboloid.

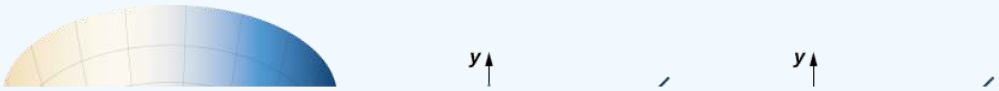


Figure 14.4.9: Finding the volume of a solid under a paraboloid and above a given triangle.

The region D is $\{(x,y)\,|\,0\le x\le 1,\ x\le y\le 2-x\}$. Converting the lines $y=x$, $x=0$, and $x+y=2$ in the xy-plane to functions of r and θ we have $\theta=\pi/4$, $\theta=\pi/2$, and $r=2/(\cos\theta+\sin\theta)$, respectively. Graphing the region on the xy-plane, we see that it looks like $D=\{(r,\theta)\,|\,\pi/4\le\theta\le\pi/2,\ 0\le r\le 2/(\cos\theta+\sin\theta)\}$.

Now converting the equation of the surface gives $z=x^2+y^2=r^2$. Therefore, the volume of the solid is given by the double integral

$$
\begin{aligned}
V &= \iint_D f(r,\theta)\, r\, dr\, d\theta \\
&= \int_{\theta=\pi/4}^{\theta=\pi/2} \int_{r=0}^{r=2/(\cos\ \theta+\sin\ \theta)} r^2 r\, dr d\theta \\
&= \int_{\pi/4}^{\pi/2} \left[\frac{r^4}{4}\right]_0^{2/(\cos\ \theta+\sin\ \theta)} d\theta \\
&= \frac{1}{4}\int_{\pi/4}^{\pi/2} \left(\frac{2}{\cos\theta+\sin\theta}\right)^4 d\theta \\
&= \frac{16}{4}\int_{\pi/4}^{\pi/2} \left(\frac{1}{\cos\theta+\sin\theta}\right)^4 d\theta \\
&= 4\int_{\pi/4}^{\pi/2} \left(\frac{1}{\cos\theta+\sin\theta}\right)^4 d\theta.
\end{aligned}
$$

As you can see, this integral is very complicated. So, we can instead evaluate this double integral in rectangular coordinates as

$$V = \int_0^1 \int_x^{2-x} (x^2+y^2)\, dy\, dx.$$

Evaluating gives

$$
\begin{aligned}
V &= \int_0^1 \int_x^{2-x} (x^2+y^2)\, dy\, dx \\
&= \int_0^1 \left[x^2 y + \frac{y^3}{3}\right]\Big|_x^{2-x} dx \\
&= \int_0^1 \frac{8}{3} - 4x + 4x^2 - \frac{8x^3}{3}\, dx \\
&= \left[\frac{8x}{3} - 2x^2 + \frac{4x^3}{3} - \frac{2x^4}{3}\right]\Big|_0^1 \\
&= \frac{4}{3}\ \text{units}^3.
\end{aligned}
$$

To answer the question of how the formulas for the volumes of different standard solids such as a sphere, a cone, or a cylinder are found, we want to demonstrate an example and find the volume of an arbitrary cone.

Example 14.4.5B: Finding a Volume Using a Double Integral

Use polar coordinates to find the volume inside the cone $z = 2 - \sqrt{x^2+y^2}$ and above the xy-plane.

Solution

The region D for the integration is the base of the cone, which appears to be a circle on the xy-plane (Figure 14.4.10).

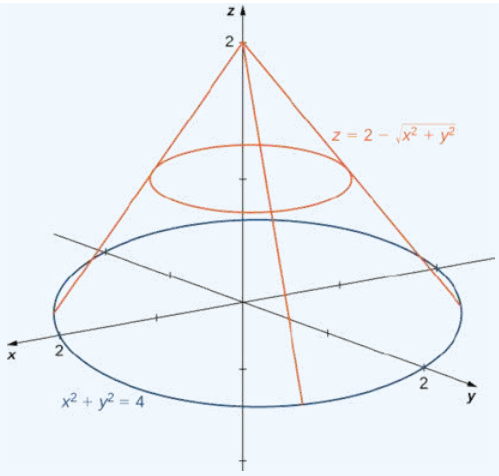


Figure 14.4.10: Finding the volume of a solid inside the cone and above the xy-plane.

We find the equation of the circle by setting $z = 0$:

$$\begin{aligned} 0 &= 2 - \sqrt{x^2 + y^2} \\ 2 &= \sqrt{x^2 + y^2} \\ x^2 + y^2 &= 4. \end{aligned}$$

This means the radius of the circle is 2 so for the integration we have $0 \le \theta \le 2\pi$ and $0 \le r \le 2$. Substituting $x = r\cos\theta$ and $y = r\sin\theta$ in the equation $z = 2 - \sqrt{x^2 + y^2}$ we have $z = 2 - r$. Therefore, the volume of the cone is

$$\int_{\theta=0}^{\theta=2\pi} \int_{r=0}^{r=2} (2 - r)\, r\, dr\, d\theta = 2\pi \frac{4}{3} = \frac{8\pi}{3} \text{ cubic units.}$$

Analysis

Note that if we were to find the volume of an arbitrary cone with radius α units and height h units, then the equation of the cone would be $z = h - \frac{h}{a}\sqrt{x^2 + y^2}$.

We can still use Figure 14.4.10and set up the integral as

$$\int_{\theta=0}^{\theta=2\pi} \int_{r=0}^{r=a} \left(h - \frac{h}{a}r\right) r\, dr\, d\theta.$$

Evaluating the integral, we get $\frac{1}{3}\pi a^2 h$.

? Exercise 14.4.5

Use polar coordinates to find an iterated integral for finding the volume of the solid enclosed by the paraboloids $z = x^2 + y^2$ and $z = 16 - x^2 - y^2$.

Hint

Sketching the graphs can help.

Answer

$$V = \int_0^{2\pi} \int_0^{2\sqrt{2}} (16 - 2r^2)\, r\, dr\, d\theta = 64\pi \text{ cubic units.}$$

As with rectangular coordinates, we can also use polar coordinates to find areas of certain regions using a double integral. As before, we need to understand the region whose area we want to compute. Sketching a graph and identifying the region can be helpful to realize the limits of integration. Generally, the area formula in double integration will look like

$$\text{Area of } A = \int_{\alpha}^{\beta} \int_{h_1(\theta)}^{h_2(\theta)} 1 \, r \, dr \, d\theta.$$

Example 14.4.6A: Finding an Area Using a Double Integral in Polar Coordinates

Evaluate the area bounded by the curve $r = \cos 4\theta$.

Solution

Sketching the graph of the function $r = \cos 4\theta$ reveals that it is a polar rose with eight petals (see the following figure).

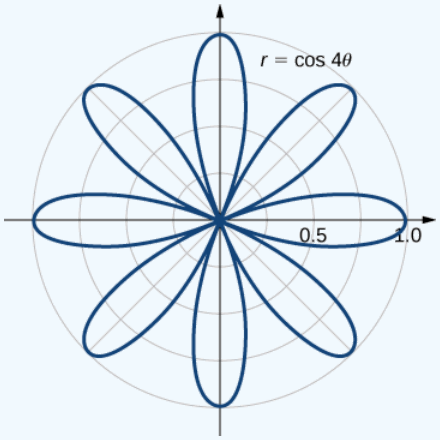


Figure 14.4.11: Finding the area of a polar rose with eight petals.

Using symmetry, we can see that we need to find the area of one petal and then multiply it by 8. Notice that the values of θ for which the graph passes through the origin are the zeros of the function $\cos 4\theta$, and these are odd multiples of $\pi/8$. Thus, one of the petals corresponds to the values of θ in the interval $[-\pi/8, \pi/8]$ Therefore, the area bounded by the curve $r = \cos 4\theta$ is

$$\begin{aligned} A &= 8 \int_{\theta=-\pi/8}^{\theta=\pi/8} \int_{r=0}^{r=\cos 4\theta} 1 \, r \, dr \, d\theta \\ &= 8 \int_{\theta=-\pi/8}^{\theta=\pi/8} \left[\frac{1}{2} r^2 \Big|_0^{\cos 4\theta} \right] d\theta \\ &= 8 \int_{-\pi/8}^{\pi/8} \frac{1}{2} \cos^2 4\theta \, d\theta \\ &= 8 \left[\frac{1}{4}\theta + \frac{1}{16} \sin 4\theta \cos 4\theta \Big|_{-\pi/8}^{\pi/8} \right] \\ &= 8 \left[\frac{\pi}{16} \right] = \frac{\pi}{2} \text{ units}^2. \end{aligned}$$

Example 14.4.6B: Finding Area Between Two Polar Curves

Find the area enclosed by the circle $r = 3 \cos \theta$ and the cardioid $r = 1 + \cos \theta$.

Solution

First and foremost, sketch the graphs of the region (Figure 14.4.12).

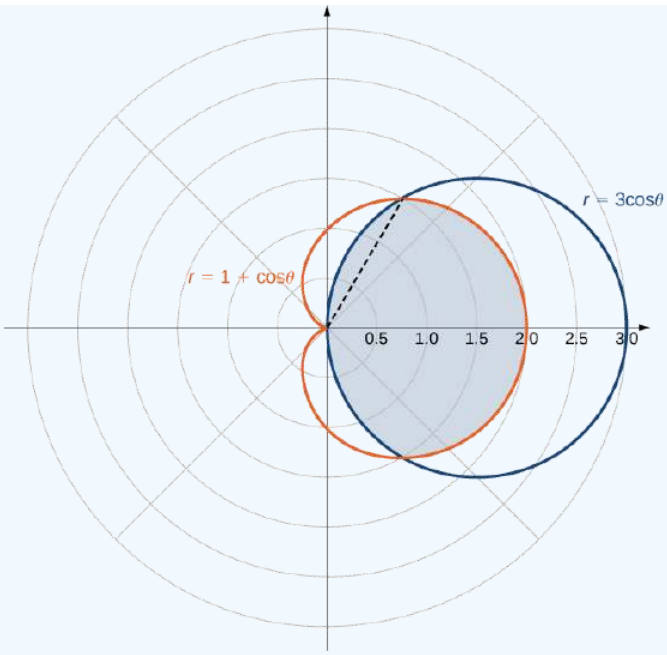


Figure 14.4.12: Finding the area enclosed by both a circle and a cardioid.

We can from see the symmetry of the graph that we need to find the points of intersection. Setting the two equations equal to each other gives

$$3\cos\theta = 1 + \cos\theta.$$

One of the points of intersection is $\theta = \pi/3$. The area above the polar axis consists of two parts, with one part defined by the cardioid from $\theta = 0$ to $\theta = \pi/3$ and the other part defined by the circle from $\theta = \pi/3$ to $\theta = \pi/2$. By symmetry, the total area is twice the area above the polar axis. Thus, we have

$$A = 2\left[\int_{\theta=0}^{\theta=\pi/3}\int_{r=0}^{r=1+\cos\theta} 1\, r\, dr\, d\theta + \int_{\theta=\pi/3}^{\theta=\pi/2}\int_{r=0}^{r=3\cos\theta} 1\, r\, dr\, d\theta\right].$$

Evaluating each piece separately, we find that the area is

$$A = 2\left(\frac{1}{4}\pi + \frac{9}{16}\sqrt{3} + \frac{3}{8}\pi - \frac{9}{16}\sqrt{3}\right) = 2\left(\frac{5}{8}\pi\right) = \frac{5}{4}\pi \text{ square units.}$$

? Exercise 14.4.6

Find the area enclosed inside the cardioid $r = 3 - 3\sin\theta$ and outside the cardioid $r = 1 + \sin\theta$.

Hint

Sketch the graph, and solve for the points of intersection.

Answer

$$A = 2\int_{-\pi/2}^{\pi/6}\int_{1+\sin\theta}^{3-3\sin\theta} r\, dr\, d\theta = (8\pi + 9\sqrt{3}) \text{ units}^2$$

✓ Example 14.4.7: Evaluating an Improper Double Integral in Polar Coordinates

Evaluate the integral

$$\iint_{R^2} e^{-10(x^2+y^2)}\, dx\, dy.$$

Solution

This is an improper integral because we are integrating over an unbounded region R^2. In polar coordinates, the entire plane R^2 can be seen as $0 \le \theta \le 2\pi,\ 0 \le r \le \infty$.

Using the changes of variables from rectangular coordinates to polar coordinates, we have

$$\begin{aligned}\iint_{R^2} e^{-10(x^2+y^2)}\,dx\,dy &= \int_{\theta=0}^{\theta=2\pi}\int_{r=0}^{r=\infty} e^{-10r^2}\,r\,dr\,d\theta = \int_{\theta=0}^{\theta=2\pi}\left(\lim_{a\to\infty}\int_{r=0}^{r=a} e^{-10r^2}\,r\,dr\right)d\theta \\ &= \left(\int_{\theta=0}^{\theta=2\pi}\right)d\theta\left(\lim_{a\to\infty}\int_{r=0}^{r=a} e^{-10r^2}\,r\,dr\right) \\ &= 2\pi\left(\lim_{a\to\infty}\int_{r=0}^{r=a} e^{-10r^2}\,r\,dr\right) \\ &= 2\pi\lim_{a\to\infty}\left(-\frac{1}{20}\right)\left(e^{-10r^2}\Big|_0^a\right) \\ &= 2\pi\left(-\frac{1}{20}\right)\lim_{a\to\infty}\left(e^{-10a^2}-1\right) \\ &= \frac{\pi}{10}.\end{aligned}$$

? Exercise 14.4.7

Evaluate the integral

$$\iint_{R^2} e^{-4(x^2+y^2)}dx\,dy.$$

Hint

Convert to the polar coordinate system.

