ACTIVE CALCULUS



Matthew Boelkins, David Austin & Steven Schlicker
Grand Valley State University



5: Active Calculus (Boelkins et al.)

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 an 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 borne 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 were 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.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 10/04/2026

TABLE OF CONTENTS

Licensing

1: Understanding the Derivative

1.1: How do we Measure Velocity?

1.2: The Notion of Limit

1.3: The Derivative of a Function at a Point

1.4: The Derivative Function

1.5: Interpretating, Estimating, and Using the Derivative

1.6: The Second Derivative

1.7: Limits, Continuity, and Differentiability

1.8: The Tangent Line Approximation

1.9: Understanding the Derivative (Exercises)

2: Computing Derivatives

2.1: Elementary Derivative Rules

2.2: The Sine and Cosine Function

2.3: The Product and Quotient Rules

2.4: Derivatives of Other Trigonometric Functions

2.5: The Chain Rule

2.6: Derivatives of Inverse Functions

2.7: Derivatives of Functions Given Implicitly

2.8: Using Derivatives to Evaluate Limits

2.9: Computing Derivatives (Exercises)

3: Using Derivatives

3.1: Using Derivatives to Identify Extreme Values

3.2: Using Derivatives to Describe Families of Functions

3.3: Global Optimization

3.4: Applied Optimization

3.5: Related Rates

3.6: Using Derivatives (Exercises)

4: The Definite Integral

4.1: Determining Distance Traveled from Velocity

4.2: Riemann Sums

4.3: The Definite Integral

4.4: The Fundamental Theorem of Calculus

4.5: The Definite Integral (Exercises)

5: Finding Antiderivatives and Evaluating Integrals

5.1: Construction Accurate Graphs of Antiderivatives

5.2: The Second Fundamental Theorem of Calculus
5.3: Integration by Substitution
5.4: Integration by Parts
5.5: Other Options for Finding Algebraic Derivatives
5.6: Numerical Integration
5.7: Finding Antiderivatives and Evaluating Integrals (Exercises)

6: Using Definite Integrals

6.1: Using Definite Integrals to Find Area and Length
6.2: Using Definite Integrals to Find Volume
6.3: Density, Mass, and Center of Mass
6.4: Physics Applications - Work, Force, and Pressure
6.5: Improper Integrals
6.6: Using Definite Integrals (Exercises)

7: Differential Equations

7.1: An Introduction to Differential Equations
7.2: Qualitative Behavior of Solutions to Differential Equations
7.3: Euler's Method
7.4: Separable Differential Equations
7.5: Modeling with Differential Equations
7.6: Population Growth and the Logistic Equation
7.7: Differential Equations (Exercises)

8: Sequences and Series

8.1: Sequences
8.2: Geometric Series
8.3: Series of Real Numbers
8.4: Alternating Series
8.5: Taylor Polynomials and Taylor Series
8.6: Power Series
8.7: Sequences and Series (Exercises)

9: Multivariable and Vector Functions

9.1: Functions of Several Variables and Three Dimensional Space
9.2: Vectors
9.3: Dot Product
9.4: The Cross Product
9.5: Lines and Planes in Space
9.6: Vector-Valued Functions
9.7: Derivatives and Integrals of Vector-Valued Functions
9.8: Arc Length and Curvature

10: Derivatives of Multivariable Functions

10.1: Limits

10.2: First-Order Partial Derivatives

10.3: Second-Order Partial Derivatives

10.4: Linearization- Tangent Planes and Differentials

10.5: The Chain Rule

10.6: Directional Derivatives and the Gradient

10.7: Optimization

10.8: Constrained Optimization - Lagrange Multipliers

11: Multiple Integrals

11.1: Double Riemann Sums and Double Integrals over Rectangles

11.2: Iterated Integrals

11.3: Double Integrals over General Regions

11.4: Applications of Double Integrals

11.5: Double Integrals in Polar Coordinates

11.6: Surfaces Defined Parametrically and Surface Area

11.7: Triple Integrals

11.8: Triple Integrals in Cylindrical and Spherical Coordinates

11.9: Change of Variables

12: Appendices

12.1: A- Short Table of Integrals

12.2: B- Answers to Activities

12.3: C- Answers to Selected Exercises

Index

Glossary

Detailed Licensing

This page titled Active Calculus (Boelkins et al.) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin & Steven Schlicker via source content that was edited to the style and standards of the LibreTexts platform.

- Active Calculus (Boelkins et al.) by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

Licensing

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

This page titled Licensing is shared under a license and was authored, remixed, and/or curated by LibreTexts.

1: UNDERSTANDING THE DERIVATIVE

1.1: How do we Measure Velocity?

1.2: The Notion of Limit

1.3: The Derivative of a Function at a Point

1.4: The Derivative Function

1.5: Interpretating, Estimating, and Using the Derivative

1.6: The Second Derivative

1.7: Limits, Continuity, and Differentiability

1.8: The Tangent Line Approximation

1.9: Understanding the Derivative (Exercises)

This page titled 1: Understanding the Derivative is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1: Understanding the Derivative by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

1.1: How do we Measure Velocity?

Motivating Questions

- How is the average velocity of a moving object connected to the values of its position function?
- How do we interpret the average velocity of an object geometrically on the graph of its position function?
- How is the notion of instantaneous velocity connected to average velocity?

Calculus can be viewed broadly as the study of change. A natural and important question to ask about any changing quantity is "how fast is the quantity changing?"

We begin with a simple problem: a ball is tossed straight up in the air. How is the ball moving? Questions like this one are central to our study of differential calculus.

Preview Activity 1.1.1

Suppose that the height s of a ball at time t (in seconds) is given in feet by the formula $s(t) = 64 - 16(t-1)^2$.

1. Construct a graph of $y = s(t)$ on the time interval $0 \leq t \leq 3$. Label at least six distinct points on the graph, including the three points showing when the ball was released, when the ball reaches its highest point, and when the ball lands.
2. Describe the behavior of the ball on the time interval $0 < t < 1$ and on time interval $1 < t < 3$. What occurs at the instant $t = 1$?
3. Consider the expression

$$AV_{[0.5,1]} = \frac{s(1) - s(0.5)}{1 - 0.5}.$$

Compute the value of $AV_{[0.5,1]}$. What does this value measure on the graph? What does this value tell us about the motion of the ball? In particular, what are the units on $AV_{[0.5,1]}$?

Position and average velocity

Any moving object has a *position* that can be considered a function of *time*. When the motion is along a straight line, the position is given by a single variable, which we denote by $s(t)$. For example, $s(t)$ might give the mile marker of a car traveling on a straight highway at time t in hours. Similarly, the function s described in Preview Activity 1.1.1 is a position function, where position is measured vertically relative to the ground.

On any time interval, a moving object also has an *average velocity*. For example, to compute a car's average velocity we divide the number of miles traveled by the time elapsed, which gives the velocity in *miles per hour*. Similarly, the value of $AV_{[0.5,1]}$ in Preview Activity 1.1.1 gave the average velocity of the ball on the time interval $[0.5, 1]$, measured in feet per second.

In general, we make the following definition:

Average Velocity

For an object moving in a straight line with position function $s(t)$, the *average velocity of the object on the interval from* $t = a$ *to* $t = b$, denoted $AV_{[a,b]}$, is given by the formula

$$AV_{[a,b]} = \frac{s(b) - s(a)}{b - a}.$$

Note well: the units on $AV_{[a,b]}$ are "units of s per unit of t," such as "miles per hour" or "feet per second."

Activity 1.1.2

The following questions concern the position function given by $s(t) = 64 - 16(t-1)^2$, considered in Preview Activity 1.1.1.

1. Compute the average velocity of the ball on each of the following time intervals: $[0.4, 0.8]$, $[0.7, 0.8]$, $[0.79, 0.8]$, $[0.799, 0.8]$, $[0.8, 1.2]$, $[0.8, 0.9]$, $[0.8, 0.81]$, $[0.8, 0.801]$. Include units for each value.
2. On the graph provided in Figure 1.1.1, sketch the line that passes through the points $A = (0.4, s(0.4))$ and $B = (0.8, s(0.8))$. What is the meaning of the slope of this line? In light of this meaning, what is a geometric way to interpret each of the values computed in the preceding question?
3. Use a graphing utility to plot the graph of $s(t) = 64 - 16(t-1)^2$ on an interval containing the value $t = 0.8$. Then, zoom in repeatedly on the point $(0.8, s(0.8))$. What do you observe about how the graph appears as you view it more and more closely?
4. What do you conjecture is the velocity of the ball at the instant $t = 0.8$? Why?

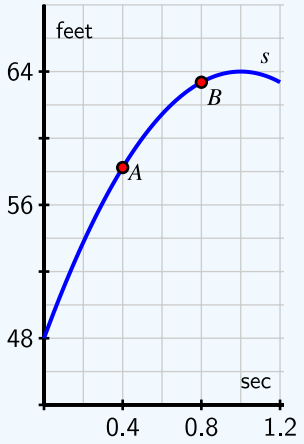


Figure 1.1.1 . A partial plot of $s(t) = 64 - 16(t-1)^2$.

Instantaneous Velocity

Whether we are driving a car, riding a bike, or throwing a ball, we have an intuitive sense that a moving object has a velocity at any given moment -- a number that measures how fast the object is moving *right now*. For instance, a car's speedometer tells the driver the car's velocity at any given instant. In fact, the velocity on a speedometer is really an average velocity that is computed over a very small time interval. If we let the time interval over which average velocity is computed become shorter and shorter, we can progress from average velocity to *instantaneous* velocity.

Informally, we define the *instantaneous velocity* of a moving object at time $t = a$ to be the value that the average velocity approaches as we take smaller and smaller intervals of time containing $t = a$. We will develop a more formal definition of instantaneous velocity soon, and this definition will be the foundation of much of our work in calculus. For now, it is fine to think of instantaneous velocity as follows: take average velocities on smaller and smaller time intervals around a specific point. If those average velocities approach a single number, then that number will be the instantaneous velocity at that point.

Activity 1.1.3

Each of the following questions concern $s(t) = 64 - 16(t-1)^2$, the position function from Preview Activity 1.1.1.