Answer

$\frac{\pi}{4}$

Key Concepts

- To apply a double integral to a situation with circular symmetry, it is often convenient to use a double integral in polar coordinates. We can apply these double integrals over a polar rectangular region or a general polar region, using an iterated integral similar to those used with rectangular double integrals.
- The area dA in polar coordinates becomes $r\,dr\,d\theta$.
- Use $x = r\cos\theta$, $y = r\sin\theta$, and $dA = r\,dr\,d\theta$ to convert an integral in rectangular coordinates to an integral in polar coordinates.
- Use $r^2 = x^2 + y^2$ and $\theta = tan^{-1}\left(\frac{y}{x}\right)$ to convert an integral in polar coordinates to an integral in rectangular coordinates, if needed.
- To find the volume in polar coordinates bounded above by a surface $z = f(r,\theta)$ over a region on the xy-plane, use a double integral in polar coordinates.

Key Equations

- **Double integral over a polar rectangular region R**

$$\iint_R f(r,\theta)dA = \lim_{m,n\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n} f(r_{ij}^*,\theta_{ij}^*)\Delta A = \lim_{m,n\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n} f(r_{ij}^*,\theta_{ij}^*)r_{ij}^*\Delta r\Delta\theta$$

- **Double integral over a general polar region**

$$\iint_D f(r,\theta)\,r\,dr\,d\theta = \int_{\theta=\alpha}^{\theta=\beta}\int_{r=h_1(\theta)}^{r_2(\theta)} f(r,\theta)\,r\,dr\,d\theta$$

Glossary

polar rectangle

the region enclosed between the circles $r = a$ and $r = b$ and the angles $\theta = \alpha$ and $\theta = \beta$; it is described as $R = \{(r, \theta) \mid a \leq r \leq b,\ \alpha \leq \theta \leq \beta\}$

14.4: Double Integrals in Polar Form is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **15.3: Double Integrals in Polar Coordinates** by Edwin “Jed” Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

14.5: Triple Integrals in Rectangular Coordinates

Learning Objectives

- Recognize when a function of three variables is integrable over a rectangular box.
- Evaluate a triple integral by expressing it as an iterated integral.
- Recognize when a function of three variables is integrable over a closed and bounded region.
- Simplify a calculation by changing the order of integration of a triple integral.
- Calculate the average value of a function of three variables.

Previously, we discussed the double integral of a function $f(x, y)$ of two variables over a rectangular region in the plane. In this section we define the triple integral of a function $f(x, y, z)$ of three variables over a rectangular solid box in space, $\mathbb{R}^3$. Later in this section we extend the definition to more general regions in $\mathbb{R}^3$.

Integrable Functions of Three Variables

We can define a rectangular box B in $\mathbb{R}^3$ as

$$B = \{(x, y, z) \mid a \leq x \leq b,\ c \leq y \leq d,\ e \leq z \leq f\}.$$

We follow a similar procedure to what we did in previously. We divide the interval $[a, b]$ into l subintervals $[x_{i-1}, x_i]$ of equal length Δx with

$$\Delta x = \frac{x_i - x_{i-1}}{l},$$

divide the interval $[c, d]$ into m subintervals $[y_{i-1}, y_i]$ of equal length Δy with

$$\Delta y = \frac{y_j - y_{j-1}}{m},$$

and divide the interval $[e, f]$ into n subintervals $[z_{i-1}, z_i]$ of equal length Δz with

$$\Delta z = \frac{z_k - z_{k-1}}{n}$$

Then the rectangular box B is subdivided into lmn subboxes:

$$B_{ijk} = [x_{i-1}, x_i] \times [y_{i-1}, y_i] \times [z_{i-1}, z_i],$$

as shown in Figure 14.5.1.

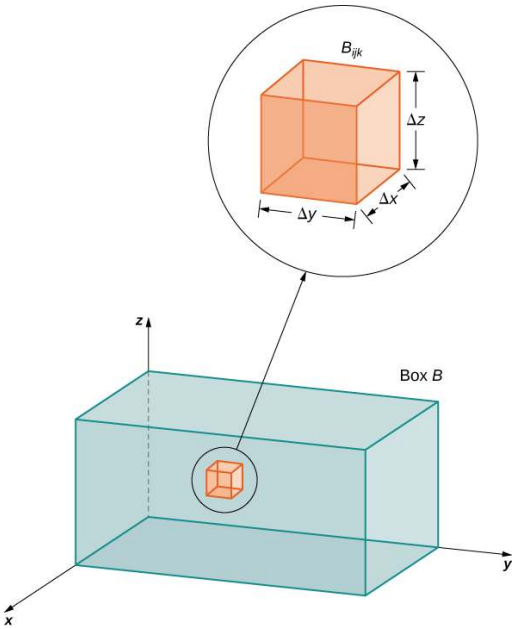


Figure 14.5.1: A rectangular box in $\mathbb{R}^3$ divided into subboxes by planes parallel to the coordinate planes.

For each i, j, and k, consider a sample point $(x^*_{ijk}, y^*_{ijk}, z^*_{ijk})$ in each sub-box B_{ijk}. We see that its volume is $\Delta V = \Delta x \Delta y \Delta z$. Form the triple Riemann sum

$$\sum_{i=1}^{l} \sum_{j=1}^{m} \sum_{k=1}^{n} f(x^*_{ijk}, y^*_{ijk}, z^*_{ijk})\, \Delta x \Delta y \Delta z.$$

We define the triple integral in terms of the limit of a triple Riemann sum, as we did for the double integral in terms of a double Riemann sum.

Definition: The triple integral

The triple integral of a function $f(x,y,z)$ over a rectangular box B is defined as

$$\lim_{l,m,n\to\infty}\sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n} f(x_{ijk}^*, y_{ijk}^*, z_{ijk}^*)\,\Delta x\Delta y\Delta z = \iiint_B f(x,y,z)\,dV$$

if this limit exists.

When the triple integral exists on B the function $f(x,y,z)$ is said to be integrable on B. Also, the triple integral exists if $f(x,y,z)$ is continuous on B. Therefore, we will use continuous functions for our examples. However, continuity is sufficient but not necessary; in other words, f is bounded on B and continuous except possibly on the boundary of B. The sample point $(x_{ijk}^*, y_{ijk}^*, z_{ijk}^*)$ can be any point in the rectangular sub-box B_{ijk} and all the properties of a double integral apply to a triple integral. Just as the double integral has many practical applications, the triple integral also has many applications, which we discuss in later sections.

Now that we have developed the concept of the triple integral, we need to know how to compute it. Just as in the case of the double integral, we can have an iterated triple integral, and consequently, a version of **Fubini's theorem** for triple integrals exists.

Fubini's Theorem for Triple Integrals

If $f(x,y,z)$ is continuous on a rectangular box $B=[a,b]\times[c,d]\times[e,f]$, then

$$\iiint_B f(x,y,z)\,dV = \int_e^f\int_c^d\int_a^b f(x,y,z)\,dx\,dy\,dz.$$

This integral is also equal to any of the other five possible orderings for the iterated triple integral.

For a, b, c, d, e and f real numbers, the iterated triple integral can be expressed in six different orderings:

$$\int_e^f\int_c^d\int_a^b f(x,y,z)\,dx\,dy\,dz = \int_e^f\left(\int_c^d\left(\int_a^b f(x,y,z)\,dx\right)dy\right)dz \tag{14.5.1}$$

$$= \int_c^d\left(\int_e^f\left(\int_a^b f(x,y,z)\,dx\right)dz\right)dy \tag{14.5.2}$$

$$= \int_a^b\left(\int_e^f\left(\int_c^d f(x,y,z)\,dy\right)dz\right)dx \tag{14.5.3}$$

$$= \int_e^f\left(\int_a^b\left(\int_c^d f(x,y,z)\,dy\right)dx\right)dz \tag{14.5.4}$$

$$= \int_c^d\left(\int_a^b\left(\int_c^d f(x,y,z)\,dz\right)dx\right)dy \tag{14.5.5}$$

$$= \int_a^b\left(\int_c^d\left(\int_e^f f(x,y,z)\,dz\right)dy\right)dx \tag{14.5.6}$$

For a rectangular box, the order of integration does not make any significant difference in the level of difficulty in computation. We compute triple integrals using Fubini's Theorem rather than using the Riemann sum definition. We follow the order of integration in the same way as we did for double integrals (that is, from inside to outside).

Example 14.5.1: Evaluating a Triple Integral

Evaluate the triple integral

$$\int_{z=0}^{z=1}\int_{y=2}^{y=4}\int_{x=-1}^{x=5}(x+yz^2)\,dx\,dy\,dz.$$

Solution

The order of integration is specified in the problem, so integrate with respect to x first, then y, and then z.

$$\int_{z=0}^{z=1}\int_{y=2}^{y=4}\int_{x=-1}^{x=5}(x+yz^2)\,dx\,dy\,dz$$

$$=\int_{z=0}^{z=1}\int_{y=2}^{y=4}\left[\frac{x^2}{2}+xyz^2\Big|_{x=-1}^{x=5}\right]dy\,dz \qquad \text{Integrate with respect to } x.$$

$$=\int_{z=0}^{z=1}\int_{y=2}^{y=4}\left[12+6yz^2\right]dy\,dz \qquad \text{Evaluate.}$$

$$=\int_{z=0}^{z=1}\left[12y+6\frac{y^2}{2}z^2\Big|_{y=2}^{y=4}\right]dz \qquad \text{Integrate with respect to } y.$$

$$=\int_{z=0}^{z=1}\left[24+36z^2\right]dz \qquad \text{Evaluate.}$$

$$=\left[24z+36\frac{z^3}{3}\right]_{z=0}^{z=1} \qquad \text{Integrate with respect to } z.$$

$$=36. \qquad \text{Evaluate.}$$

Example 14.5.2: Evaluating a Triple Integral

Evaluate the triple integral

$$\iiint_B x^2yz\,dV$$

where $B=\{(x,y,z)\mid -2\le x\le 1,\ 0\le y\le 3,\ 1\le z\le 5\}$ as shown in Figure 14.5.2

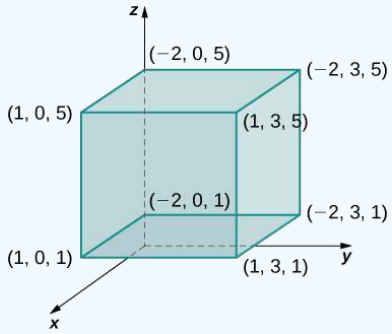


Figure 14.5.2: Evaluating a triple integral over a given rectangular box.

Solution

The order is not specified, but we can use the iterated integral in any order without changing the level of difficulty. Choose, say, to integrate y first, then x, and then z.

$$\begin{aligned}\iiint_B x^2yz\,dV &= \int_1^5\int_{-2}^1\int_0^3 [x^2yz]\,dy\,dx\,dz\\ &=\int_1^5\int_{-2}^1\left[x^2\frac{y^3}{3}z\Big|_0^3\right]dx\,dz\\ &=\int_1^5\int_{-2}^1\frac{y}{2}x^2z\,dx\,dz\\ &=\int_1^5\left[\frac{9}{2}\frac{x^3}{3}z\Big|_{-2}^1\right]dz\\ &=\int_1^5\frac{27}{2}z\,dz\\ &=\frac{27}{2}\frac{z^2}{2}\Big|_1^5=162.\end{aligned}$$

Now try to integrate in a different order just to see that we get the same answer. Choose to integrate with respect to x first, then z, then y

$$\begin{aligned}\iiint_B x^2yz\,dV &= \int_0^3\int_1^5\int_{-2}^1 [x^2yz]\,dx\,dz\,dy \\ &= \int_0^3\int_1^5 \left[\frac{x^3}{3}yz\Big|_{-2}^1\right] dz\,dy \\ &= \int_0^3\int_1^5 3yz\,dz\,dy \\ &= \int_0^3 \left[3y\frac{z^2}{2}\Big|_1^5\right] dy \\ &= \int_0^3 36y\,dy \\ &= 36\frac{y^2}{2}\Big|_0^3 = 18(9-0) = 162.\end{aligned}$$

? Exercise 14.5.1

Evaluate the triple integral

$$\iiint_B z\sin x\,\cos y\,dV$$

where $B = \left\{(x,y,z)\,|\,0\le x\le\pi,\ \frac{3\pi}{2}\le y\le 2\pi,\ 1\le z\le 3\right\}$.

Hint

Follow the steps in the previous example.

Answer

$$\iiint_B z\sin x\,\cos y\,dV = 8$$

Triple Integrals over a General Bounded Region

We now expand the definition of the triple integral to compute a triple integral over a more **general bounded region E** in $\mathbb{R}^3$. The general bounded regions we will consider are of three types. First, let D be the bounded region that is a projection of E onto the xy-plane. Suppose the region E in $\mathbb{R}^3$ has the form

$$E = \left\{(x,y,z)\,|\,(x,y)\in D, u_1(x,y)\le z\le u_2(x,y)\right\}.$$

For two functions $z = u_1(x,y)$ and $u_2(x,y)$, such that $u_1(x,y)\le u_2(x,y)$ for all (x,y) in D as shown in the following figure.

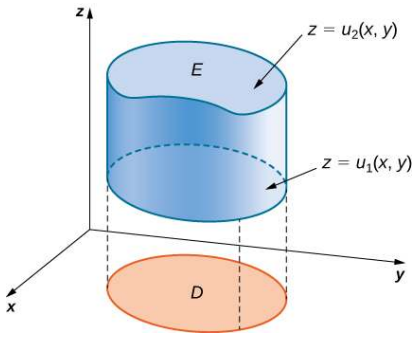


Figure 14.5.3: We can describe region E as the space between $u_1(x,y)$ and $u_2(x,y)$ above the projection D of E onto the xy-plane.

Triple Integral over a General Region

The triple integral of a continuous function $f(x,y,z)$ over a general three-dimensional region

$$E = \left\{(x,y,z)\,|\,(x,y)\in D,\ u_1(x,y)\le z\le u_2(x,y)\right\}$$

in $\mathbb{R}^3$, where D is the projection of E onto the xy-plane, is

$$\iiint_E f(x,y,z)\,dV = \iint_D \left[\int_{u_1(x,y)}^{u_2(x,y)} f(x,y,z)\,dz\right] dA.$$

Similarly, we can consider a general bounded region D in the xy-plane and two functions $y = u_1(x, z)$ and $y = u_2(x, z)$ such that $u_1(x, z) \le u_2(x, z)$ for all (x, z) in D. Then we can describe the solid region E in $\mathbb{R}^3$ as

$$E = \{(x, y, z) \mid (x, z) \in D,\ u_1(x, z) \le z \le u_2(x, z)\}$$

where D is the projection of E onto the xy-plane and the triple integral is

$$\iiint_E f(x, y, z)\, dV = \iint_D \left[\int_{u_1(x,z)}^{u_2(x,z)} f(x, y, z)\, dy \right] dA.$$

Finally, if D is a general bounded region in the xy-plane and we have two functions $x = u_1(y, z)$ and $x = u_2(y, z)$ such that $u_1(y, z) \le u_2(y, z)$ for all (y, z) in D, then the solid region E in $\mathbb{R}^3$ can be described as

$$E = \{(x, y, z) \mid (y, z) \in D,\ u_1(y, z) \le z \le u_2(y, z)\}$$

where D is the projection of E onto the xy-plane and the triple integral is

$$\iiint_E f(x, y, z)\, dV = \iint_D \left[\int_{u_1(y,z)}^{u_2(y,z)} f(x, y, z)\, dx \right] dA.$$

Note that the region D in any of the planes may be of Type I or Type II as described in previously. If D in the xy-plane is of Type I (Figure 14.5.4), then

$$E = \{(x, y, z) \mid a \le x \le b,\ g_1(x) \le y \le g_2(x),\ u_1(x, y) \le z \le u_2(x, y)\}.$$

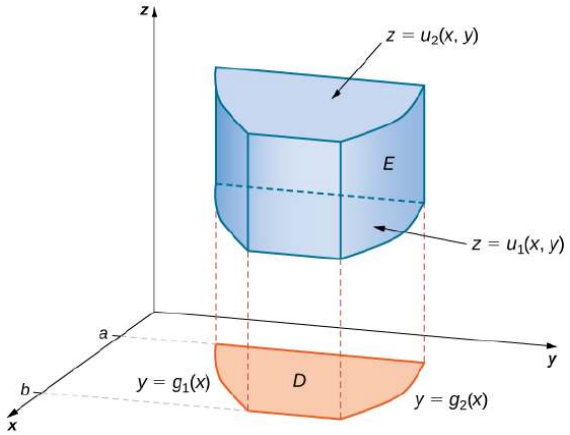


Figure 14.5.4: A box E where the projection D in the xy-plane is of Type I.

Then the triple integral becomes

$$\iiint_E f(x, y, z)\, dV = \int_a^b \int_{g_1(x)}^{g_2(x)} \int_{u_1(x,y)}^{u_2(x,y)} f(x, y, z)\, dz\, dy\, dx.$$

If D in the xy-plane is of Type II (Figure 14.5.5), then

$$E = \{(x, y, z) \mid c \le x \le d, h_1(x) \le y \le h_2(x),\ u_1(x, y) \le z \le u_2(x, y)\}.$$

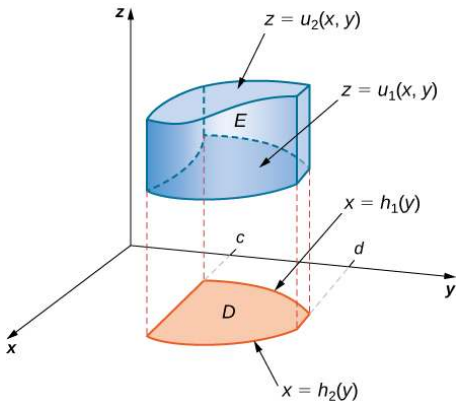


Figure 14.5.5: A box E where the projection D in the xy-plane is of Type II.

Then the triple integral becomes

$$\iiint_E f(x,y,z)\,dV = \int_{y=c}^{y=d}\int_{x=h_1(y)}^{x=h_2(y)}\int_{z=u_1(x,y)}^{z=u_2(x,y)} f(x,y,z)\,dz\,dx\,dy.$$

Example 14.5.3.A: Evaluating a Triple Integral over a General Bounded Region

Evaluate the triple integral of the function $f(x,y,z) = 5x - 3y$ over the solid tetrahedron bounded by the planes $x = 0$, $y = 0$, $z = 0$, and $x + y + z = 1$.

Solution

Figure 14.5.6 shows the solid tetrahedron E and its projection D on the xy-plane.

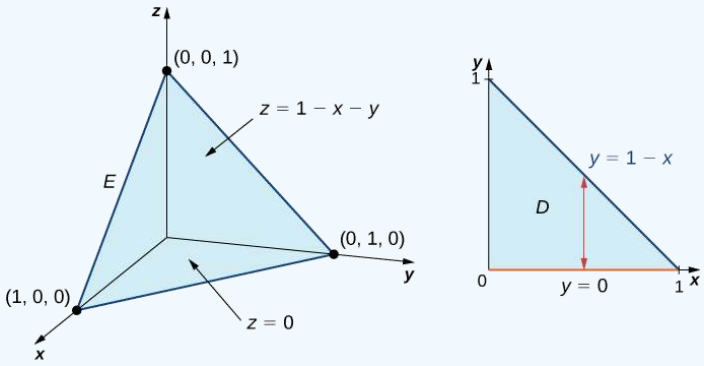


Figure 14.5.6: The solid E has a projection D on the xy-plane of Type I.

We can describe the solid region tetrahedron as

$$E = \{(x,y,z) \mid 0 \le x \le 1,\ 0 \le y \le 1-x,\ 0 \le z \le 1-x-y\}.$$

Hence, the triple integral is

$$\iiint_E f(x,y,z)\,dV = \int_{x=0}^{x=1}\int_{y=0}^{y=1-x}\int_{z=0}^{z=1-x-y} (5x-3y)\,dz\,dy\,dx.$$

To simplify the calculation, first evaluate the integral $\int_{z=0}^{z=1-x-y} (5x-3y)\,dz$. We have

$$\int_{z=0}^{z=1-x-y} (5x-3y)\,dz = (5x-3y)z\Big|_{z=0}^{z=1-x-y} = (5x-3y)(1-x-y).$$

Now evaluate the integral

$$\int_{y=0}^{y=1-x} (5x-3y)(1-x-y)\,dy,$$

obtaining

$$\int_{y=0}^{y=1-x} (5x-3y)(1-x-y)\,dy = \frac{1}{2}(x-1)^2(6x-1).$$

Finally evaluate

$$\int_{x=0}^{x=1} \frac{1}{2}(x-1)^2(6x-1)\,dx = \frac{1}{12}.$$

Putting it all together, we have

$$\iiint_E f(x,y,z)\,dV = \int_{x=0}^{x=1}\int_{y=0}^{y=1-x}\int_{z=0}^{z=1-x-y} (5x-3y)\,dz\,dy\,dx = \frac{1}{12}.$$

Just as we used the double integral

$$\iint_D 1\,dA$$

to find the area of a general bounded region D we can use

$$\iiint_E 1\,dV$$

to find the volume of a general solid bounded region E. The next example illustrates the method.

Example 14.5.3B: Finding a Volume by Evaluating a Triple Integral

Find the volume of a right pyramid that has the square base in the xy-plane $[-1,1]\times[-1,1]$ and vertex at the point $(0,0,1)$ as shown in the following figure.

Figure 14.5.7: Finding the volume of a pyramid with a square base.

Solution

In this pyramid the value of z changes from 0 to 1 and at each height z the cross section of the pyramid for any value of z is the square

$$[-1+z,\,1-z]\times[-1+z,\,1-z].$$

Hence, the volume of the pyramid is

$$\iiint_E 1\,dV$$

where

$$E=\{(x,y,z)\,|\,0\le z\le 1,\,-1+z\le y\le 1-z,\,-1+z\le x\le 1-z\}.$$

Thus, we have

$$\begin{aligned}\iiint_E 1\,dV &= \int_{z=0}^{z=1}\int_{y=-1+z}^{y=1-z}\int_{x=-1+z}^{x=1-z} 1\,dx\,dy\,dz\\ &= \int_{z=0}^{z=1}\int_{y=-1+z}^{y=1-z}(2-2z)\,dy\,dz\\ &= \int_{z=0}^{z=1}(2-2z)^2\,dz=\frac{4}{3}.\end{aligned}$$

Hence, the volume of the pyramid is $\frac{4}{3}$ cubic units.

Exercise 14.5.3

Consider the solid sphere $E=\{(x,y,z)\,|\,x^2+y^2+z^2=9\}$. Write the triple integral

$$\iiint_E f(x,y,z)\,dV$$

for an arbitrary function f as an iterated integral. Then evaluate this triple integral with $f(x,y,z)=1$. Notice that this gives the volume of a sphere using a triple integral.

Hint

Follow the steps in the previous example. Use symmetry.

Answer

$$\begin{aligned}\iiint_E 1\,dV = 8\int_{x=-3}^{x=3}\int_{y=-\sqrt{9-z^2}}^{y=\sqrt{9-z^2}}\int_{z=-\sqrt{9-x^2-y^2}}^{z=\sqrt{9-x^2-y^2}} 1\,dz\,dy\,dx\\ = 36\pi \text{ cubic units.}\end{aligned}$$

Changing the Order of Integration

As we have already seen in double integrals over general bounded regions, changing the order of the integration is done quite often to simplify the computation. With a triple integral over a rectangular box, the order of integration does not change the level of difficulty of the calculation. However, with a triple integral over a general bounded region, choosing an appropriate order of integration can simplify the computation quite a bit. Sometimes making the change to polar coordinates can also be very helpful. We demonstrate two examples here.

Example 14.5.4: Changing the Order of Integration

Consider the iterated integral

$$\int_{x=0}^{x=1}\int_{y=0}^{y=x^2}\int_{z=0}^{z=y} f(x,y,z)\,dz\,dy\,dx.$$

The order of integration here is first with respect to z, then y, and then x. Express this integral by changing the order of integration to be first with respect to x, then z, and then y. Verify that the value of the integral is the same if we let $f(x,y,z)=xyz$.

Solution

The best way to do this is to sketch the region E and its projections onto each of the three coordinate planes. Thus, let

$$E = \{(x, y, z) \mid 0 \le x \le 1,\ 0 \le y \le x^2,\ 0 \le z \le y\}.$$

and

$$\int_{x=0}^{x=1} \int_{y=0}^{y=x^2} \int_{z=0}^{z=x^2} f(x, y, z)\, dz\, dy\, dx = \iiint_E f(x, y, z)\, dV.$$

We need to express this triple integral as

$$\int_{y=c}^{y=d} \int_{z=v_1(y)}^{z=v_2(y)} \int_{x=u_1(y,z)}^{x=u_2(y,z)} f(x, y, z)\, dx\, dz\, dy.$$

Knowing the region E we can draw the following projections (Figure 14.5.8):

on the xy-plane is $D_1 = \{(x, y) \mid 0 \le x \le 1,\ 0 \le y \le x^2\} = \{(x, y) \mid 0 \le y \le 1,\ \sqrt{y} \le x \le 1\}$,

on the yz-plane is $D_2 = \{(y, z) \mid 0 \le y \le 1,\ 0 \le z \le y^2\}$, and

on the xz-plane is $D_3 = \{(x, z) \mid 0 \le x \le 1,\ 0 \le z \le x^2\}$.