1. Compute the average velocity of the ball on the time interval $[1.5, 2]$. What is different between this value and the average velocity on the interval $[0, 0.5]$?

2. Use appropriate computing technology to estimate the instantaneous velocity of the ball at $t = 1.5$. Likewise, estimate the instantaneous velocity of the ball at $t = 2$. Which value is greater?
3. How is the sign of the instantaneous velocity of the ball related to its behavior at a given point in time? That is, what does positive instantaneous velocity tell you the ball is doing? Negative instantaneous velocity?
4. Without doing any computations, what do you expect to be the instantaneous velocity of the ball at $t = 1$? Why?

At this point we have started to see a close connection between average velocity and instantaneous velocity. Each is connected not only to the physical behavior of the moving object but also to the geometric behavior of the graph of the position function. We are interested in computing average velocities on the interval $[a, b]$ for smaller and smaller intervals. In order to make the link between average and instantaneous velocity more formal, think of the value b as $b = a + h$, where h is a small (non-zero) number that is allowed to vary. Then the average velocity of the object on the interval $[a, a + h]$ is

$$AV_{[a,a+h]} = \frac{s(a+h) - s(a)}{h},$$

with the denominator being simply h because $(a + h) - a = h$. Note that when $h < 0$, $AV_{[a,a+h]}$ measures the average velocity on the interval $[a + h, a]$.

To find the instantaneous velocity at $t = a$, we investigate what happens as the value of h approaches zero.

Example 1.1.2. Computing instantaneous velocity for a falling ball

The position function for a falling ball is given by $s(t) = 16 - 16t^2$ (where s is measured in feet and t in seconds).

1. Find an expression for the average velocity of the ball on a time interval of the form $[0.5, 0.5 + h]$ where $-0.5 < h < 0.5$ and $h \neq 0$.
2. Use this expression to compute the average velocity on $[0.5, 0.75]$ and $[0.4, 0.5]$.
3. Make a conjecture about the instantaneous velocity at $t = 0.5$.

Answer available at the end of this section.

Activity 1.1.4

For the function given by $s(t) = 64 - 16(t - 1)^2$ from Preview Activity 1.1.1, find the most simplified expression you can for the average velocity of the ball on the interval $[2, 2 + h]$. Use your result to compute the average velocity on $[1.5, 2]$ and to estimate the instantaneous velocity at $t = 2$. Finally, compare your earlier work in Activity 1.1.2.

Summary

- For an object moving in a straight line with position function $s(t)$, the *average velocity of the object on the interval from* $t = a$ *to* $t = b$, denoted $AV_{[a,b]}$, is given by the formula

$$AV_{[a,b]} = \frac{s(b) - s(a)}{b - a}.$$

- The average velocity on $[a, b]$ can be viewed geometrically as the slope of the line between the points $(a, s(a))$ and $(b, s(b))$ on the graph of $y = s(t)$, as shown in Figure 1.1.2.

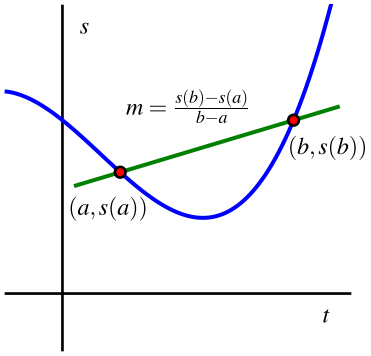


Figure 1.1.2. The graph of position function s together with the line through $(a, s(a))$ and $(b, s(b))$ whose slope is $m = \frac{s(b)-s(a)}{b-a}$. The line's slope is the average rate of change of s on the interval $[a, b]$.

- Given a moving object whose position at time t is given by a function s, the average velocity of the object on the time interval $[a, b]$ is given by $AV_{[a,b]} = \frac{s(b)-s(a)}{b-a}$. Viewing the interval $[a, b]$ as having the form $[a, a + h]$, we equivalently compute average velocity by the formula $AV_{[a,a+h]} = \frac{s(a+h)-s(a)}{h}$.
- The instantaneous velocity of a moving object at a fixed time is estimated by considering average velocities on shorter and shorter time intervals that contain the instant of interest.

Answers

Example 1.1.2. Computing instantaneous velocity for a falling ball

a. We make the assumptions that $-0.5 < h < 0.5$ and $h \neq 0$ because h cannot be zero (otherwise there is no interval on which to compute average velocity) and because the function only makes sense on the time interval $0 \leq t \leq 1$, as this is the duration of time during which the ball is falling. We want to compute and simplify

$$AV_{[0.5,0.5+h]} = \frac{s(0.5+h) - s(0.5)}{(0.5+h) - 0.5}.$$

We start by finding $s(0.5 + h)$. To do so, we follow the rule that defines the function s.

$$s(0.5+h) = 16 - 16(0.5+h)^2 \backslash[4pt] \quad = 16 - 16(0.25 + h + h^2) \backslash[4pt] = 16 - 4 - 16h - 16h^2 \backslash[4pt] \quad = 12 - 16h - 16h^2.$$

Now, returning to our computation of the average velocity, we find that

$$AV_{[0.5,0.5+h]} = \frac{s(0.5+h) - s(0.5)}{(0.5+h) - 0.5} \backslash[4pt] \quad = \frac{(12 - 16h - 16h^2) - (16 - 16(0.5)^2)}{0.5 + h - 0.5} \backslash[4pt] = \frac{12 - 16h - 16h^2 - 12}{h} \backslash[4pt] \quad = \frac{-16h - 16h^2}{h}.$$

At this point, we note two things: first, the expression for average velocity clearly depends on h, which it must, since as h changes the average velocity will change. Further, we note that since h can never equal zero, we may remove the common factor of h from the numerator and denominator. It follows that

$$AV_{[0.5,0.5+h]} = -16 - 16h.$$

b. From this expression we can compute the average for any small positive or negative value of h. For instance, to obtain the average velocity on $[0.5, 0.75]$, we let $h = 0.25$, and the average velocity is $-16 - 16(0.25) = -20$ ft/sec. To get the average velocity on $[0.4, 0.5]$, we let $h = -0.1$, and compute the average velocity as

$$-16 - 16(-0.1) = -14.4 \text{ ft/sec}.$$

c. We can even explore what happens to $AV_{[0.5,0.5+h]}$ as h gets closer and closer to zero. As h approaches zero, $-16h$ will also approach zero, so it appears that the instantaneous velocity of the ball at $t = 0.5$ should be -16 ft/sec.

This page titled 1.1: How do we Measure Velocity? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.1: How do we Measure Velocity? by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

1.2: The Notion of Limit

Motivating Questions

- What is the mathematical notion of *limit* and what role do limits play in the study of functions?
- What is the meaning of the notation $\lim_{x\to a} f(x) = L$?
- How do we go about determining the value of the limit of a function at a point?
- How do we manipulate average velocity to compute instantaneous velocity?

In Section 1.1 we used a function, $s(t)$, to model the location of a moving object at a given time. Functions can model other interesting phenomena, such as the rate at which an automobile consumes gasoline at a given velocity, or the reaction of a patient to a given dosage of a drug. We can use calculus to study how a function value changes in response to changes in the input variable.

Think about the falling ball whose position function is given by $s(t) = 64 - 16t^2$. Its average velocity on the interval $[1, x]$ is given by

$$AV_{[1,x]} = \frac{s(x) - s(1)}{x - 1} = \frac{(64 - 16x^2) - (64 - 16)}{x - 1} = \frac{16 - 16x^2}{x - 1}.$$

Note that the average velocity is a function of x. That is, the function $g(x) = \frac{16-16x^2}{x-1}$ tells us the average velocity of the ball on the interval from $t = 1$ to $t = x$. To find the instantaneous velocity of the ball when $t = 1$, we need to know what happens to $g(x)$ as x gets closer and closer to 1. But also notice that $g(1)$ is not defined, because it leads to the quotient $0/0$.

This is where the notion of a *limit* comes in. By using a limit, we can investigate the behavior of $g(x)$ as x gets arbitrarily close, but not equal, to 1. We first use the graph of a function to explore points where interesting behavior occurs.

Preview Activity 1.2.1.

Suppose that g is the function given by the graph below. Use the graph in Figure1.2.1 to answer each of the following questions.

1. Determine the values $g(-2)$, $g(-1)$, $g(0)$, $g(1)$, and $g(2)$, if defined. If the function value is not defined, explain what feature of the graph tells you this.
2. For each of the values $a = -1$, $a = 0$, and $a = 2$, complete the following sentence: "As x gets closer and closer (but not equal) to a, $g(x)$ gets as close as we want to _____."
3. What happens as x gets closer and closer (but not equal) to $a = 1$? Does the function $g(x)$ get as close as we would like to a single value?

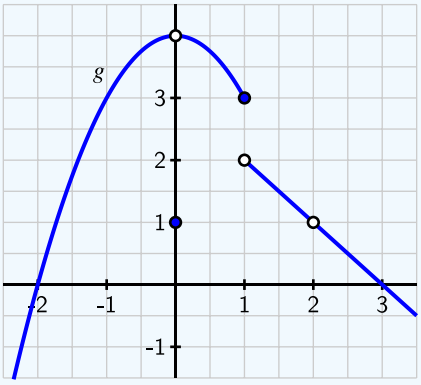


Figure1.2.1. Graph of $y = g(x)$ for Preview Activity 1.2.1.

Notion of Limit

Limits give us a way to identify a trend in the values of a function as its input variable approaches a particular value of interest. We need a precise understanding of what it means to say "a function f has limit L as x approaches a." To begin, think about a recent example.

In Preview Activity $1.2.1$, we saw that as x gets closer and closer (but not equal) to 0, $g(x)$ gets as close as we want to the value 4. At first, this may feel counterintuitive, because the value of $g(0)$ is 1, not 4. But limits describe the behavior of a function *arbitrarily close to* a fixed input, and the value of the function *at* the fixed input does not matter. More formally,[1] we say the following.

What follows here is not what mathematicians consider the formal definition of a limit. To be completely precise, it is necessary to quantify both what it means to say "as close to L as we like" and "sufficiently close to a." That can be accomplished through what is traditionally called the epsilon-delta definition of limits. The definition presented here is sufficient for the purposes of this text.

Definition 1.2.1

Given a function f, a fixed input $x = a$, and a real number L, we say that f *has limit* L *as* x *approaches* a, and write

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

provided that we can make $f(x)$ as close to L as we like by taking x sufficiently close (but not equal) to a. If we cannot make $f(x)$ as close to a single value as we would like as x approaches a, then we say that f ***does not have a limit as*** x ***approaches*** a.

Example 1.2.1

For the function g pictured in Figure$1.2.1$, we make the following observations:

$$\lim_{x \to -1} g(x) = 3, \quad \lim_{x \to 0} g(x) = 4, \quad \text{and} \quad \lim_{x \to 2} g(x) = 1.$$

When working from a graph, it suffices to ask if the function *approaches* a single value from each side of the fixed input. The function value *at* the fixed input is irrelevant. This reasoning explains the values of the three limits stated above.

However, g does not have a limit as $x \to 1$. There is a jump in the graph at $x = 1$. If we approach $x = 1$ from the left, the function values tend to get close to 3, but if we approach $x = 1$ from the right, the function values get close to 2. There is no single number that all of these function values approach. This is why the limit of g does not exist at $x = 1$.

For any function f, there are typically three ways to answer the question "does f have a limit at $x = a$, and if so, what is the limit?" The first is to reason graphically as we have just done with the example from Preview Activity $1.2.1$. If we have a formula for $f(x)$, there are two additional possibilities:

1. Evaluate the function at a sequence of inputs that approach a on either side (typically using some sort of computing technology), and ask if the sequence of outputs seems to approach a single value.
2. Use the algebraic form of the function to understand the trend in its output values as the input values approach a.

The first approach produces only an approximation of the value of the limit, while the latter can often be used to determine the limit exactly.

Example 1.2.2: Limits of Two Functions

For each of the following functions, we'd like to know whether or not the function has a limit at the stated a-values. Use both numerical and algebraic approaches to investigate and, if possible, estimate or determine the value of the limit. Compare the results with a careful graph of the function on an interval containing the points of interest.

1. $f(x) = \frac{4-x^2}{x+2}$; $a = -1$, $a = -2$
2. $g(x) = \sin\left(\frac{\pi}{x}\right)$; $a = 3$, $a = 0$

Answer available at the end of this section.

An important lesson to take from Example 1.2.2 is that tables can be misleading when determining the value of a limit. While a table of values is useful for investigating the possible value of a limit, we should also use other tools to confirm the value.

Activity 1.2.2

Estimate the value of each of the following limits by constructing appropriate tables of values. Then determine the exact value of the limit by using algebra to simplify the function. Finally, plot each function on an appropriate interval to check your result visually.

1. $\lim\limits_{x\to 1} \dfrac{x^2-1}{x-1}$
2. $\lim\limits_{x\to 0} \dfrac{(2+x)^3-8}{x}$
3. $\lim\limits_{x\to 0} \dfrac{\sqrt{x+1}-1}{x}$

Recall that our primary motivation for considering limits of functions comes from our interest in studying the rate of change of a function. To that end, we close this section by revisiting our previous work with average and instantaneous velocity and highlighting the role that limits play.

Instantaneous Velocity

Suppose that we have a moving object whose position at time t is given by a function s. We know that the average velocity of the object on the time interval $[a, b]$ is $AV_{[a,b]} = \frac{s(b)-s(a)}{b-a}$. We define the *instantaneous velocity* at a to be the limit of average velocity as b approaches a. Note particularly that as $b \to a$, the length of the time interval gets shorter and shorter (while always including a). We will write $IV_{t=a}$ for the instantaneous velocity at $t = a$, and thus

$$IV_{t=a} = \lim_{b\to a} AV_{[a,b]} = \lim_{b\to a} \frac{s(b)-s(a)}{b-a}.$$

Equivalently, if we think of the changing value b as being of the form $b = a + h$, where h is some small number, then we may instead write

$$IV_{t=a} = \lim_{h\to 0} AV_{[a,a+h]} = \lim_{h\to 0} \frac{s(a+h)-s(a)}{h}.$$

Again, the most important idea here is that to compute instantaneous velocity, we take a limit of average velocities as the time interval shrinks.

Activity 1.2.3

Consider a moving object whose position function is given by $s(t) = t^2$, where s is measured in meters and t is measured in minutes.

1. Determine the most simplified expression for the average velocity of the object on the interval $[3, 3 + h]$, where $h > 0$.
2. Determine the average velocity of the object on the interval $[3, 3.2]$. Include units on your answer.
3. Determine the instantaneous velocity of the object when $t = 3$. Include units on your answer.

The closing activity of this section asks you to make some connections among average velocity, instantaneous velocity, and slopes of certain lines.

Activity 1.2.4

For the moving object whose position s at time t is given by the graph in Figure1.2.4, answer each of the following questions. Assume that s is measured in feet and t is measured in seconds.

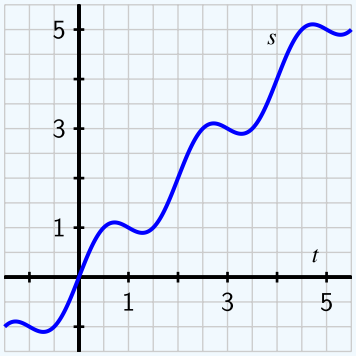


Figure1.2.4 . Plot of the position function $y = s(t)$ in Activity 1.2.4 .

a. Use the graph to estimate the average velocity of the object on each of the following intervals: $[0.5, 1]$, $[1.5, 2.5]$, $[0, 5]$. Draw each line whose slope represents the average velocity you seek.

b. How could you use average velocities or slopes of lines to estimate the instantaneous velocity of the object at a fixed time?

c. Use the graph to estimate the instantaneous velocity of the object when $t = 2$. Should this instantaneous velocity at $t = 2$ be greater or less than the average velocity on $[1.5, 2.5]$ that you computed in (a)? Why?

Summary

- Limits enable us to examine trends in function behavior near a specific point. In particular, taking a limit at a given point asks if the function values nearby tend to approach a particular fixed value.
- We read $\lim_{x\to a} f(x) = L$, as "the limit of f as x approaches a is L," which means that we can make the value of $f(x)$ as close to L as we want by taking x sufficiently close (but not equal) to a.
- To find $\lim_{x\to a} f(x)$ for a given value of a and a known function f, we can estimate this value from the graph of f, or we can make a table of function values for x-values that are closer and closer to a. If we want the exact value of the limit, we can work with the function algebraically to understand how different parts of the formula for f change as $x \to a$.
- We find the instantaneous velocity of a moving object at a fixed time by taking the limit of average velocities of the object over shorter and shorter time intervals containing the time of interest.

Answers

Example 1.2.2: Limits of Two Functions

a. We first construct a graph of f along with tables of values near $a = -1$ and $a = -2$.

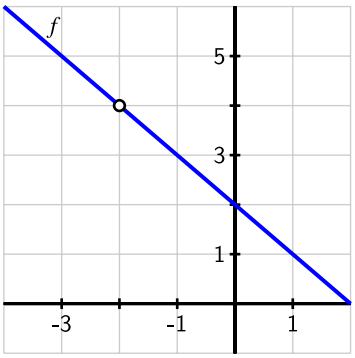
Table 1.2.1. Table of f values near $x = -1$.

x	$f(x)$
-0.9	2.9
-0.99	2.99
-0.999	2.999
-0.9999	2.9999

x	$f(x)$
-1.1	3.1
-1.01	3.01
-1.001	3.001
-1.0001	3.0001

Table $1.2.2$. *Table of* f *values near* $x = -2$.

x	$f(x)$
-1.9	3.9
-1.99	3.99
-1.999	3.999
-1.9999	3.9999
-2.1	4.1
-2.01	4.01
-2.001	4.001
-2.0001	4.0001



Figure$1.2.2$. Plot of $f(x)$ on $[-4, 2]$.

From Table $1.2.1$, it appears that we can make f as close as we want to 3 by taking x sufficiently close to -1, which suggests that $\lim_{x\to -1} f(x) = 3$. This is also consistent with the graph of f. To see this a bit more rigorously and from an algebraic point of view, consider the formula for f: $f(x) = \frac{4-x^2}{x+2}$. As $x \to -1$, $(4 - x^2)$ $\to (4 - (-1)^2) = 3$, and $(x + 2) \to (-1 + 2) = 1$, so as $x \to -1$, the numerator of f tends to 3 and the denominator tends to 1, hence $\lim_{x\to -1} f(x) = \frac{3}{1} = 3$.

The situation is more complicated when $x \to -2$, because $f(-2)$ is not defined. If we try to use a similar algebraic argument regarding the numerator and denominator, we observe that as $x \to -2$, $(4 - x^2) \to (4 - (-2)^2) = 0$, and $(x + 2) \to (-2 + 2) = 0$, so as $x \to -2$, the numerator and denominator of f both tend to 0. We call $0/0$ an *indeterminate form*. This tells us that there is somehow more work to do. From Table $1.2.2$ and Figure$1.2.2$, it appears that f should have a limit of 4 at $x = -2$.

To see algebraically why this is the case, observe that

$$\begin{aligned}\lim_{x\to -2} f(x) &= \lim_{x\to -2} \frac{4-x^2}{x+2} \\ &= \lim_{x\to -2} \frac{(2-x)(2+x)}{x+2}.\end{aligned}$$

It is important to observe that, since we are taking the limit as $x \to -2$, we are considering x values that are close, but not equal, to -2. Because we never actually allow x to equal -2, the quotient $\frac{2+x}{x+2}$ has value 1 for every possible value of x. Thus, we can simplify the most recent expression above, and find that

$$\lim_{x\to -2} f(x) = \lim_{x\to -2} 2 - x.$$

This limit is now easy to determine, and its value clearly is 4. Thus, from several points of view we've seen that $\lim_{x\to -2} f(x) = 4$.

b. Next we turn to the function g, and construct two tables and a graph.