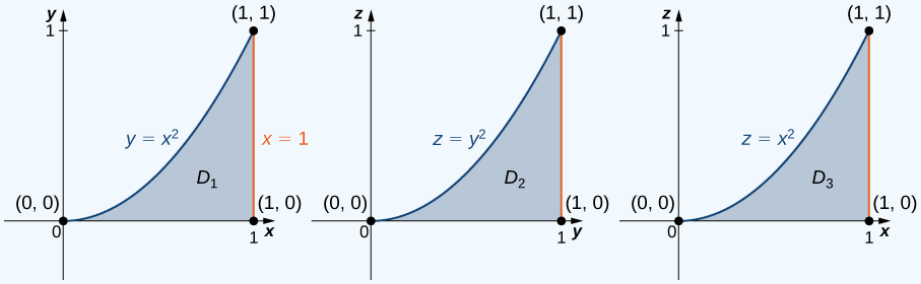


Figure 14.5.8. The three cross sections of E on the three coordinate planes.

Now we can describe the same region E as $\{(x, y, z) \mid 0 \le y \le 1,\ 0 \le z \le y^2,\ \sqrt{y} \le x \le 1\}$, and consequently, the triple integral becomes

$$\int_{y=c}^{y=d} \int_{z=v_1(y)}^{z=v_2(y)} \int_{x=u_1(y,z)}^{x=u_2(y,z)} f(x, y, z)\, dx\, dz\, dy = \int_{y=0}^{y=1} \int_{z=0}^{z=x^2} \int_{x=\sqrt{y}}^{x=1} f(x, y, z)\, dx\, dz\, dy$$

Now assume that $f(x, y, z) = xyz$ in each of the integrals. Then we have

$$\begin{aligned}
\int_{x=0}^{x=1} \int_{y=0}^{y=x^2} \int_{z=0}^{z=y^2} xyz\, dz\, dy\, dx &= \int_{x=0}^{x=1} \int_{y=0}^{y=x^2} \left[xy \frac{z^2}{2} \Big|_{z=0}^{z=y^2} \right] dy\, dx \\
&= \int_{x=0}^{x=1} \int_{y=0}^{y=x^2} \left(x \frac{y^5}{2} \right) dy\, dx \\
&= \int_{x=0}^{x=1} \left[x \frac{y^6}{12} \Big|_{y=0}^{y=x^2} \right] dx \\
&= \int_{x=0}^{x=1} \frac{x^{13}}{12} dx = \frac{x^{14}}{168} \Big|_{x=0}^{x=1} \\
&= \frac{1}{168},
\end{aligned}$$

$$\int_{y=0}^{y=1}\int_{z=0}^{z=y^2}\int_{x=\sqrt{y}}^{x=1} xyz\,dx\,dz\,dy = \int_{y=0}^{y=1}\int_{z=0}^{z=y^2}\left[yz\frac{x^2}{2}\Big|_{\sqrt{y}}^{1}\right]dz\,dy$$
$$= \int_{y=0}^{y=1}\int_{z=0}^{z=y^2}\left(\frac{yz}{2}-\frac{y^2z}{2}\right)dz\,dy$$
$$= \int_{y=0}^{y=1}\left[\frac{yz^2}{4}-\frac{y^2z^2}{4}\Big|_{z=0}^{z=y^2}\right]dy$$
$$= \int_{y=0}^{y=1}\left(\frac{y^5}{4}-\frac{y^6}{4}\right)dy$$
$$= \left(\frac{y^6}{24}-\frac{y^7}{28}\right)\Big|_{y=0}^{y=1}$$
$$= \frac{1}{168}.$$

The answers match.

Exercise 14.5.4

Write five different iterated integrals equal to the given integral

$$\int_{z=0}^{z=4}\int_{y=0}^{y=4-z}\int_{x=0}^{x=\sqrt{y}} f(x,y,z)\,dx\,dy\,dz.$$

Hint

Follow the steps in the previous example, using the region E as $\{(x,y,z)\,|\,0\le z\le 4,\ 0\le y\le 4-z,\ 0\le x\le\sqrt{y}\}$, and describe and sketch the projections onto each of the three planes, five different times.

Answer

$$(i)\ \int_{z=0}^{z=4}\int_{x=0}^{x=\sqrt{4-z}}\int_{y=x^2}^{y=4-z} f(x,y,z)\,dy\,dx\,dz,\ (ii)\ \int_{y=0}^{y=4}\int_{z=0}^{z=4-y}\int_{x=0}^{x=\sqrt{y}} f(x,y,z)\,dx\,dz\,dy,\ (iii)\ \int_{y=0}^{y=4}\int_{x=0}^{x=\sqrt{y}}\int_{z=0}^{Z=4-y} f(x,y,z)\,dz\,dx\,dy,$$
$$(iv)\ \int_{x=0}^{x=2}\int_{y=x^2}^{y=4}\int_{z=0}^{z=4-y} f(x,y,z)\,dz\,dy\,dx,\ (v)\ \int_{x=0}^{x=2}\int_{z=0}^{z=4-x^2}\int_{y=x^2}^{y=4-z} f(x,y,z)\,dy\,dz\,dx$$

Example 14.5.5: Changing Integration Order and Coordinate Systems

Evaluate the triple integral

$$\iiint_E \sqrt{x^2+z^2}\,dV,$$

where E is the region bounded by the paraboloid $y=x^2+z^2$ (Figure 14.5.9) and the plane $y=4$.

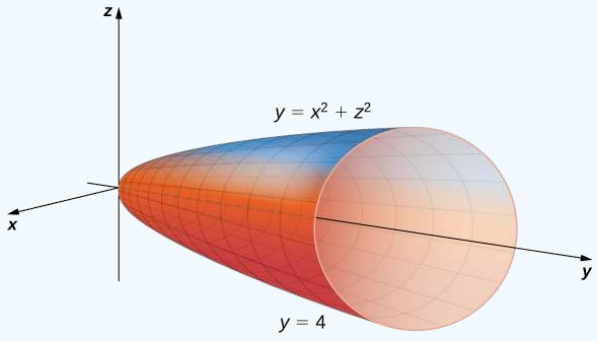


Figure 14.5.9. Integrating a triple integral over a paraboloid.

Solution

The projection of the solid region E onto the xy-plane is the region bounded above by $y=4$ and below by the parabola $y=x^2$ as shown.

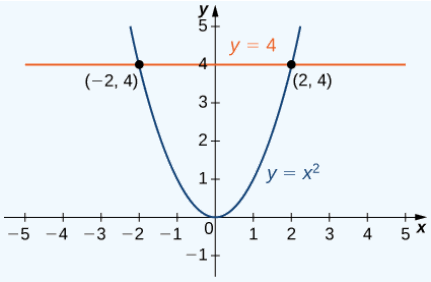


Figure 14.5.10. Cross section in the xy-plane of the paraboloid in Figure 14.5.9.

Thus, we have

$$E = \{(x, y, z) \mid -2 \le x \le 2,\ x^2 \le y \le 4,\ -\sqrt{y - x^2} \le z \le \sqrt{y - x^2}\}.$$

The triple integral becomes

$$\iiint_E \sqrt{x^2 + z^2}\, dV = \int_{x=-2}^{x=2} \int_{y=x^2}^{y=4} \int_{z=-\sqrt{y-x^2}}^{z=\sqrt{y-x^2}} \sqrt{x^2 + z^2}\, dz\, dy\, dx.$$

This expression is difficult to compute, so consider the projection of E onto the xz-plane. This is a circular disc $x^2 + z^2 \le 4$. So we obtain

$$\iiint_E \sqrt{x^2 + z^2}\, dV = \int_{x=-2}^{x=2} \int_{y=x^2}^{y=4} \int_{z=-\sqrt{y-x^2}}^{z=\sqrt{y-x^2}} \sqrt{x^2 + z^2}\, dz\, dy\, dx = \int_{x=-2}^{x=2} \int_{z=-\sqrt{4-x^2}}^{z=\sqrt{4-x^2}} \int_{y=x^2+z^2}^{y=4} \sqrt{x^2 + z^2}\, dy\, dz\, dx.$$

Here the order of integration changes from being first with respect to z then y and then x to being first with respect to y then to z and then to x. It will soon be clear how this change can be beneficial for computation. We have

$$\int_{x=-2}^{x=2} \int_{z=-\sqrt{4-x^2}}^{z=\sqrt{4-x^2}} \int_{y=x^2+z^2}^{y=4} \sqrt{x^2 + z^2}\, dy\, dz\, dx = \int_{x=-2}^{x=2} \int_{z=-\sqrt{4-x^2}}^{z=\sqrt{4-x^2}} (4 - x^2 - z^2)\sqrt{x^2 + z^2}\, dz\, dx.$$

Now use the polar substitution $x = r\cos\theta$, $z = r\sin\theta$, and $dz\, dx = r\, dr\, d\theta$ in the xz-plane. This is essentially the same thing as when we used polar coordinates in the xy-plane, except we are replacing y by z. Consequently the limits of integration change and we have, by using $r^2 = x^2 + z^2$,

$$\int_{x=-2}^{x=2} \int_{z=-\sqrt{4-x^2}}^{z=\sqrt{4-x^2}} (4 - x^2 - z^2)\sqrt{x^2 + z^2}\, dz\, dx = \int_{\theta=0}^{\theta=2\pi} \int_{r=0}^{r=2} (4 - r^2) r r\, dr\, d\theta = \int_0^{2\pi} \left[\frac{4r^3}{3} - \frac{r^5}{5}\Big|_0^2\right] d\theta =$$

$$\int_0^{2\pi} \frac{64}{15}\, d\theta = \frac{128\pi}{15}$$

Average Value of a Function of Three Variables

Recall that we found the average value of a function of two variables by evaluating the double integral over a region on the plane and then dividing by the area of the region. Similarly, we can find the average value of a function in three variables by evaluating the triple integral over a solid region and then dividing by the volume of the solid.

Average Value of a Function of Three Variables

If $f(x, y, z)$ is integrable over a solid bounded region E with positive volume $V(E)$, then the average value of the function is

$$f_{ave} = \frac{1}{V(E)} \iiint_E f(x, y, z)\, dV.$$

Note that the volume is

$$V(E) = \iiint_E 1\, dV.$$

Example 14.5.6: Finding an Average Temperature

The temperature at a point (x, y, z) of a solid E bounded by the coordinate planes and the plane $x + y + z = 1$ is $T(x, y, z) = (xy + 8z + 20)\ ^\circ\text{C}$. Find the average temperature over the solid.

Solution

Use the theorem given above and the triple integral to find the numerator and the denominator. Then do the division. Notice that the plane $x + y + z = 1$ has intercepts $(1, 0, 0)$, $(0, 1, 0)$,and $(0, 0, 1)$. The region E looks like

$$E = \{(x, y, z) \mid 0 \le x \le 1,\ 0 \le y \le 1 - x,\ 0 \le z \le 1 - x - y\}.$$

Hence the triple integral of the temperature is

$$\iiint_E f(x,y,z)\,dV = \int_{x=0}^{x=1}\int_{y=0}^{y=1-x}\int_{z=0}^{z=1-x-y}(xy+8z+20)\,dz\,dy\,dx = \frac{147}{40}.$$

The volume evaluation is

$$V(E) = \iiint_E 1\,dV = \int_{x=0}^{x=1}\int_{y=0}^{y=1-x}\int_{z=0}^{z=1-x-y} 1\,dz\,dy\,dx = \frac{1}{6}.$$

Hence the average value is

$$T_{ave} = \frac{147/40}{1/6} = \frac{6(147)}{40} = \frac{441}{20}\,^{\circ}\mathrm{C}$$

.

? Exercise 14.5.6

Find the average value of the function $f(x,y,z) = xyz$ over the cube with sides of length 4 units in the first octant with one vertex at the origin and edges parallel to the coordinate axes.

Hint

Follow the steps in the previous example.

Answer

$f_{ave} = 8$

Key Concepts

- To compute a triple integral we use Fubini's theorem, which states that if $f(x,y,z)$ is continuous on a rectangular box $B = [a,b]\times[c,d]\times[e,f]$, then

$$\iiint_B f(x,y,z)\,dV = \int_e^f\int_c^d\int_a^b f(x,y,z)\,dx\,dy\,dz$$

  and is also equal to any of the other five possible orderings for the iterated triple integral.
- To compute the volume of a general solid bounded region E we use the triple integral

$$V(E) = \iiint_E 1\,dV.$$

- Interchanging the order of the iterated integrals does not change the answer. As a matter of fact, interchanging the order of integration can help simplify the computation.
- To compute the average value of a function over a general three-dimensional region, we use

$$f_{ave} = \frac{1}{V(E)}\iiint_E f(x,y,z)\,dV.$$

Key Equations

- **Triple integral**

$$\lim_{l,m,n\to\infty}\sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n} f(x^*_{ijk},y^*_{ijk},z^*_{ijk})\,\Delta x\Delta y\Delta z = \iiint_B f(x,y,z)\,dV$$

Glossary

triple integral

the triple integral of a continuous function $f(x,y,z)$ over a rectangular solid box B is the limit of a Riemann sum for a function of three variables, if this limit exists

14.5: Triple Integrals in Rectangular Coordinates is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **15.4: Triple Integrals** by Edwin "Jed" Herman, Gilbert Strang is licensed CC BY-NC-SA 4.0. Original source: https://openstax.org/details/books/calculus-volume-1.