Table 1.2.3 . *Table of* g *values near* $x = 3$.

x	$g(x)$
2.9	0.84864
2.99	0.86428
2.999	0.86585
2.9999	0.86601
3.1	0.88351
3.01	0.86777
3.001	0.86620
3.0001	0.86604

Table 1.2.4. *Table of* g *values near* $x = 0$.

x	$g(x)$
-0.1	0
-0.01	0
-0.001	0
-0.0001	0
0.1	0
0.01	0
0.001	0
0.0001	0

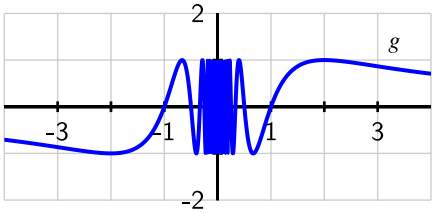


Figure1.2.3 . Plot of $g(x)$ on $[-4, 4]$.

First, as $x \to 3$, it appears from the table values that the function is approaching a number between 0.86601 and 0.86604. From the graph it appears that $g(x) \to g(3)$ as $x \to 3$. The exact value of $g(3) = \sin(\frac{\pi}{3})$ is $\frac{\sqrt{3}}{2}$, which is approximately 0.8660254038. This is convincing evidence that

$$\lim_{x \to 3} g(x) = \frac{\sqrt{3}}{2}.$$

As $x \to 0$, we observe that $\frac{\pi}{x}$ does not behave in an elementary way. When x is positive and approaching zero, we are dividing by smaller and smaller positive values, and $\frac{\pi}{x}$ increases without bound. When x is negative and approaching zero, $\frac{\pi}{x}$ decreases without bound. In this sense, as we get close to $x = 0$, the inputs to the sine function are growing rapidly, and this leads to increasingly rapid oscillations in the graph of g betweem 1 and -1. If we plot the function $g(x) = \sin\left(\frac{\pi}{x}\right)$ with a graphing utility and then zoom in on $x = 0$, we see that the function never settles down to a single value near the origin, which suggests that g does not have a limit at $x = 0$.

How do we reconcile the graph with the righthand table above, which seems to suggest that the limit of g as x approaches 0 may in fact be 0? The data misleads us because of the special nature of the sequence of input values $\{0.1, 0.01, 0.001, ...\}$. When we evaluate $g(10^{-k})$, we get $g(10^{-k}) = \sin\left(\frac{\pi}{10^{-k}}\right) = \sin(10^k \pi) = 0$ for each positive integer value of k. But if we take a different sequence of values approaching zero, say $\{0.3, 0.03, 0.003, ...\}$, then we find that

$$g(3 \cdot 10^{-k}) = \sin\left(\frac{\pi}{3 \cdot 10^{-k}}\right) = \sin\left(\frac{10^k \pi}{3}\right) = \frac{\sqrt{3}}{2} \approx 0.866025.$$

That sequence of function values suggests that the value of the limit is $\frac{\sqrt{3}}{2}$. Clearly the function cannot have two different values for the limit, so g has no limit as $x \to 0$.

This page titled 1.2: The Notion of Limit is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.2: The Notion of Limit by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

1.3: The Derivative of a Function at a Point

Motivating Questions

- How is the average rate of change of a function on a given interval defined, and what does this quantity measure?
- How is the instantaneous rate of change of a function at a particular point defined? How is the instantaneous rate of change linked to average rate of change?
- What is the derivative of a function at a given point? What does this derivative value measure? How do we interpret the derivative value graphically?
- How are limits used formally in the computation of derivatives?

The *instantaneous rate of change* of a function is an idea that sits at the foundation of calculus. It is a generalization of the notion of instantaneous velocity and measures how fast a particular function is changing at a given point. If the original function represents the position of a moving object, this instantaneous rate of change is precisely the velocity of the object. In other contexts, instantaneous rate of change could measure the number of cells added to a bacteria culture per day, the number of additional gallons of gasoline consumed by increasing a car's velocity one mile per hour, or the number of dollars added to a mortgage payment for each percentage point increase in interest rate. The instantaneous rate of change can also be interpreted geometrically on the function's graph, and this connection is fundamental to many of the main ideas in calculus.

Recall that for a moving object with position function s, its average velocity on the time interval $t = a$ to $t = a + h$ is given by the quotient

$$AV_{[a,a+h]} = \frac{s(a+h) - s(a)}{h}.$$

In a similar way, we make the following definition for an arbitrary function $y = f(x)$.

Definition 1.3.1: Average Rate of Change

For a function f, the ***average rate of change*** of f on the interval $[a, a + h]$ is given by the value

$$AV_{[a,a+h]} = \frac{f(a+h) - f(a)}{h}.$$

Equivalently, if we want to consider the average rate of change of f on $[a, b]$, we compute

$$AV_{[a,b]} = \frac{f(b) - f(a)}{b - a}.$$

It is essential that you understand how the average rate of change of f on an interval is connected to its graph.

Preview Activity 1.3.1

Suppose that f is the function given by the graph below and that a and $a + h$ are the input values as labeled on the x-axis. Use the graph in Figure 1.3.1 to answer the following questions.

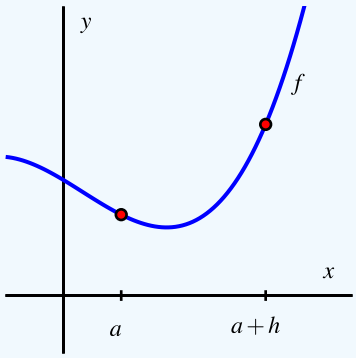


Figure 1.3.1. Plot of $y = f(x)$ for Preview Activity 1.3.1.

1. Locate and label the points $(a, f(a))$ and $(a + h, f(a + h))$ on the graph.
2. Construct a right triangle whose hypotenuse is the line segment from $(a, f(a))$ to $(a + h, f(a + h))$. What are the lengths of the respective legs of this triangle?
3. What is the slope of the line that connects the points $(a, f(a))$ and $(a + h, f(a + h))$?
4. Write a meaningful sentence that explains how the average rate of change of the function on a given interval and the slope of a related line are connected.

The Derivative of a Function at a Point

Just as we defined instantaneous velocity in terms of average velocity, we now define the instantaneous rate of change of a function at a point in terms of the average rate of change of the function f over related intervals. This instantaneous rate of change of f at a is called "the *derivative* of f at a," and is denoted by $f'(a)$.

Definition 1.3.2: Derivative

Let f be a function and $x = a$ a value in the function's domain. We define the *derivative of f with respect to x evaluated at $x = a$*, denoted $f'(a)$, by the formula

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

provided this limit exists.

Aloud, we read the symbol $f'(a)$ as either "f-prime at a" or "the derivative of f evaluated at $x = a$." Much of the next several chapters will be devoted to understanding, computing, applying, and interpreting derivatives. For now, we observe the following important things.

Note 1.3.1

- The derivative of f at the value $x = a$ is defined as the limit of the average rate of change of f on the interval $[a, a + h]$ as $h \to 0$. This limit may not exist, so not every function has a derivative at every point.
- We say that a function is *differentiable* at $x = a$ if it has a derivative at $x = a$.
- The derivative is a generalization of the instantaneous velocity of a position function: if $y = s(t)$ is a position function of a moving body, $s'(a)$ tells us the instantaneous velocity of the body at time $t = a$.
- Because the units on $\frac{f(a+h)-f(a)}{h}$ are "units of $f(x)$ per unit of x," the derivative has these very same units. For instance, if s measures position in feet and t measures time in seconds, the units on $s'(a)$ are feet per second.

- Because the quantity $\frac{f(a+h)-f(a)}{h}$ represents the slope of the line through $(a, f(a))$ and $(a+h, f(a+h))$, when we compute the derivative we are taking the limit of a collection of slopes of lines. Thus, the derivative itself represents the slope of a particularly important line.

We first consider the derivative at a given value as the slope of a certain line.

When we compute an instantaneous rate of change, we allow the interval $[a, a+h]$ to shrink as $h \to 0$. We can think of one endpoint of the interval as "sliding towards" the other. In particular, provided that f has a derivative at $(a, f(a))$, the point $(a+h, f(a+h))$ will approach $(a, f(a))$ as $h \to 0$. Because the process of taking a limit is a dynamic one, it can be helpful to use computing technology to visualize it. One option is a java applet in which the user is able to control the point that is moving. For a helpful collection of examples, consider the work of David Austin of Grand Valley State University, and this particularly relevant example. For applets that have been built in Geogebra [1], see Marc Renault's library via Shippensburg University, with this example being especially fitting for our work in this section.

You can even consider building your own examples; the fantastic program Geogebra is available for free download and is easy to learn and use.

Figure 1.3.2 shows a sequence of figures with several different lines through the points $(a, f(a))$ and $(a+h, f(a+h))$, generated by different values of h. These lines (shown in the first three figures in magenta), are often called *secant lines* to the curve $y = f(x)$. A secant line to a curve is simply a line that passes through two points on the curve. For each such line, the slope of the secant line is $m = \frac{f(a+h)-f(a)}{h}$, where the value of h depends on the location of the point we choose. We can see in the diagram how, as $h \to 0$, the secant lines start to approach a single line that passes through the point $(a, f(a))$. If the limit of the slopes of the secant lines exists, we say that the resulting value is the slope of the *tangent line* to the curve. This tangent line (shown in the right-most figure in green) to the graph of $y = f(x)$ at the point $(a, f(a))$ has slope $m = f'(a)$.

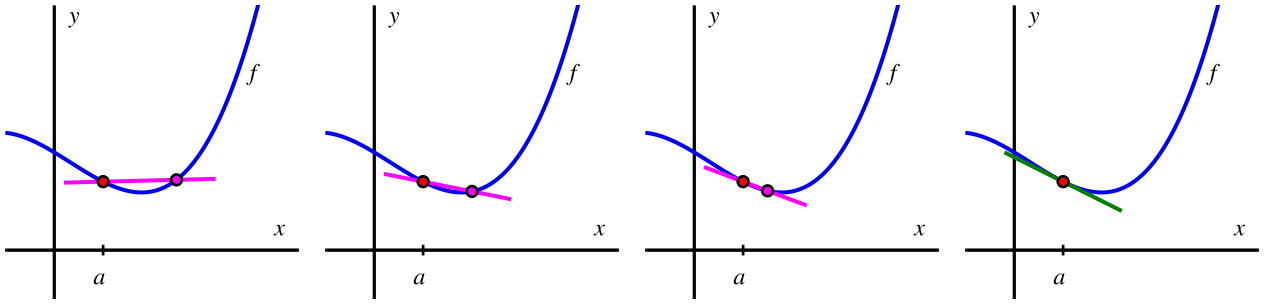


Figure 1.3.2. A sequence of secant lines approaching the tangent line to f at $(a, f(a))$.