14.6: Moments and Centers of Mass

Using a single integral we were able to compute the center of mass for a one-dimensional object with variable density, and a two dimensional object with constant density. With a double integral we can handle two dimensions and variable density.

Just as before, the coordinates of the center of mass are

$$\bar{x} = \frac{M_y}{M} \qquad \bar{y} = \frac{M_x}{M},$$

where M is the total mass, M_y is the moment around the y-axis, and M_x is the moment around the x-axis. (You may want to review the concepts in Section 9.6.)

The key to the computation, just as before, is the approximation of mass. In the two-dimensional case, we treat density σ as mass per square area, so when density is constant, mass is $(\text{density})(\text{area})$. If we have a two-dimensional region with varying density given by $\sigma(x, y)$, and we divide the region into small subregions with area ΔA, then the mass of one subregion is approximately $\sigma(x_i, y_j)\Delta A$, the total mass is approximately the sum of many of these, and as usual the sum turns into an integral in the limit:

$$M = \int_{x_0}^{x_1} \int_{y_0}^{y_1} \sigma(x, y)\, dy\, dx,$$

and similarly for computations in cylindrical coordinates. Then as before

$$M_x = \int_{x_0}^{x_1} \int_{y_0}^{y_1} y\sigma(x, y)\, dy\, dx$$
$$M_y = \int_{x_0}^{x_1} \int_{y_0}^{y_1} x\sigma(x, y)\, dy\, dx.$$

Example 14.6.1

Find the center of mass of a thin, uniform plate whose shape is the region between $y = \cos x$ and the x-axis between $x = -\pi/2$ and $x = \pi/2$. Since the density is constant, we may take $\sigma(x, y) = 1$.

It is clear that $\bar{x} = 0$, but for practice let's compute it anyway. First we compute the mass:

$$\begin{aligned} M &= \int_{-\pi/2}^{\pi/2} \int_0^{\cos x} 1\, dy\, dx \\ &= \int_{-\pi/2}^{\pi/2} \cos x\, dx \\ &= \sin x \Big|_{-\pi/2}^{\pi/2} = 2. \end{aligned}$$

Next,

$$\begin{aligned} M_x &= \int_{-\pi/2}^{\pi/2} \int_0^{\cos x} y\, dy\, dx \\ &= \int_{-\pi/2}^{\pi/2} \frac{1}{2} \cos^2 x\, dx \\ &= \frac{\pi}{4}. \end{aligned}$$

Finally,

$$M_y = \int_{-\pi/2}^{\pi/2} \int_0^{\cos x} x \, dy \, dx$$
$$= \int_{-\pi/2}^{\pi/2} x \cos x \, dx$$
$$= 0.$$

So $\bar{x} = 0$ as expected, and $\bar{y} = \pi/4/2 = \pi/8$. This is the same problem as in example 9.6.4; it may be helpful to compare the two solutions.

Example 14.6.2

Find the center of mass of a two-dimensional plate that occupies the quarter circle $x^2 + y^2 \leq 1$ in the first quadrant and has density $k(x^2 + y^2)$. It seems clear that because of the symmetry of both the region and the density function (both are important!), $\bar{x} = \bar{y}$. We'll do both to check our work.

Jumping right in:

$$M = \int_0^1 \int_0^{\sqrt{1-x^2}} k(x^2 + y^2) \, dy \, dx$$
$$= k \int_0^1 x^2 \sqrt{1-x^2} + \frac{(1-x^2)^{3/2}}{3} \, dx.$$

This integral is something we can do, but it's a bit unpleasant. Since everything in sight is related to a circle, let's back up and try polar coordinates. Then $x^2 + y^2 = r^2$ and

$$M = \int_0^{\pi/2} \int_0^1 k(r^2) \, r \, dr \, d\theta$$
$$= k \int_0^{\pi/2} \frac{r^4}{4} \Big|_0^1 \, d\theta$$
$$= k \int_0^{\pi/2} \frac{1}{4} \, d\theta$$
$$= k \frac{\pi}{8}.$$

Much better. Next, since $y = r \sin\theta$,

$$M_x = k \int_0^{\pi/2} \int_0^1 r^4 \sin\theta \, dr \, d\theta$$
$$= k \int_0^{\pi/2} \frac{1}{5} \sin\theta \, d\theta$$
$$= k - \frac{1}{5} \cos\theta \Big|_0^{\pi/2} = \frac{k}{5}.$$

Similarly,

$$M_y = k \int_0^{\pi/2} \int_0^1 r^4 \cos\theta \, dr \, d\theta$$
$$= k \int_0^{\pi/2} \frac{1}{5} \cos\theta \, d\theta$$
$$= k \frac{1}{5} \sin\theta \Big|_0^{\pi/2} = \frac{k}{5}.$$

Finally, $\bar{x} = \bar{y} = \frac{8}{5\pi}$.

Contributors

- David Guichard (Whitman College)
- Integrated by Justin Marshall.

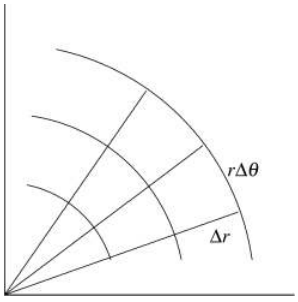
14.6: Moments and Centers of Mass is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **15.3: Moment and Center of Mass** by David Guichard is licensed CC BY-NC-SA 4.0. Original source: https://www.whitman.edu/mathematics/multivariable/.

14.7: Triple Integrals in Cylindrical and Spherical Coordinates

We have seen that sometimes double integrals are simplified by doing them in polar coordinates; not surprisingly, triple integrals are sometimes simpler in cylindrical coordinates or spherical coordinates. To set up integrals in polar coordinates, we had to understand the shape and area of a typical small region into which the region of integration was divided. We need to do the same thing here, for three dimensional regions.

The cylindrical coordinate system is the simplest, since it is just the polar coordinate system plus a z coordinate. A typical small unit of volume is the shape shown below "fattened up" in the z direction, so its volume is $r\Delta r\Delta\theta\Delta z$, or in the limit, $r\,dr\,d\theta\,dz$.



A polar coordinates "grid".

Example 14.7.1

Find the volume under $z=\sqrt{4-r^2}$ above the quarter circle inside $x^2+y^2=4$ in the first quadrant.

Solution

We could of course do this with a double integral, but we'll use a triple integral:

$$\int_0^{\pi/2}\int_0^2\int_0^{\sqrt{4-r^2}} r\,dz\,dr\,d\theta = \int_0^{\pi/2}\int_0^2 \sqrt{4-r^2}\;r\,dr\,d\theta = \frac{4\pi}{3}.$$

Compare this to Example 15.2.1.

Example 14.7.2

An object occupies the space inside both the cylinder $x^2+y^2=1$ and the sphere $x^2+y^2+z^2=4$, and has density x^2 at (x,y,z). Find the total mass.

Solution

We set this up in cylindrical coordinates, recalling that $x=r\cos\theta$:

$$\begin{aligned}\int_0^{2\pi}\int_0^1\int_{-\sqrt{4-r^2}}^{\sqrt{4-r^2}} r^3\cos^2(\theta)\,dz\,dr\,d\theta &= \int_0^{2\pi}\int_0^1 2\sqrt{4-r^2}\;r^3\cos^2(\theta)\,dr\,d\theta\\ &= \int_0^{2\pi}\left(\frac{128}{15}-\frac{22}{5}\sqrt{3}\right)\cos^2(\theta)\,d\theta\\ &= \left(\frac{128}{15}-\frac{22}{5}\sqrt{3}\right)\pi\end{aligned}$$

Spherical coordinates are somewhat more difficult to understand. The small volume we want will be defined by $\Delta\rho$, $\Delta\phi$, and $\Delta\theta$, as pictured in Figure 14.7.1.

The small volume is nearly box shaped, with 4 flat sides and two sides formed from bits of concentric spheres. When $\Delta\rho$, $\Delta\phi$, and $\Delta\theta$ are all very small, the volume of this little region will be nearly the volume we get by treating it as a box. One dimension of the box is simply $\Delta\rho$, the change in distance from the origin. The other two dimensions are the lengths of small circular arcs, so they are $r\Delta\alpha$ for some suitable r and α, just as in the polar coordinates case.

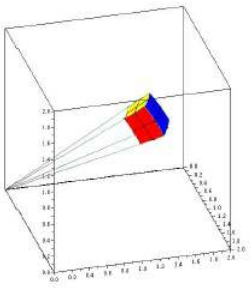


Figure 14.7.1: A small unit of volume for a spherical coordinates (AP)

The easiest of these to understand is the arc corresponding to a change in ϕ, which is nearly identical to the derivation for polar coordinates, as shown in the left graph in Figure 14.7.2 In that graph we are looking "face on" at the side of the box we are interested in, so the small angle pictured is precisely $\Delta\phi$, the vertical axis really is the z axis, but the horizontal axis is *not* a real axis---it is just some line in the x-y plane. Because the other arc is governed by θ, we need to imagine looking straight down the z axis, so that the apparent angle we see is $\Delta\theta$. In this view, the axes really are the x and y axes. In this graph, the apparent distance from the origin is not ρ but $\rho\sin\phi$, as indicated in the left graph.

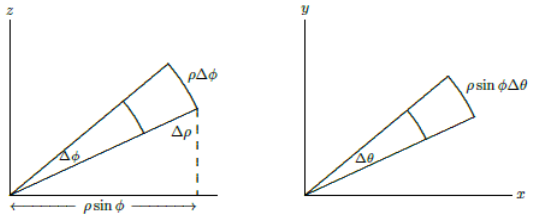


Figure 14.7.2: *Setting up integration in spherical coordinates.*

The upshot is that the volume of the little box is approximately $\Delta\rho(\rho\Delta\phi)(\rho\sin\phi\Delta\theta) = \rho^2\sin\phi\Delta\rho\Delta\phi\Delta\theta$, or in the limit $\rho^2\sin\phi\,d\rho\,d\phi\,d\theta$.

Example 14.7.3

Suppose the temperature at (x, y, z) is

$$T = \frac{1}{1+x^2+y^2+z^2}.$$

Find the average temperature in the unit sphere centered at the origin.

Solution

In two dimensions we add up the temperature at "each" point and divide by the area; here we add up the temperatures and divide by the volume, $(4/3)\pi$:

$$\frac{3}{4\pi}\int_{-1}^{1}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\int_{-\sqrt{1-x^2-y^2}}^{\sqrt{1-x^2-y^2}}\frac{1}{1+x^2+y^2+z^2}\,dz\,dy\,dx$$

This looks quite messy; since everything in the problem is closely related to a sphere, we'll convert to spherical coordinates.

$$\frac{3}{4\pi}\int_0^{2\pi}\int_0^{\pi}\int_0^1\frac{1}{1+\rho^2}\,\rho^2\sin\phi\,d\rho\,d\phi\,d\theta = \frac{3}{4\pi}(4\pi-\pi^2) = 3-\frac{3\pi}{4}.$$

Contributors

- David Guichard (Whitman College)
- Integrated by Justin Marshall.

14.7: Triple Integrals in Cylindrical and Spherical Coordinates is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **15.6: Cylindrical and Spherical Coordinates** by David Guichard is licensed CC BY-NC-SA 4.0. Original source: https://www.whitman.edu/mathematics/multivariable/.

14.8: Substitutions in Multiple Integrals

This section discusses the translation of a graph from the xy Cartesian plane to the uv Cartesian plane and defines the Jacobian.

Introduction

As observed in other sections regarding polar coordinates, some integration of functions on the xyz-space are more easily integrated by translating them to different coordinate systems. These substitutions can make the integrand and/or the limits of integration easier to work with, as "U" Substitution did for single integrals. In this section, we will translate functions from the x-y-z Cartesian coordinate plane to the u-v-w Cartesian coordinate plane to make some integrations easier to solve.

One key component of this translation is called the Jacobian determinant, or simply the Jacobian, which measures how much the volume at a certain point changes when being transformed from one coordinate system to another.

It is important to note that although we are changing the coordinate system on which we graph our function, the theory behind multiple integrals has not changed. The limits of integration still create the domain under the curve, and the integration will help us find the volume of the figure created by the original function and the domain.

Theoretical discussion with Descriptive Elaboration

For any given function $f(x,y)$, we can define x and y as a function of other variables $g(u,v)$. To do this, we first find u and v as a function of x and y that will allow for an easier integrand. Then solve for x and y in order to translate the function so that $x=g(u,v)$ and $y=h(u,v)$. This translates the are region from R in the x-y plane to D in the u-v plane.