If the tangent line at $x = a$ exists, the graph of f looks like a straight line when viewed up close at $(a, f(a))$. In Figure 1.3.3 we combine the four graphs in Figure 1.3.2 into the single one on the left, and zoom in on the box centered at $(a, f(a))$ on the right. Note how the tangent line sits relative to the curve $y = f(x)$ at $(a, f(a))$ and how closely it resembles the curve near $x = a$.

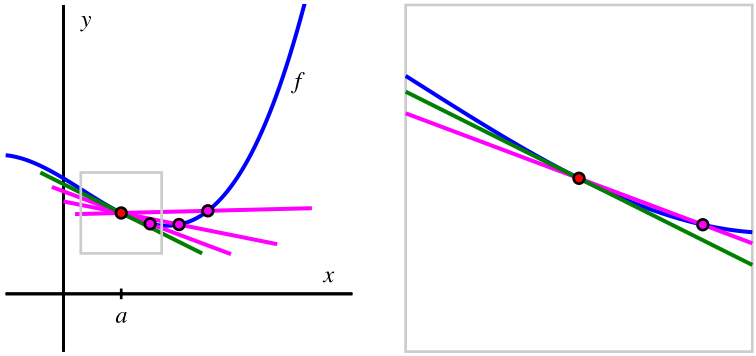


Figure 1.3.3. A sequence of secant lines approaching the tangent line to f at $(a, f(a))$. At right, we zoom in on the point $(a, f(a))$. The slope of the tangent line (in green) to f at $(a, f(a))$ is given by $f'(a)$.

Note 1.3.2

The instantaneous rate of change of f with respect to x at $x = a$, $f'(a)$, also measures the slope of the tangent line to the curve $y = f(x)$ at $(a, f(a))$.

The following example demonstrates several key ideas involving the derivative of a function.

Example 1.3.1

For the function $f(x) = x - x^2$, use the limit definition of the derivative to compute $f'(2)$. In addition, discuss the meaning of this value and draw a labeled graph that supports your explanation.

Answer available at the end of this section.

The following activities will help you explore a variety of key ideas related to derivatives.

Activity 1.3.2

Consider the function f whose formula is $f(x) = 3 - 2x$.

1. What familiar type of function is f? What can you say about the slope of f at every value of x?
2. Compute the average rate of change of f on the intervals $[1, 4]$, $[3, 7]$, and $[5, 5 + h]$; simplify each result as much as possible. What do you notice about these quantities?
3. Use the limit definition of the derivative to compute the exact instantaneous rate of change of f with respect to x at the value $a = 1$. That is, compute $f'(1)$ using the limit definition. Show your work. Is your result surprising?
4. Without doing any additional computations, what are the values of $f'(2)$, $f'(\pi)$, and $f'(-\sqrt{2})$? Why?

Activity 1.3.3

A water balloon is tossed vertically in the air from a window. The balloon's height in feet at time t in seconds after being launched is given by $s(t) = -16t^2 + 16t + 32$. Use this function to respond to each of the following questions.

a. Sketch an accurate, labeled graph of s on the axes provided in Figure 1.3.5. You should be able to do this without using computing technology.

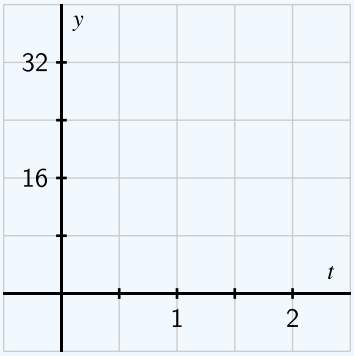


Figure 1.3.5. Axes for plotting $y = s(t)$ in Activity 1.3.3.

b. Compute the average rate of change of s on the time interval $[1, 2]$. Include units on your answer and write one sentence to explain the meaning of the value you found.

c. Use the limit definition to compute the instantaneous rate of change of s with respect to time, t, at the instant $a = 1$. Show your work using proper notation, include units on your answer, and write one sentence to explain the meaning of the value you found.

d. On your graph in (a), sketch two lines: one whose slope represents the average rate of change of s on $[1, 2]$, the other whose slope represents the instantaneous rate of change of s at the instant $a = 1$. Label each line clearly.

e. For what values of a do you expect $s'(a)$ to be positive? Why? Answer the same questions when "positive" is replaced by "negative" and "zero."

Activity 1.3.4

A rapidly growing city in Arizona has its population P at time t, where t is the number of decades after the year 2010, modeled by the formula $P(t) = 25000e^{t/5}$. Use this function to respond to the following questions.

a. Sketch an accurate graph of P for $t = 0$ to $t = 5$ on the axes provided in Figure 1.3.6. Label the scale on the axes carefully.

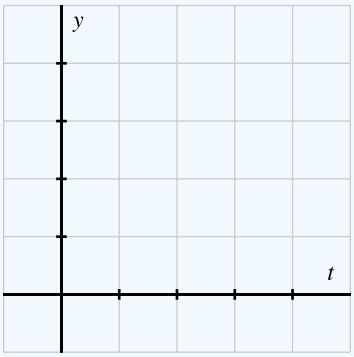


Figure 1.3.6. Axes for plotting $y = P(t)$ in Activity 1.3.4.

b. Compute the average rate of change of P between 2030 and 2050. Include units on your answer and write one sentence to explain the meaning (in everyday language) of the value you found.

c. Use the limit definition to write an expression for the instantaneous rate of change of P with respect to time, t, at the instant $a = 2$. Explain why this limit is difficult to evaluate exactly.

d. Estimate the limit in (c) for the instantaneous rate of change of P at the instant $a = 2$ by using several small h values. Once you have determined an accurate estimate of $P'(2)$, include units on your answer, and write one sentence (using everyday language) to explain the meaning of the value you found.

e. On your graph above, sketch two lines: one whose slope represents the average rate of change of P on $[2, 4]$, the other whose slope represents the instantaneous rate of change of P at the instant $a = 2$.

f. In a carefully-worded sentence, describe the behavior of $P'(a)$ as a increases in value. What does this reflect about the behavior of the given function P?

Summary

- The average rate of change of a function f on the interval $[a, b]$ is $\frac{f(b)-f(a)}{b-a}$. The units on the average rate of change are units of $f(x)$ per unit of x, and the numerical value of the average rate of change represents the slope of the secant line between the

points $(a, f(a))$ and $(b, f(b))$ on the graph of $y = f(x)$. If we view the interval as being $[a, a + h]$ instead of $[a, b]$, the meaning is still the same, but the average rate of change is now computed by $\frac{f(a+h)-f(a)}{h}$.

- The instantaneous rate of change with respect to x of a function f at a value $x = a$ is denoted $f'(a)$ (read "the derivative of f evaluated at a" or "f-prime at a") and is defined by the formula

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

provided the limit exists. Note particularly that the instantaneous rate of change at $x = a$ is the limit of the average rate of change on $[a, a + h]$ as $h \to 0$.

- Provided the derivative $f'(a)$ exists, its value tells us the instantaneous rate of change of f with respect to x at $x = a$, which geometrically is the slope of the tangent line to the curve $y = f(x)$ at the point $(a, f(a))$. We even say that $f'(a)$ is the "slope of the curve" at the point $(a, f(a))$.
- Limits allow us to move from the rate of change over an interval to the rate of change at a single point.

Answers

Example 1.3.1

From the limit definition, we know that

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

Now we use the rule for f, and observe that $f(2) = 2 - 2^2 = -2$ and $f(2 + h) = (2 + h) - (2 + h)^2$. Substituting these values into the limit definition, we have that

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

In order to let $h \to 0$, we must simplify the quotient. Expanding and distributing in the numerator,

$$f'(2) = \lim_{h \to 0} \frac{2 + h - 4 - 4h - h^2 + 2}{h}.$$

Combining like terms, we have

$$f'(2) = \lim_{h \to 0} \frac{-3h - h^2}{h}.$$

Next, we remove a common factor of h in both the numerator and denominator and find that

$$f'(2) = \lim_{h \to 0} (-3 - h).$$

Finally, we are able to take the limit as $h \to 0$, and thus conclude that $f'(2) = -3$. We note that $f'(2)$ is the instantaneous rate of change of f at the point $(2, -2)$. It is also the slope of the tangent line to the graph of $y = x - x^2$ at the point $(2, -2)$. Figure 1.3.4 shows both the function and the line through $(2, -2)$ with slope $m = f'(2) = -3$.

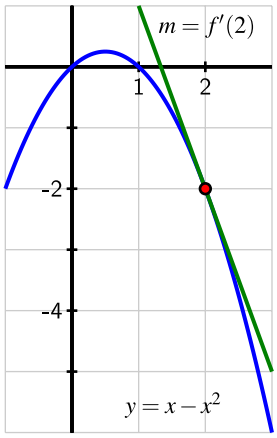


Figure 1.3.4. The tangent line to $y = x - x^2$ at the point $(2, -2)$.

This page titled 1.3: The Derivative of a Function at a Point is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.3: The Derivative of a Function at a Point by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

1.4: The Derivative Function

Motivating Questions

- How does the limit definition of the derivative of a function f lead to an entirely new (but related) function f'?
- What is the difference between writing $f'(a)$ and $f'(x)$?
- How is the graph of the derivative function $f'(x)$ related to the graph of $f(x)$?
- What are some examples of functions f for which f' is not defined at one or more points?