Remember:

$$I=\iint_R f(x,y)\,dA$$

So we must find dA:

dA changes from $dxdy$ to $|J(u,v)|\,dudv$. Each change in u (Δu) and change in v (Δv) create parallelograms that are small areas ΔA or dA . We can find the area of each of these parallelograms (P) by taking the cross product of the two vectors that create it (Δu and Δv).

$$\text{Area of P}=\left|\vec{V_1}\times\vec{V_2}\right|=J(u,v)$$

$$\begin{aligned} J(u,v)&=\begin{vmatrix}\frac{\partial x}{\partial u} & \frac{\partial x}{\partial v}\\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v}\end{vmatrix}\\ &=\frac{\partial x}{\partial u}\frac{\partial y}{\partial v}-\frac{\partial y}{\partial u}\frac{\partial x}{\partial v}\\ &=\frac{\partial(x,y)}{\partial(u,v)}\end{aligned}$$

$|J(u,v)|$ represents the area of the parallelogram, and it is the determinant of the Jacobian matrix, shown above. The Jacobian measures how much the transformation is changing from the region R to the region G. Therefore, the translation of the integration of $f(x,y)$ is represented by

$$\iint_R f(x,y)\,dx\,dy=\iint_G f(g(u,v),\ h(u,v))|J(u,v)|du\,dv.$$

The same can be applied for triple integrals, where the translations are represented by

$$x=g(u,v,w),\ y=h(u,v,w),\ z=k(u,v,w)$$

$$
\begin{aligned}
J(u,v,w) &= \frac{\partial(x,y,z)}{\partial(u,v,w)} \\
&= \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{vmatrix} \\
&= \frac{\partial x}{\partial u}\begin{vmatrix} \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{vmatrix} - \frac{\partial x}{\partial v}\begin{vmatrix} \frac{\partial y}{\partial u} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial w} \end{vmatrix} + \frac{\partial x}{\partial w}\begin{vmatrix} \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} \end{vmatrix} \\
&= \frac{\partial x}{\partial u}\left(\frac{\partial y}{\partial v}\cdot\frac{\partial z}{\partial w} - \frac{\partial z}{\partial v}\cdot\frac{\partial y}{\partial w}\right) - \frac{\partial x}{\partial v}\left(\frac{\partial y}{\partial u}\cdot\frac{\partial z}{\partial w} - \frac{\partial z}{\partial u}\cdot\frac{\partial y}{\partial w}\right) + \frac{\partial x}{\partial w}\left(\frac{\partial y}{\partial u}\cdot\frac{\partial z}{\partial v} - \frac{\partial z}{\partial u}\cdot\frac{\partial y}{\partial v}\right)
\end{aligned}
$$

This method for getting the Jacobian is called *cofactor expansion*.

Although the introduction focused primarily on translating a Cartesian function to a different Cartesian coordinate system, the concept of the Jacobian can also be used to explain how translations into polar coordinates work as well.

For cylindrical coordinates

$$x = r\cos\theta,\ y = r\sin\theta,\ z = z$$

Therefore:

$$J(u,v,w) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin & r\cos\theta & 0 \\ 0 & 0 & 1 \end{vmatrix}$$

$$\iiint_D F(x,y,z)\,dx\,dy\,dz = \iiint_G H(r,\theta,z)|r|\,dr\,d\theta\,dz$$

For spherical coordinates

$$J(\rho,\phi,\theta) = \begin{vmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \phi} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \phi} & \frac{\partial y}{\partial \theta} \\ \frac{\partial z}{\partial \rho} & \frac{\partial z}{\partial \phi} & \frac{\partial z}{\partial \theta} \end{vmatrix} = \rho^2\,\sin\phi$$

$$\iiint_D F(x,y,z)\,dx\,dy\,dz = \iiint_G H(\rho,\phi,\theta)|\rho^2\sin\phi\,|\,d\rho\,d\phi\,d\theta$$

Hence, $dx\,dy\,dz$ becomes $rd\,rd\,\theta$ in cylindrical coordinates and $\rho^2\sin\phi\,d\rho\,d\phi\,d\theta$ in spherical coordinates.

Example 14.8.1

Use the following transformation to evaluate the integral.

$$u = \frac{y}{x} \text{ and } v = xy$$

$$\iint_R \frac{y}{x}dA$$

Where R is bounded by: $1 \le u \le 2$ and $1 \le v \le 2$

Solution

First find x and y as functions of u and v:

$$u = \frac{y}{x} \; v = xy$$
$$y = xu \rightarrow x = \frac{v}{y}$$
$$x = \frac{v}{xu}$$
$$x^2 = \frac{v}{u}$$
$$x = \sqrt{\frac{v}{u}}$$
$$y = xu$$
$$y = \left(\sqrt{\frac{v}{u}}\right) u$$
$$y = \sqrt{vu}$$

$$x = g(u,v) = \sqrt{\frac{v}{u}} \text{ and } y = h(u,v) = \sqrt{vu}$$

Using $x = g(u,v)$ and $y = h(u,v)$, find the integrand in terms of u and v:

$$\frac{y}{x} = \frac{\sqrt{vu}}{\sqrt{\frac{v}{u}}} = u$$

And dA changes from $dxdy$ to $|J(u,v)|\ du\ dv$. The Jacobian equals:

$$J(u,v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v}$$

$$\frac{\partial x}{\partial u} = \frac{1}{2}u^{-\frac{3}{2}}v^{\frac{1}{2}} \quad \frac{\partial x}{\partial v} = \frac{1}{2}u^{-\frac{1}{2}}v^{-\frac{1}{2}} \quad \frac{\partial y}{\partial u} = \frac{1}{2}u^{-\frac{1}{2}}v^{\frac{1}{2}} \quad \frac{\partial y}{\partial v} = \frac{1}{2}u^{\frac{1}{2}}v^{-\frac{1}{2}}$$

$$\begin{aligned} J(u,v) &= \begin{vmatrix} \frac{1}{2}u^{-\frac{3}{2}}v^{\frac{1}{2}} & \frac{1}{2}u^{-\frac{1}{2}}v^{-\frac{1}{2}} \\ \frac{1}{2}u^{-\frac{1}{2}}v^{\frac{1}{2}} & \frac{1}{2}u^{\frac{1}{2}}v^{-\frac{1}{2}} \end{vmatrix} \\ &= \left(-\frac{1}{4}u^{-1} - \frac{1}{4}u^{-1}\right) \\ &= \frac{1}{2u} \end{aligned}$$

Therefore, evaluate:

$$\int_1^2 \int_1^2 u \left(\frac{1}{2u}\right) du\, dv$$
$$= \int_1^2 \int_1^2 \frac{1}{2}\, du\, dv$$
$$= \int_1^2 \frac{1}{2}u\Big|_1^2 dv$$
$$= \int_1^2 1 - \frac{1}{2}\, dv$$
$$= \frac{1}{2}v\Big|_1^2$$
$$= 1 - \frac{1}{2} = \frac{1}{2}$$

Example 14.8.2

Use the following transformation to evaluate the integral.

$$u = x - \frac{1}{2}y \text{ and } v = y$$
$$\int_0^{\frac{1}{2}} \int_{\frac{y}{2}}^{\frac{y+4}{2}} y^3(2x-y)e^{2x-y^2}\, dx\, dy$$

Solution

First solve for x and y:

$$u = x - \frac{1}{2}y \;\; v = y$$
$$u = x - \frac{1}{2}v \;\; y = v$$
$$x = u + \frac{1}{2}v\,.$$

Then substitute x and y for $g(u,v)$ and $h(u,v)$:

The integrand:

$$y^3(2x-y)e^{2x-y^2} \;\rightarrow\; v^3[2(u+\frac{1}{2}v)-v]e^{[2(u+\frac{1}{2}v)-v]^2}$$
$$= v^3(2u)e^{2u^2}$$
$$= (2uv^3)e^{4u^2}$$

The transformation also changes the bounds of integration

\(
$$x = \frac{y+4}{2} \;\rightarrow\; u + \frac{1}{2}v = \frac{v+4}{2}$$
$$= \frac{y}{4} \;\rightarrow\; u + \frac{v}{2} = \frac{v}{2}$$
\]
$$u = \frac{4}{2}$$
$$u = 0$$

$$u = 2$$

And $dxdy$ changes to $|J(u,v)|\ du\ dv$. The Jacobian equals:

$$J(u,v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial u} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v}$$
$$= \begin{vmatrix} 1 & 5 \\ 0 & 1 \end{vmatrix} = 1$$

Thus,

$$\begin{aligned} \iint_R f(x,y)\,dx\,dy &= \iint_G f(g(u,v),\ h(u,v))|J(u,v)|du\,dv \\ &= \int_0^{\frac{1}{2}} \int_0^2 2uv^3 e^{4u^2}\ (1)\,dv\,du \\ &= \int_0^{\frac{1}{2}} \left.\frac{ue^{4u^2}v^4}{2}\right|_0^2 du \\ &= \int_0^{\frac{1}{2}} 8ue^{4u^2}\,du \\ &= e^{4u^2}\Big|_0^{\frac{1}{2}} \\ &= e-1 \end{aligned}$$

Example 14.8.3

Find the mass of an object bounded by

$$1 \le x \le 2,\ \ 0 \le xy \le 1,\ \ 0 \le z \le 2$$

with a density that can be described by the formula $x^2y + 2xyz$ by using the transformation $u = x,\ \ v = xy,\ \ w = 3z$.

Solution

Set up the integral in cartesian coordinates:

$$\int_1^2 \int_0^{\frac{1}{x}} \int_0^2 x^2y + 2xyz\,dzdydx.$$

To apply the substitution, first solve for x and y using the given transformations:

$$\begin{array}{lll} u = x & v = xy & w = 3z \\ x = u & y = \frac{v}{x} & z = \frac{w}{3} \\ & y = \frac{v}{u}. & \end{array}$$

Then make the appropriate substitutions within the integrand:

$$x^2y + 2xyz \to u^2\left(\frac{v}{u}\right) + 2u\left(\frac{v}{u}\right)\left(\frac{w}{3}\right)\ \to\ uv + \frac{2vw}{3}.$$

Next, find the new boundaries to the region we want to integrate:

$$1 \le x \le 2\ \to\ 1 \le u \le 2$$

$$0 \le xy \le 1 \;\rightarrow\; 0 \le v \le 1$$

$$0 \le z \le 2 \;\rightarrow\; 0 \le \frac{w}{3} \le 2 \;\rightarrow\; 0 \le w \le 6 \quad .$$

To complete the transformation, find the Jacobian:

$$\begin{aligned} J(u,v,w) = \frac{\partial(x,y,z)}{\partial(u,v,w)} &= \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{vmatrix} \\ &= \begin{vmatrix} 1 & 0 & 0 \\ -\frac{v}{u^2} & \frac{1}{u} & 0 \\ 0 & 0 & \frac{1}{3} \end{vmatrix} \\ &= \frac{1}{3u}. \end{aligned}$$

Notice the Jacobian of a lower triangular matrix (the values above the diagonal are all zero) is the multiplication of the diagonal entries. You can confirm this with cofactor expansion.

Using all of our calculated transformations, we can compute the new integral:

$$\begin{aligned} \text{Mass} &= \int_0^6 \int_0^1 \int_1^2 \left(uv + \frac{2vw}{3}\right) \frac{1}{3u}\, du dv dw \\ &= \frac{1}{3}\int_0^6 \int_0^1 \int_1^2 v + \frac{2vw}{3u}\, du dv dw \\ &= \frac{1}{2}\int_0^6 \int_0^1 vu + \frac{2vw}{3}\ln|u|\Big|_1^2\, dv dw \\ &= \frac{1}{3}\int_0^6 \int_0^1 v + \frac{2vw}{3}\ln 2\, dv dw \\ &= \frac{1}{3}\int_0^6 \frac{v^2}{2} + \frac{2w\ln 2}{3}\left(\frac{v^2}{2}\right)\Big|_0^1\, dw \\ &= \frac{1}{3}\int_0^6 \frac{1}{2} + \frac{w\ln 2}{3}\, dw \\ &= \frac{1}{3}\left[\frac{1}{2}w + \frac{w^2\ln 2}{6}\right]_0^6 \\ &= \frac{1}{3}[3 + 6\ln 2] \\ &= 1 + 2\ln 2. \end{aligned}$$

Contributors and Attributions

- Sydney Wong (UCD), Alagu Chidambaram (UCD)
- Integrated by Justin Marshall.

14.8: Substitutions in Multiple Integrals is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **3.9: Substitutions in Multiple Integrals** is licensed CC BY 4.0.

CHAPTER OVERVIEW

15: Integration in Vector Fields

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

15.1: Line Integrals
15.2: Vector Fields and Line Integrals- Work, Circulation, and Flux
15.3: Path Independence, Conservative Fields, and Potential Functions
15.4: Green's Theorem in the Plane
15.5: Surfaces and Area
15.6: Surface Integrals
15.7: Stokes' Theorem
15.8: The Divergence Theorem and a Unified Theory

15: Integration in Vector Fields is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

15.1: Line Integrals

15.1: Line Integrals is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

15.2: Vector Fields and Line Integrals- Work, Circulation, and Flux

15.2: Vector Fields and Line Integrals- Work, Circulation, and Flux is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

15.3: Path Independence, Conservative Fields, and Potential Functions

15.3: Path Independence, Conservative Fields, and Potential Functions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

15.4: Green's Theorem in the Plane

15.4: Green's Theorem in the Plane is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

15.5: Surfaces and Area

15.5: Surfaces and Area is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

15.6: Surface Integrals

15.6: Surface Integrals is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

15.7: Stokes' Theorem

15.7: Stokes' Theorem is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

15.8: The Divergence Theorem and a Unified Theory

15.8: The Divergence Theorem and a Unified Theory is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

16: First-Order Differential Equations

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

16.1: Solutions, Slope Fields, and Euler's Method
16.2: First-Order Linear Equations
16.3: Applications
16.4: Graphical Solutions of Autonomous Equations
16.5: Systems of Equations and Phase Planes

16: First-Order Differential Equations is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.1: Solutions, Slope Fields, and Euler's Method

16.1: Solutions, Slope Fields, and Euler's Method is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.2: First-Order Linear Equations

16.2: First-Order Linear Equations is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.3: Applications

16.3: Applications is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.4: Graphical Solutions of Autonomous Equations

16.4: Graphical Solutions of Autonomous Equations is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.5: Systems of Equations and Phase Planes

16.5: Systems of Equations and Phase Planes is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

17: Second-Order Differential Equations

A general Calculus Textmap organized around the textbook

University Calculus: Early Transcendentals

by Joel R. Hass, Maurice D. Weir, George B. Thomas

This Textmap is currently under construction... please be patient with us.