We now know that the instantaneous rate of change of a function $f(x)$ at $x = a$, or equivalently the slope of the tangent line to the graph of $y = f(x)$ at $x = a$, is given by the value $f'(a)$. In all of our examples so far, we have identified a particular value of a as our point of interest: $a = 1$, $a = 3$, etc. But it is not hard to imagine that we will often be interested in the derivative value for more than just one a-value, and possibly for many of them. In this section, we explore how we can move from computing the derivative at a single point to computing a formula for $f'(a)$ at any point a. Indeed, the process of "taking the derivative" generates a new function, denoted by $f'(x)$, derived from the original function $f(x)$.

Preview Activity 1.4.1

Consider the function $f(x) = 4x - x^2$.

1. Use the limit definition to compute the derivative values: $f'(0)$, $f'(1)$, $f'(2)$, and $f'(3)$.
2. Observe that the work to find $f'(a)$ is the same, regardless of the value of a. Based on your work in (a), what do you conjecture is the value of $f'(4)$? How about $f'(5)$? (Note: you should *not* use the limit definition of the derivative to find either value.)
3. Conjecture a formula for $f'(a)$ that depends only on the value a. That is, in the same way that we have a formula for $f(x)$ (recall $f(x) = 4x - x^2$), see if you can use your work above to guess a formula for $f'(a)$ in terms of a.

How the derivative is itself a function

In your work in Preview Activity 1.4.1 with $f(x) = 4x - x^2$, you may have found several patterns. One comes from observing that $f'(0) = 4$, $f'(1) = 2$, $f'(2) = 0$, and $f'(3) = -2$. That sequence of values leads us naturally to conjecture that $f'(4) = -4$ and $f'(5) = -6$. We also observe that the particular value of a has very little effect on the process of computing the value of the derivative through the limit definition. To see this more clearly, we compute $f'(a)$, where a represents a number to be named later. Following the now standard process of using the limit definition of the derivative,

$$\begin{aligned} f'(a) &= \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} = \lim_{h\to 0} \frac{4(a+h) - (a+h)^2 - (4a - a^2)}{h} \\ &= \lim_{h\to 0} \frac{4a + 4h - a^2 - 2ha - h^2 - 4a + a^2}{h} = \lim_{h\to 0} \frac{4h - 2ha - h^2}{h} \\ &= \lim_{h\to 0} \frac{h(4 - 2a - h)}{h} = \lim_{h\to 0} (4 - 2a - h). \end{aligned}$$

Here we observe that neither 4 nor $2a$ depend on the value of h, so as $h \to 0$, $(4 - 2a - h) \to (4 - 2a)$. Thus, $f'(a) = 4 - 2a$.

This result is consistent with the specific values we found above: e.g., $f'(3) = 4 - 2(3) = -2$. And indeed, our work confirms that the value of a has almost no bearing on the process of computing the derivative. We note further that the letter being used is immaterial: whether we call it a, x, or anything else, the derivative at a given value is simply given by "4 minus 2 times the value." We choose to use x for consistency with the original function given by $y = f(x)$, as well as for the purpose of graphing the derivative function. For the function $f(x) = 4x - x^2$, it follows that $f'(x) = 4 - 2x$.

Because the value of the derivative function is linked to the graph of the original function, it makes sense to look at both of these functions plotted on the same domain.

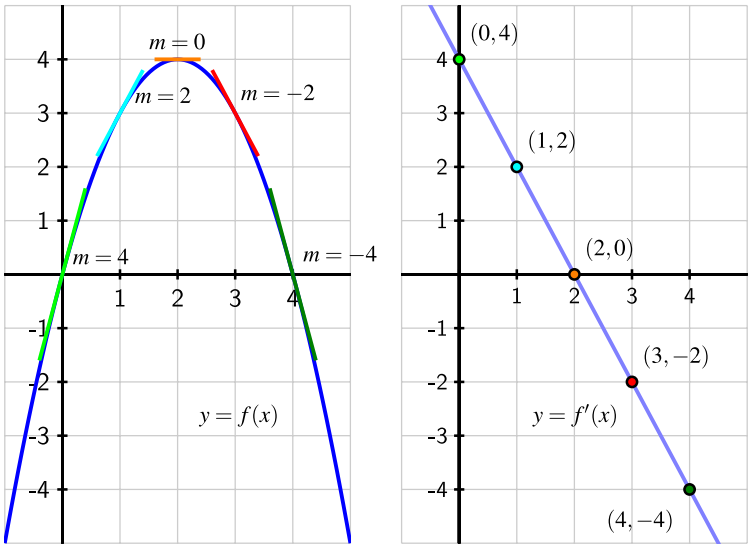


Figure 1.4.1 . The graphs of $f(x) = 4x - x^2$ (at left) and $f'(x) = 4 - 2x$ (at right). Slopes on the graph of f correspond to heights on the graph of f'.

In Figure 1.4.1 , on the left we show a plot of $f(x) = 4x - x^2$ together with a selection of tangent lines at the points we've considered above. On the right, we show a plot of $f'(x) = 4 - 2x$ with emphasis on the heights of the derivative graph at the same selection of points. Notice the connection between colors in the left and right graphs: the green tangent line on the original graph is tied to the green point on the right graph in the following way: *the slope of the tangent line* at a point on the lefthand graph is the same as the *height* at the corresponding point on the righthand graph. That is, at each respective value of x, the slope of the tangent line to the original function is the same as the height of the derivative function. Do note, however, that the units on the vertical axes are different: in the left graph, the vertical units are simply the output units of f. On the righthand graph of $y = f'(x)$, the units on the vertical axis are units of f per unit of x.

An excellent way to explore how the graph of $f(x)$ generates the graph of $f'(x)$ is through a java applet. See, for instance, the applets at http://gvsu.edu/s/5C or http://gvsu.edu/s/5D, via the sites of Austin and Renault [1].

David Austin, http://gvsu.edu/s/5r; Marc Renault, http://gvsu.edu/s/5p.

In Section 1.3 when we first defined the derivative, we wrote the definition in terms of a value a to find $f'(a)$. As we have seen above, the letter a is merely a placeholder, and it often makes more sense to use x instead. For the record, here we restate the definition of the derivative.

Definition 1.4.1

Let f be a function and x a value in the function's domain. We define the *derivative of* f, a new function called f', by the formula $f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$, provided this limit exists.

We now have two different ways of thinking about the derivative function:

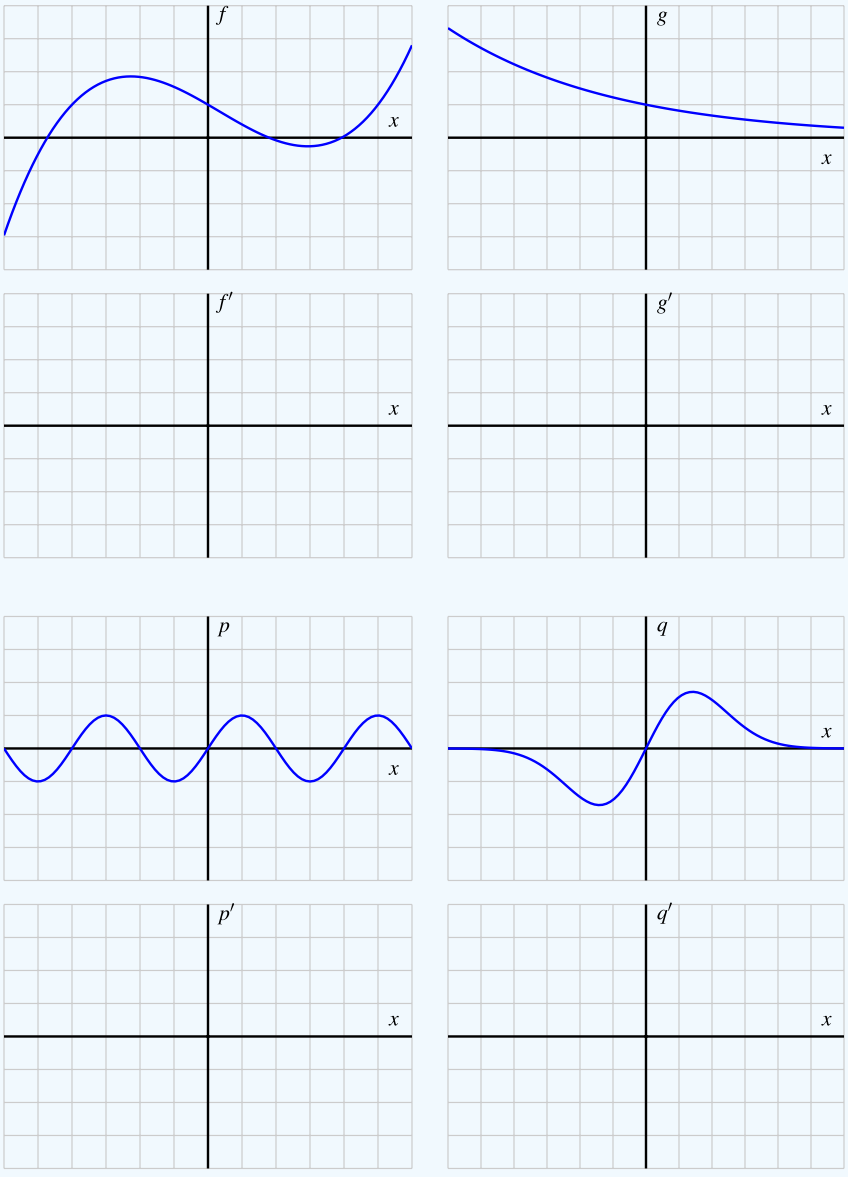
1. given a graph of $y = f(x)$, how does this graph lead to the graph of the derivative function $y = f'(x)$? and
2. given a formula for $y = f(x)$, how does the limit definition of derivative generate a formula for $y = f'(x)$?

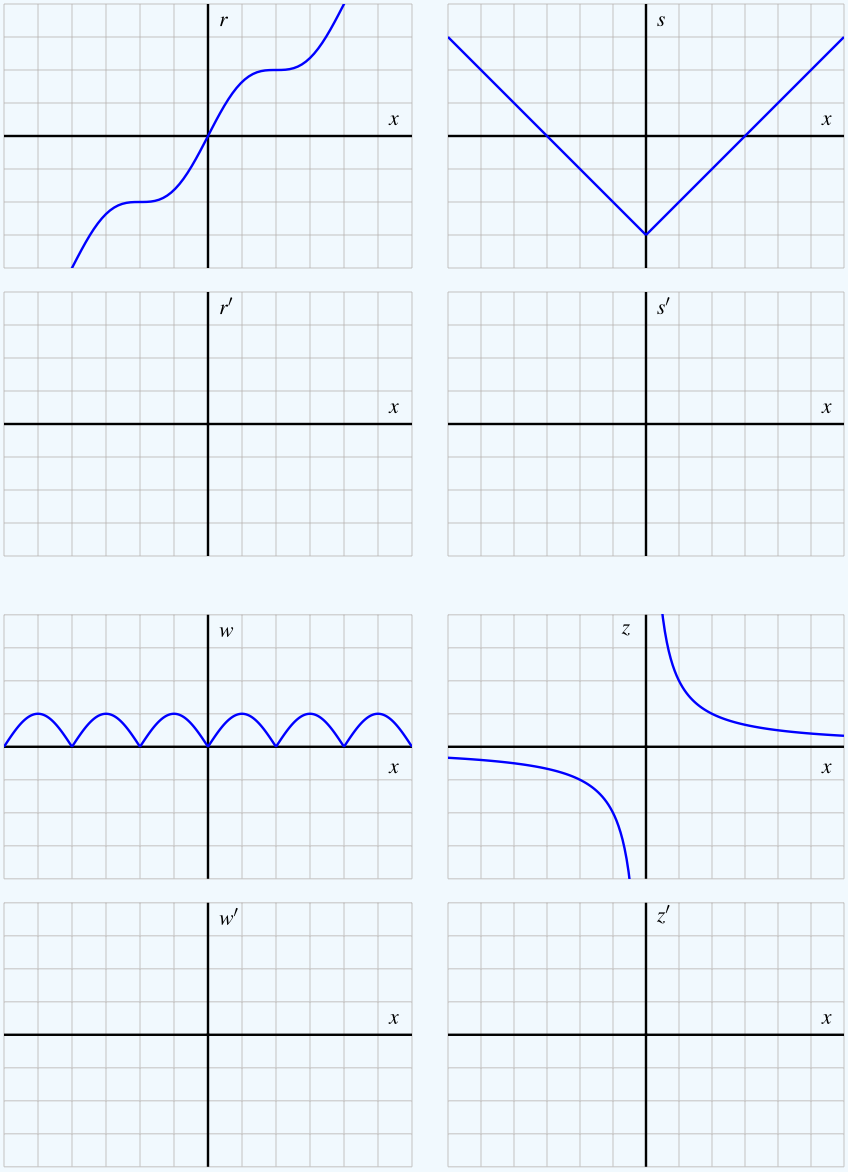
Both of these issues are explored in the following activities.

Activity 1.4.2

For each given graph of $y = f(x)$, sketch an approximate graph of its derivative function, $y = f'(x)$, on the axes immediately below. The scale of the grid for the graph of f is 1×1; assume the horizontal scale of the grid for the graph of f' is identical to that for f. If necessary, adjust and label the vertical scale on the axes for f'.

When you are finished with all 8 graphs, write several sentences that describe your overall process for sketching the graph of the derivative function, given the graph the original function. What are the values of the derivative function that you tend to identify first? What do you do thereafter? How do key traits of the graph of the derivative function exemplify properties of the graph of the original function?





For a dynamic investigation that allows you to experiment with graphing f' when given the graph of f, see http://gvsu.edu/s/8y.[2]

Marc Renault, Calculus Applets Using Geogebra.

Now, recall the opening example of this section: we began with the function $y = f(x) = 4x - x^2$ and used the limit definition of the derivative to show that $f'(a) = 4 - 2a$, or equivalently that $f'(x) = 4 - 2x$. We subsequently graphed the functions f and f' as shown in Figure $1.4.1$. Following Activity $1.4.2$, we now understand that we could have constructed a fairly accurate graph of $f'(x)$ *without* knowing a formula for either f or f'. At the same time, it is useful to know a formula for the derivative function whenever it is possible to find one.

In the next activity, we further explore the more algebraic approach to finding $f'(x)$: given a formula for $y = f(x)$, the limit definition of the derivative will be used to develop a formula for $f'(x)$.

Activity 1.4.3

For each of the listed functions, determine a formula for the derivative function. For the first two, determine the formula for the derivative by thinking about the nature of the given function and its slope at various points; do not use the limit definition. For the latter four, use the limit definition. Pay careful attention to the function names and independent variables. It is important to be comfortable with using letters other than f and x. For example, given a function $p(z)$, we call its derivative $p'(z)$.

1. $f(x) = 1$
2. $g(t) = t$
3. $p(z) = z^2$
4. $q(s) = s^3$
5. $F(t) = \dfrac{1}{t}$
6. $G(y) = \sqrt{y}$

Summary

- The limit definition of the derivative, $f'(x) = \lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$, produces a value for each x at which the derivative is defined, and this leads to a new function $y = f'(x)$. It is especially important to note that taking the derivative is a process that starts with a given function (f) and produces a new, related function (f').
- There is essentially no difference between writing $f'(a)$ (as we did regularly in Section 1.3) and writing $f'(x)$. In either case, the variable is just a placeholder that is used to define the rule for the derivative function.
- Given the graph of a function $y = f(x)$, we can sketch an approximate graph of its derivative $y = f'(x)$ by observing that *heights* on the derivative's graph correspond to *slopes* on the original function's graph.
- In Activity 1.4.2 , we encountered some functions that had sharp corners on their graphs, such as the shifted absolute value function. At such points, the derivative fails to exist, and we say that f is not differentiable there. For now, it suffices to understand this as a consequence of the jump that must occur in the derivative function at a sharp corner on the graph of the original function.

This page titled 1.4: The Derivative Function is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.4: The Derivative Function by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

1.5: Interpretating, Estimating, and Using the Derivative

Motivating Questions

- In contexts other than the position of a moving object, what does the derivative of a function measure?
- What are the units on the derivative function f', and how are they related to the units of the original function f?
- What is a central difference, and how can one be used to estimate the value of the derivative at a point from given function data?
- Given the value of the derivative of a function at a point, what can we infer about how the value of the function changes nearby?

It is a powerful feature of mathematics that it can be studied both as abstract discipline and as an applied one. For instance, calculus can be developed almost entirely as an abstract collection of ideas that focus on properties of functions. At the same time, if we consider functions that represent meaningful processes, calculus can describe our experience of physical reality. We have already seen that for the position function $y = s(t)$ of a ball being tossed straight up in the air, the derivative of the position function, $v(t) = s'(t)$, gives the ball's velocity at time t.

In this section, we investigate several functions with specific physical meaning, and consider how the units on the independent variable, dependent variable, and the derivative function add to our understanding. To start, we consider the familiar problem of a position function of a moving object.

Preview Activity 1.5.1

One of the longest stretches of straight (and flat) road in North America can be found on the Great Plains in the state of North Dakota on state highway 46, which lies just south of the interstate highway I-94 and runs through the town of Gackle. A car leaves town (at time $t = 0$) and heads east on highway 46; its position in miles from Gackle at time t in minutes is given by the graph of the function in Figure 1.5.1 . Three important points are labeled on the graph; where the curve looks linear, assume that it is indeed a straight line.

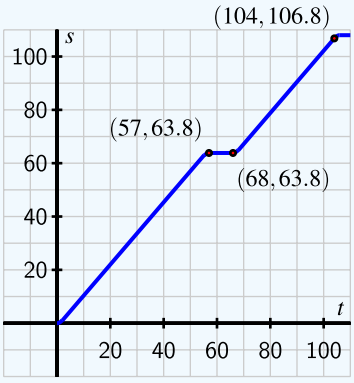


Figure 1.5.1. The graph of $y = s(t)$, the position of the car along highway 46, which tells its distance in miles from Gackle, ND, at time t in minutes.

1. In everyday language, describe the behavior of the car over the provided time interval. In particular, discuss what is happening on the time intervals $[57, 68]$ and $[68, 104]$.
2. Find the slope of the line between the points $(57, 63.8)$ and $(104, 106.8)$. What are the units on this slope? What does the slope represent?
3. Find the average rate of change of the car's position on the interval $[68, 104]$. Include units on your answer.
4. Estimate the instantaneous rate of change of the car's position at the moment $t = 80$. Write a sentence to explain your reasoning and the meaning of this value.

Units of the derivative function

As we now know, the derivative of the function f at a fixed value x is given by

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

and this value has several different interpretations. If we set $x = a$, one meaning of $f'(a)$ is the slope of the tangent line at the point $(a, f(a))$.

We also sometimes write $\frac{df}{dx}$ or $\frac{dy}{dx}$ instead of $f'(x)$, and these alternate notations help us see the units (and thus the meaning) of the derivative as *the instantaneous rate of change of f with respect to x*. The units on the slope of the secant line, $\frac{f(x+h)-f(x)}{h}$, are "units of y per unit of x," and when we take the limit as h goes to zero, the derivative $f'(x)$ has the same units: units of y per unit of x. It is helpful to remember that the units on the derivative function are "units of output per unit of input," for the variables of the original function.

For example, suppose that the function $y = P(t)$ measures the population of a city (in thousands) at the start of year t (where $t = 0$ corresponds to 2010 AD). We are told that $P'(2) = 21.37$. What is the meaning of this value? Well, since P is measured in thousands and t is measured in years, we can say that the instantaneous rate of change of the city's population with respect to time at the start of 2012 is 21.37 thousand people per year. We therefore expect that in the coming year, about 21,370 people will be added to the city's population.

Toward more accurate derivative estimates

Recall that to estimate the value of $f'(x)$ at a given x, we calculate a *difference quotient* $\frac{f(x+h)-f(x)}{h}$ with a relatively small value of h. We should use both positive and negative values of h in order to account for the behavior of the function on both sides of the point of interest. To that end, we introduce the notion of a *central difference* and its role in estimating derivatives.

Example 1.5.1

Suppose that $y = f(x)$ is a function for which three values are known: $f(1) = 2.5$, $f(2) = 3.25$, and $f(3) = 3.625$. Estimate $f'(2)$.

Answer available at the end of this section.