Template:HideTOC

17.1: Second-Order Linear Equations
17.2: Nonhomogeneous Linear Equations
17.3: Applications
17.4: Euler Equations
17.5: Power Series Solutions

17: Second-Order Differential Equations is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

17.1: Second-Order Linear Equations

17.1: Second-Order Linear Equations is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

17.2: Nonhomogeneous Linear Equations

17.2: Nonhomogeneous Linear Equations is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

17.3: Applications

17.3: Applications is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

17.4: Euler Equations

17.4: Euler Equations is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

17.5: Power Series Solutions

17.5: Power Series Solutions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

Index

A

Absolute Value Function
1.1: Review of Functions
algebraic function
1.2: Basic Classes of Functions
antiderivative
5.1: Antiderivatives
arc length
12.3: Arc Length in Space
average value of a function
5.3: The Definite Integral
average velocity
3.1: Defining the Derivative

C

chain rule
3.6: The Chain Rule
change of variables
5.5: Substitution with Indefinite Integrals
Closeness
2.2: The Precise Definition of a Limit
composite function
1.1: Review of Functions
concave down
4.3: The First Derivative Test
4.4: The Second Derivative Test and Curve Sketching
concave up
4.3: The First Derivative Test
4.4: The Second Derivative Test and Curve Sketching
concavity
4.3: The First Derivative Test
4.4: The Second Derivative Test and Curve Sketching
concavity test
4.3: The First Derivative Test
4.4: The Second Derivative Test and Curve Sketching
cross product
11.4: The Cross Product
cubic function
1.2: Basic Classes of Functions

D

decreasing on the interval I
1.1: Review of Functions
definite integral
5.3: The Definite Integral
degree
1.2: Basic Classes of Functions
dependent variable
1.1: Review of Functions
derivative
3.1: Defining the Derivative
derivative function
3.2: The Derivative as a Function
Derivative of exponential functions
3.8: Derivatives of Inverse and Logarithmic Functions
Derivative of logarithmic functions
3.8: Derivatives of Inverse and Logarithmic Functions
Derivative of the general exponential function
3.8: Derivatives of Inverse and Logarithmic Functions
Derivative of the general logarithmic function
3.8: Derivatives of Inverse and Logarithmic Functions
Derivative of the natural exponential function
3.8: Derivatives of Inverse and Logarithmic Functions
Derivative of the natural logarithmic function
3.8: Derivatives of Inverse and Logarithmic Functions
difference quotient
3.1: Defining the Derivative
differentiable function
3.2: The Derivative as a Function
differentiable on S
3.2: The Derivative as a Function
Directional Derivatives
13.5: Directional Derivatives and Gradient Vectors
domain
1.1: Review of Functions
dot product
11.3: The Dot Product

E

even function
1.1: Review of Functions
exponent
1.5: Exponential and Logarithmic Functions
extrema
13.7: Extreme Values and Saddle Points

F

First Derivative Test
4.3: The First Derivative Test
4.4: The Second Derivative Test and Curve Sketching
function
1.1: Review of Functions
fundamental theorem of calculus
5.4: The Fundamental Theorem of Calculus
fundamental theorem of calculus, part 1
5.4: The Fundamental Theorem of Calculus
fundamental theorem of calculus, part 2
5.4: The Fundamental Theorem of Calculus

G

Gradient Vectors
13.5: Directional Derivatives and Gradient Vectors
graph of a function
1.1: Review of Functions

H

horizontal line test
1.4: Inverse Functions
hyperbolic functions
1.5: Exponential and Logarithmic Functions
7.3: Hyperbolic Functions

I

implicit differentiation
3.7: Implicit Differentiation
increasing on the interval I
1.1: Review of Functions
indefinite integral
5.1: Antiderivatives
independent variable
1.1: Review of Functions
inflection point
4.3: The First Derivative Test
4.4: The Second Derivative Test and Curve Sketching
initial value problem
5.1: Antiderivatives
instantaneous rate of change
3.1: Defining the Derivative
instantaneous velocity
3.1: Defining the Derivative
integrable function
5.3: The Definite Integral
Integral Test
9.4: The Integral Test
integrand
5.3: The Definite Integral
integration by substitution
5.5: Substitution with Indefinite Integrals
inverse function
1.4: Inverse Functions
inverse hyperbolic functions
1.5: Exponential and Logarithmic Functions
inverse trigonometric functions
1.4: Inverse Functions

L

line
11.5: Lines and Planes in Space
linear function
1.2: Basic Classes of Functions
Logarithmic Function
1.2: Basic Classes of Functions
lower sum
5.2: Approximating Areas with Limits of Finite Sums

M

mathematical model
1.2: Basic Classes of Functions
Mean Value Theorem
4.2: The Mean Value Theorem
Mean Value Theorem for Integrals
5.4: The Fundamental Theorem of Calculus
method of exhaustion
5.2: Approximating Areas with Limits of Finite Sums
Moments of Inertia
6.6: Moments and Centers of Mass
14.6: Moments and Centers of Mass

N

natural exponential function
1.5: Exponential and Logarithmic Functions
natural logarithm
1.5: Exponential and Logarithmic Functions

net signed area
5.3: The Definite Integral
Newton’s method
4.7: Newton's Method
number e
1.5: Exponential and Logarithmic Functions

O

odd function
1.1: Review of Functions

P

partition
5.2: Approximating Areas with Limits of Finite Sums
periodic function
1.3: Trigonometric Functions
plane
11.5: Lines and Planes in Space
points of inflection
4.3: The First Derivative Test
4.4: The Second Derivative Test and Curve Sketching
polynomial function
1.2: Basic Classes of Functions
power function
1.2: Basic Classes of Functions

Q

quadratic function
1.2: Basic Classes of Functions

R

radians
1.3: Trigonometric Functions
range
1.1: Review of Functions
rational function
1.2: Basic Classes of Functions
regular partition
5.2: Approximating Areas with Limits of Finite Sums
related rates
3.10: Related Rates
restricted domain
1.4: Inverse Functions
Riemann sum
5.2: Approximating Areas with Limits of Finite Sums
Rolle’s Theorem
4.2: The Mean Value Theorem
root function
1.2: Basic Classes of Functions

S

saddle point
13.7: Extreme Values and Saddle Points
second derivative test
4.3: The First Derivative Test
4.4: The Second Derivative Test and Curve Sketching
sigma notation
5.2: Approximating Areas with Limits of Finite Sums
slope
1.2: Basic Classes of Functions
summation notation
5.2: Approximating Areas with Limits of Finite Sums
symmetry about the origin
1.1: Review of Functions

T

table of values
1.1: Review of Functions
tangent line
3.1: Defining the Derivative
The power rule for functions
3.6: The Chain Rule
total area
5.3: The Definite Integral
transcendental function
1.2: Basic Classes of Functions
transformation of a function
1.2: Basic Classes of Functions
trigonometric functions
1.3: Trigonometric Functions
trigonometric identity
1.3: Trigonometric Functions

U

unit circle
1.3: Trigonometric Functions
upper sum
5.2: Approximating Areas with Limits of Finite Sums

V

variable of integration
5.3: The Definite Integral
vertical line test
1.1: Review of Functions
Volume by Shells
6.2: Volumes Using Cylindrical Shells

Z

zeros of a function
1.1: Review of Functions

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Calculus (Hutchinson)

Webpages: 140

Applicable Restrictions: Noncommercial

All licenses found:

- Undeclared: 86.4% (121 pages)
- CC BY-NC-SA 4.0: 13.6% (19 pages)

By Page

- Calculus (Hutchinson) - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Functions and Graphs (Precalculus Review) - *Undeclared*
    - 1.1: Review of Functions - *CC BY-NC-SA 4.0*
    - 1.2: Basic Classes of Functions - *CC BY-NC-SA 4.0*
    - 1.3: Trigonometric Functions - *CC BY-NC-SA 4.0*
    - 1.4: Inverse Functions - *CC BY-NC-SA 4.0*
    - 1.5: Exponential and Logarithmic Functions - *CC BY-NC-SA 4.0*
  - 2: Limits and Continuity - *Undeclared*
    - 2.1: Limit of a Function and Limit Laws - *Undeclared*
    - 2.2: The Precise Definition of a Limit - *CC BY-NC-SA 4.0*
    - 2.3: One-Sided Limits - *Undeclared*
    - 2.4: Continuity - *Undeclared*
    - 2.5: Limits Involving Infinity; Asymptotes of Graphs - *Undeclared*
  - 3: Differentiation - *Undeclared*
    - 3.1: Defining the Derivative - *CC BY-NC-SA 4.0*
    - 3.2: The Derivative as a Function - *CC BY-NC-SA 4.0*
    - 3.3: Differentiation Rules - *Undeclared*
    - 3.4: The Derivative as a Rate of Change - *Undeclared*
    - 3.5: Derivatives of Trigonometric Functions - *Undeclared*
    - 3.6: The Chain Rule - *CC BY-NC-SA 4.0*
    - 3.7: Implicit Differentiation - *CC BY-NC-SA 4.0*
    - 3.8: Derivatives of Inverse and Logarithmic Functions - *CC BY-NC-SA 4.0*
    - 3.9: Inverse Trigonometric Functions - *Undeclared*
    - 3.10: Related Rates - *CC BY-NC-SA 4.0*
    - 3.11: Linearization and Differentials - *Undeclared*
  - 4: Applications of Derivatives - *Undeclared*
    - 4.1: Maxima and Minima - *Undeclared*
    - 4.2: The Mean Value Theorem - *Undeclared*
    - 4.3: The First Derivative Test - *CC BY-NC-SA 4.0*
    - 4.4: The Second Derivative Test and Curve Sketching - *CC BY-NC-SA 4.0*
    - 4.5: Indeterminate Forms and L'Hôpital's Rule - *Undeclared*
    - 4.6: Applied Optimization Problems - *Undeclared*
    - 4.7: Newton's Method - *Undeclared*
  - 5: Integration - *Undeclared*
    - 5.1: Antiderivatives - *CC BY-NC-SA 4.0*
    - 5.2: Approximating Areas with Limits of Finite Sums - *CC BY-NC-SA 4.0*
    - 5.3: The Definite Integral - *CC BY-NC-SA 4.0*
    - 5.4: The Fundamental Theorem of Calculus - *CC BY-NC-SA 4.0*
    - 5.5: Substitution with Indefinite Integrals - *CC BY-NC-SA 4.0*
    - 5.6: Substitution with Definite Integrals - *Undeclared*
  - 6: Applications of Definite Integrals - *Undeclared*
    - 6.1: Volumes Using Cross-Sections - *Undeclared*
    - 6.2: Volumes Using Cylindrical Shells - *Undeclared*
    - 6.3: Arc Length - *Undeclared*
    - 6.4: Areas of Surfaces of Revolution - *Undeclared*
    - 6.5: Work - *Undeclared*
    - 6.6: Moments and Centers of Mass - *Undeclared*
  - 7: Integrals and Transcendental Functions - *Undeclared*
    - 7.1: The Logarithm Defined as an Integral - *Undeclared*
    - 7.2: Exponential Change and Separable Differential Equations - *Undeclared*
    - 7.3: Hyperbolic Functions - *Undeclared*
  - 8: Techniques of Integration - *Undeclared*
    - 8.1: Integration by Parts - *Undeclared*
    - 8.2: Trigonometric Integrals - *Undeclared*
    - 8.3: Trigonometric Substitutions - *Undeclared*

- 8.4: Integration of Rational Functions by Partial Fractions - *Undeclared*
- 8.5: Integral Tables and Computer Algebra Systems - *Undeclared*
- 8.6: Numerical Integration - *Undeclared*
- 8.7: Improper Integrals - *Undeclared*

- 9: Infinite Sequence and Series - *Undeclared*
  - 9.1: The Binomial Series and Applications of Taylor Series - *Undeclared*
  - 9.2: Sequences - *Undeclared*
  - 9.3: Infinite Series - *Undeclared*
  - 9.4: The Integral Test - *Undeclared*
  - 9.5: Comparison Tests - *Undeclared*
  - 9.6: The Ratio and Root Tests - *Undeclared*
  - 9.7: Alternating Series, Absolute and Conditional Convergence - *Undeclared*
  - 9.8: Power Series - *Undeclared*
  - 9.9: Taylor and Maclaurin Series - *Undeclared*
  - 9.10: Convergence of Taylor Series - *Undeclared*
- 10: Parametric Equations and Polar Coordinates - *Undeclared*
  - 10.1: Parametrizations of Plane Curves - *Undeclared*
  - 10.2: Calculus with Parametric Curves - *Undeclared*
  - 10.3: Polar Coordinates - *Undeclared*
  - 10.4: Graphing in Polar Coordinates - *Undeclared*
  - 10.5: Areas and Lengths in Polar Coordinates - *Undeclared*
  - 10.6: Conics in Polar Coordinates - *Undeclared*
- 11: Vectors and the Geometry of Space - *Undeclared*
  - 11.1: Three-Dimensional Coordinate Systems - *Undeclared*
  - 11.2: Vectors - *Undeclared*
  - 11.3: The Dot Product - *Undeclared*
  - 11.4: The Cross Product - *Undeclared*
  - 11.5: Lines and Planes in Space - *Undeclared*
  - 11.6: Cylinders and Quadric Surfaces - *Undeclared*
- 12: Vector-Valued Functions and Motion in Space - *Undeclared*
  - 12.1: Curves in Space and Their Tangents - *Undeclared*
  - 12.2: Integrals of Vector Functions; Projectile Motion - *Undeclared*
  - 12.3: Arc Length in Space - *Undeclared*
  - 12.4: Curvature and Normal Vectors of a Curve - *Undeclared*
  - 12.5: Tangential and Normal Components of Acceleration - *Undeclared*
  - 12.6: Velocity and Acceleration in Polar Coordinates - *Undeclared*
- 13: Partial Derivatives - *Undeclared*
  - 13.1: Functions of Several Variables - *Undeclared*
  - 13.2: Limits and Continuity in Higher Dimensions - *Undeclared*
  - 13.3: Partial Derivatives - *Undeclared*
  - 13.4: The Chain Rule - *Undeclared*
  - 13.5: Directional Derivatives and Gradient Vectors - *Undeclared*
  - 13.6: Tangent Planes and Differentials - *Undeclared*
  - 13.7: Extreme Values and Saddle Points - *Undeclared*
  - 13.8: Lagrange Multipliers - *Undeclared*
- 14: Multiple Integrals - *Undeclared*
  - 14.1: Double and Iterated Integrals over Rectangles - *Undeclared*
  - 14.2: Double Integrals over General Regions - *Undeclared*
  - 14.3: Area by Double Integration - *Undeclared*
  - 14.4: Double Integrals in Polar Form - *Undeclared*
  - 14.5: Triple Integrals in Rectangular Coordinates - *Undeclared*
  - 14.6: Moments and Centers of Mass - *Undeclared*
  - 14.7: Triple Integrals in Cylindrical and Spherical Coordinates - *Undeclared*
  - 14.8: Substitutions in Multiple Integrals - *Undeclared*
- 15: Integration in Vector Fields - *Undeclared*
  - 15.1: Line Integrals - *Undeclared*
  - 15.2: Vector Fields and Line Integrals- Work, Circulation, and Flux - *Undeclared*
  - 15.3: Path Independence, Conservative Fields, and Potential Functions - *Undeclared*
  - 15.4: Green's Theorem in the Plane - *Undeclared*
  - 15.5: Surfaces and Area - *Undeclared*
  - 15.6: Surface Integrals - *Undeclared*
  - 15.7: Stokes' Theorem - *Undeclared*
  - 15.8: The Divergence Theorem and a Unified Theory - *Undeclared*
- 16: First-Order Differential Equations - *Undeclared*
  - 16.1: Solutions, Slope Fields, and Euler's Method - *Undeclared*
  - 16.2: First-Order Linear Equations - *Undeclared*
  - 16.3: Applications - *Undeclared*
  - 16.4: Graphical Solutions of Autonomous Equations - *Undeclared*
  - 16.5: Systems of Equations and Phase Planes - *Undeclared*
- 17: Second-Order Differential Equations - *Undeclared*
  - 17.1: Second-Order Linear Equations - *Undeclared*
  - 17.2: Nonhomogeneous Linear Equations - *Undeclared*
  - 17.3: Applications - *Undeclared*
  - 17.4: Euler Equations - *Undeclared*
  - 17.5: Power Series Solutions - *Undeclared*

- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: Calculus (Hutchinson)

Webpages: 140

Applicable Restrictions: Noncommercial

All licenses found:

- Undeclared: 86.4% (121 pages)
- CC BY-NC-SA 4.0: 13.6% (19 pages)

By Page

- Calculus (Hutchinson) - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Functions and Graphs (Precalculus Review) - *Undeclared*
    - 1.1: Review of Functions - *CC BY-NC-SA 4.0*
    - 1.2: Basic Classes of Functions - *CC BY-NC-SA 4.0*
    - 1.3: Trigonometric Functions - *CC BY-NC-SA 4.0*
    - 1.4: Inverse Functions - *CC BY-NC-SA 4.0*
    - 1.5: Exponential and Logarithmic Functions - *CC BY-NC-SA 4.0*
  - 2: Limits and Continuity - *Undeclared*
    - 2.1: Limit of a Function and Limit Laws - *Undeclared*
    - 2.2: The Precise Definition of a Limit - *CC BY-NC-SA 4.0*
    - 2.3: One-Sided Limits - *Undeclared*
    - 2.4: Continuity - *Undeclared*
    - 2.5: Limits Involving Infinity; Asymptotes of Graphs - *Undeclared*
  - 3: Differentiation - *Undeclared*
    - 3.1: Defining the Derivative - *CC BY-NC-SA 4.0*
    - 3.2: The Derivative as a Function - *CC BY-NC-SA 4.0*
    - 3.3: Differentiation Rules - *Undeclared*
    - 3.4: The Derivative as a Rate of Change - *Undeclared*
    - 3.5: Derivatives of Trigonometric Functions - *Undeclared*
    - 3.6: The Chain Rule - *CC BY-NC-SA 4.0*
    - 3.7: Implicit Differentiation - *CC BY-NC-SA 4.0*
    - 3.8: Derivatives of Inverse and Logarithmic Functions - *CC BY-NC-SA 4.0*
    - 3.9: Inverse Trigonometric Functions - *Undeclared*
    - 3.10: Related Rates - *CC BY-NC-SA 4.0*
    - 3.11: Linearization and Differentials - *Undeclared*
  - 4: Applications of Derivatives - *Undeclared*
    - 4.1: Maxima and Minima - *Undeclared*
    - 4.2: The Mean Value Theorem - *Undeclared*
    - 4.3: The First Derivative Test - *CC BY-NC-SA 4.0*
    - 4.4: The Second Derivative Test and Curve Sketching - *CC BY-NC-SA 4.0*
    - 4.5: Indeterminate Forms and L'Hôpital's Rule - *Undeclared*
    - 4.6: Applied Optimization Problems - *Undeclared*
    - 4.7: Newton's Method - *Undeclared*
  - 5: Integration - *Undeclared*
    - 5.1: Antiderivatives - *CC BY-NC-SA 4.0*
    - 5.2: Approximating Areas with Limits of Finite Sums - *CC BY-NC-SA 4.0*
    - 5.3: The Definite Integral - *CC BY-NC-SA 4.0*
    - 5.4: The Fundamental Theorem of Calculus - *CC BY-NC-SA 4.0*
    - 5.5: Substitution with Indefinite Integrals - *CC BY-NC-SA 4.0*
    - 5.6: Substitution with Definite Integrals - *Undeclared*
  - 6: Applications of Definite Integrals - *Undeclared*
    - 6.1: Volumes Using Cross-Sections - *Undeclared*
    - 6.2: Volumes Using Cylindrical Shells - *Undeclared*
    - 6.3: Arc Length - *Undeclared*
    - 6.4: Areas of Surfaces of Revolution - *Undeclared*
    - 6.5: Work - *Undeclared*
    - 6.6: Moments and Centers of Mass - *Undeclared*
  - 7: Integrals and Transcendental Functions - *Undeclared*
    - 7.1: The Logarithm Defined as an Integral - *Undeclared*
    - 7.2: Exponential Change and Separable Differential Equations - *Undeclared*
    - 7.3: Hyperbolic Functions - *Undeclared*
  - 8: Techniques of Integration - *Undeclared*
    - 8.1: Integration by Parts - *Undeclared*
    - 8.2: Trigonometric Integrals - *Undeclared*
    - 8.3: Trigonometric Substitutions - *Undeclared*

- 8.4: Integration of Rational Functions by Partial Fractions - *Undeclared*
- 8.5: Integral Tables and Computer Algebra Systems - *Undeclared*
- 8.6: Numerical Integration - *Undeclared*
- 8.7: Improper Integrals - *Undeclared*

- 9: Infinite Sequence and Series - *Undeclared*
  - 9.1: The Binomial Series and Applications of Taylor Series - *Undeclared*
  - 9.2: Sequences - *Undeclared*
  - 9.3: Infinite Series - *Undeclared*
  - 9.4: The Integral Test - *Undeclared*
  - 9.5: Comparison Tests - *Undeclared*
  - 9.6: The Ratio and Root Tests - *Undeclared*
  - 9.7: Alternating Series, Absolute and Conditional Convergence - *Undeclared*
  - 9.8: Power Series - *Undeclared*
  - 9.9: Taylor and Maclaurin Series - *Undeclared*
  - 9.10: Convergence of Taylor Series - *Undeclared*
- 10: Parametric Equations and Polar Coordinates - *Undeclared*
  - 10.1: Parametrizations of Plane Curves - *Undeclared*
  - 10.2: Calculus with Parametric Curves - *Undeclared*
  - 10.3: Polar Coordinates - *Undeclared*
  - 10.4: Graphing in Polar Coordinates - *Undeclared*
  - 10.5: Areas and Lengths in Polar Coordinates - *Undeclared*
  - 10.6: Conics in Polar Coordinates - *Undeclared*
- 11: Vectors and the Geometry of Space - *Undeclared*
  - 11.1: Three-Dimensional Coordinate Systems - *Undeclared*
  - 11.2: Vectors - *Undeclared*
  - 11.3: The Dot Product - *Undeclared*
  - 11.4: The Cross Product - *Undeclared*
  - 11.5: Lines and Planes in Space - *Undeclared*
  - 11.6: Cylinders and Quadric Surfaces - *Undeclared*
- 12: Vector-Valued Functions and Motion in Space - *Undeclared*
  - 12.1: Curves in Space and Their Tangents - *Undeclared*
  - 12.2: Integrals of Vector Functions; Projectile Motion - *Undeclared*
  - 12.3: Arc Length in Space - *Undeclared*
  - 12.4: Curvature and Normal Vectors of a Curve - *Undeclared*
  - 12.5: Tangential and Normal Components of Acceleration - *Undeclared*
  - 12.6: Velocity and Acceleration in Polar Coordinates - *Undeclared*
- 13: Partial Derivatives - *Undeclared*
  - 13.1: Functions of Several Variables - *Undeclared*
  - 13.2: Limits and Continuity in Higher Dimensions - *Undeclared*
  - 13.3: Partial Derivatives - *Undeclared*
  - 13.4: The Chain Rule - *Undeclared*
  - 13.5: Directional Derivatives and Gradient Vectors - *Undeclared*
  - 13.6: Tangent Planes and Differentials - *Undeclared*
  - 13.7: Extreme Values and Saddle Points - *Undeclared*
  - 13.8: Lagrange Multipliers - *Undeclared*
- 14: Multiple Integrals - *Undeclared*
  - 14.1: Double and Iterated Integrals over Rectangles - *Undeclared*
  - 14.2: Double Integrals over General Regions - *Undeclared*
  - 14.3: Area by Double Integration - *Undeclared*
  - 14.4: Double Integrals in Polar Form - *Undeclared*
  - 14.5: Triple Integrals in Rectangular Coordinates - *Undeclared*
  - 14.6: Moments and Centers of Mass - *Undeclared*
  - 14.7: Triple Integrals in Cylindrical and Spherical Coordinates - *Undeclared*
  - 14.8: Substitutions in Multiple Integrals - *Undeclared*
- 15: Integration in Vector Fields - *Undeclared*
  - 15.1: Line Integrals - *Undeclared*
  - 15.2: Vector Fields and Line Integrals- Work, Circulation, and Flux - *Undeclared*
  - 15.3: Path Independence, Conservative Fields, and Potential Functions - *Undeclared*
  - 15.4: Green's Theorem in the Plane - *Undeclared*
  - 15.5: Surfaces and Area - *Undeclared*
  - 15.6: Surface Integrals - *Undeclared*
  - 15.7: Stokes' Theorem - *Undeclared*
  - 15.8: The Divergence Theorem and a Unified Theory - *Undeclared*
- 16: First-Order Differential Equations - *Undeclared*
  - 16.1: Solutions, Slope Fields, and Euler's Method - *Undeclared*
  - 16.2: First-Order Linear Equations - *Undeclared*
  - 16.3: Applications - *Undeclared*
  - 16.4: Graphical Solutions of Autonomous Equations - *Undeclared*
  - 16.5: Systems of Equations and Phase Planes - *Undeclared*
- 17: Second-Order Differential Equations - *Undeclared*
  - 17.1: Second-Order Linear Equations - *Undeclared*
  - 17.2: Nonhomogeneous Linear Equations - *Undeclared*
  - 17.3: Applications - *Undeclared*
  - 17.4: Euler Equations - *Undeclared*
  - 17.5: Power Series Solutions - *Undeclared*

- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*