The intuitive approach to average the two estimates found in Example 1.5.1 is in fact the best possible way estimate to a derivative when we have just two function values for f on opposite sides of the point of interest.

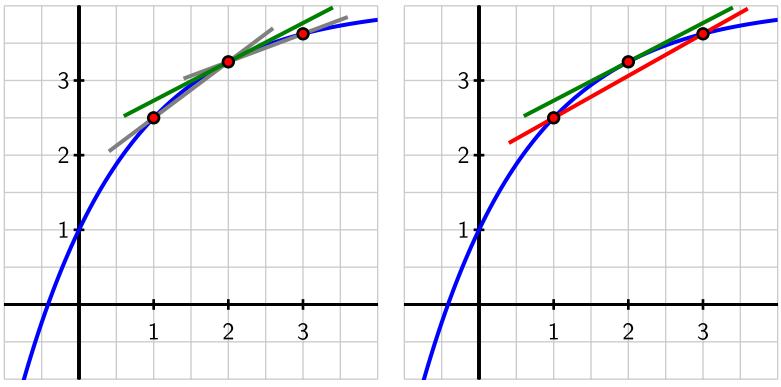


Figure 1.5.2 . At left, the graph of $y = f(x)$ along with the secant line through $(1, 2.5)$ and $(2, 3.25)$, the secant line through $(2, 3.25)$ and $(3, 3.625)$, as well as the tangent line. At right, the same graph along with the secant line through $(1, 2.5)$ and $(3, 3.625)$, plus the tangent line.

To see why, we think about the diagram in Figure 1.5.2 . On the left, we see the two secant lines with slopes that come from computing the *backward difference* $\frac{f(1)-f(2)}{1-2} = 0.75$ and from the *forward difference* $\frac{f(3)-f(2)}{3-2} = 0.375$. Note how the first slope

over-estimates the slope of the tangent line at $(2, f(2))$, while the second slope underestimates $f'(2)$. On the right, we see the secant line whose slope is given by the *central difference*

$$\frac{f(3)-f(1)}{3-1} = \frac{3.625-2.5}{2} = \frac{1.125}{2} = 0.5625.$$

Note that this central difference has the same value as the average of the forward and backward differences (and it is straightforward to explain why this always holds). The central difference yields a very good approximation to the derivative's value, because it yields a line closer to being parallel to the tangent line.

The ***central difference approximation*** to the value of the first derivative is given by

$$f'(a) \approx \frac{f(a+h)-f(a-h)}{2h}.$$

This quantity measures the slope of the secant line to $y = f(x)$ through the points $(a-h, f(a-h))$ and $(a+h, f(a+h))$.

Activity 1.5.2

A potato is placed in an oven, and the potato's temperature F (in degrees Fahrenheit) at various points in time is taken and recorded in the following table. Time t is measured in minutes.

Table 1.5.1. Temperature data in degrees Fahrenheit.

t	0	15	30	45	60	75	90
$F(t)$	70	180.5	251	296	324.5	342.8	354.5

1. Use a central difference to estimate the instantaneous rate of change of the temperature of the potato at $t = 30$. Include units on your answer.
2. Use a central difference to estimate the instantaneous rate of change of the temperature of the potato at $t = 60$. Include units on your answer.
3. Without doing any calculation, which do you expect to be greater: $F'(75)$ or $F'(90)$? Why?
4. Suppose it is given that $F(64) = 330.28$ and $F'(64) = 1.341$. What are the units on these two quantities? What do you expect the temperature of the potato to be when $t = 65$? when $t = 66$? Why?
5. Write a couple of careful sentences that describe the behavior of the temperature of the potato on the time interval $[0, 90]$, as well as the behavior of the instantaneous rate of change of the temperature of the potato on the same time interval.

Activity 1.5.3

A company manufactures rope, and the total cost of producing r feet of rope is $C(r)$ dollars.

1. What does it mean to say that $C(2000) = 800$?
2. What are the units of $C'(r)$?
3. Suppose that $C(2000) = 800$ and $C'(2000) = 0.35$. Estimate $C(2100)$, and justify your estimate by writing at least one sentence that explains your thinking.
4. Do you think $C'(2000)$ is less than, equal to, or greater than $C'(3000)$? Why?
5. Suppose someone claims that $C'(5000) = -0.1$. What would the practical meaning of this derivative value tell you about the approximate cost of the next foot of rope? Is this possible? Why or why not?

Activity 1.5.4

Researchers at a major car company have found a function that relates gasoline consumption to speed for a particular model of car. In particular, they have determined that the consumption C, in *liters per kilometer*, at a given speed s, is given by a function $C = f(s)$, where s is the car's speed in *kilometers per hour*.

1. Data provided by the car company tells us that $f(80) = 0.015$, $f(90) = 0.02$, and $f(100) = 0.027$. Use this information to estimate the instantaneous rate of change of fuel consumption with respect to speed at $s = 90$. Be as accurate as possible, use proper notation, and include units on your answer.
2. By writing a complete sentence, interpret the meaning (in the context of fuel consumption) of "$f(80) = 0.015$."
3. Write at least one complete sentence that interprets the meaning of the value of $f'(90)$ that you estimated in (a).

In Section 1.4, we learned how use to the graph of a given function f to plot the graph of its derivative, f'. It is important to remember that when we do so, the scale and the units on the vertical axis often have to change to represent f'. For example, suppose that $P(t) = 400 - 330e^{-0.03t}$ tells us the temperature in degrees Fahrenheit of a potato in an oven at time t in minutes. In Figure $1.5.3$, we sketch the graph of P on the left and the graph of P' on the right.

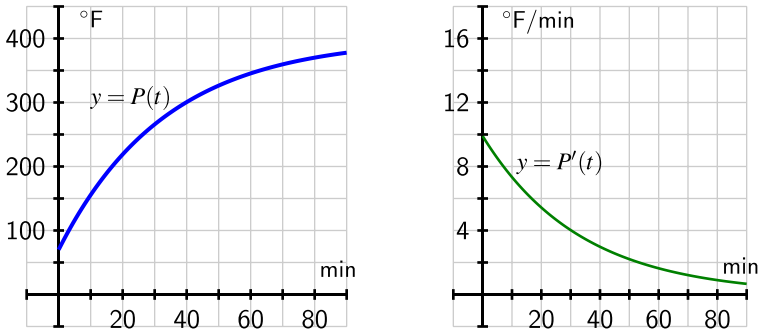


Figure $1.5.3$. Plot of $P(t) = 400 - 330e^{-0.03t}$ at left, and its derivative $P'(t)$ at right.

Notice that the vertical scales are different in size and different in units, as the units of P are °F, while those of P' are °F/min.

Summary

- The derivative of a given function $y = f(x)$ measures the instantaneous rate of change of the output variable with respect to the input variable.
- The units on the derivative function $y = f'(x)$ are units of y per unit of x. Again, this measures how fast the output of the function f changes when the input of the function changes.
- The central difference approximation to the value of the first derivative is given by

$$f'(a) \approx \frac{f(a+h) - f(a-h)}{2h}.$$

This quantity measures the slope of the secant line to $y = f(x)$ through the points $(a-h, f(a-h))$ and $(a+h, f(a+h))$. The central difference generates a good approximation of the derivative's value.

Answers

Example $1.5.1$

We know that $f'(2) = \lim_{h\to 0} \frac{f(2+h)-f(2)}{h}$. But since we don't have a graph or a formula for the function, we can neither sketch a tangent line nor evaluate the limit algebraically. We can't even use smaller and smaller values of h to estimate the limit. Instead, we have just two choices: using $h = -1$ or $h = 1$, depending on which point we pair with $(2, 3.25)$.

So, one estimate is

$$f'(2) \approx \frac{f(1) - f(2)}{1 - 2} = \frac{2.5 - 3.25}{-1} = 0.75.$$

The other is

$$f'(2) \approx \frac{f(3) - f(2)}{3 - 2} = \frac{3.625 - 3.25}{1} = 0.375.$$

Because the first approximation looks backward from the point $(2, 3.25)$ and the second approximation looks forward, it makes sense to average these two estimates in order to account for behavior on both sides of $x = 2$. Doing so, we find that

$$f'(2) \approx \frac{0.75 + 0.375}{2} = 0.5625.$$

This page titled 1.5: Interpretating, Estimating, and Using the Derivative is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.5: Interpretating, Estimating, and Using the Derivative by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

1.6: The Second Derivative

Motivating Questions

- How does the derivative of a function tell us whether the function is increasing or decreasing on an interval?
- What can we learn by taking the derivative of the derivative (the *second* derivative) of a function f?
- What does it mean to say that a function is concave up or concave down? How are these characteristics connected to certain properties of the derivative of the function?
- What are the units of the second derivative? How do they help us understand the rate of change of the rate of change?

Given a differentiable function $y = f(x)$, we know that its derivative, $y = f'(x)$, is a related function whose output at $x = a$ tells us the slope of the tangent line to $y = f(x)$ at the point $(a, f(a))$. That is, heights on the derivative graph tell us the values of slopes on the original function's graph.

At a point where $f'(x)$ is positive, the slope of the tangent line to f is positive. Therefore, on an interval where $f'(x)$ is positive, the function f is increasing (or rising). Similarly, if $f'(x)$ is negative on an interval, the graph of f is decreasing (or falling).

The derivative of f tells us not only *whether* the function f is increasing or decreasing on an interval, but also *how* the function f is increasing or decreasing. Look at the two tangent lines shown in Figure 1.6.1. We see that near point A the value of $f'(x)$ is positive and relatively close to zero, and near that point the graph is rising slowly. By contrast, near point B, the derivative is negative and relatively large in absolute value, and f is decreasing rapidly near B.

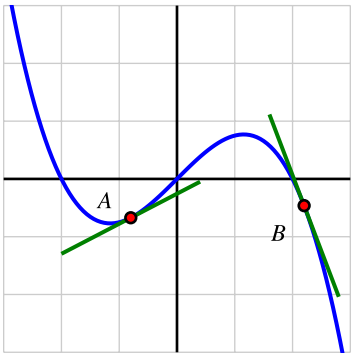


Figure 1.6.1. Two tangent lines on a graph.

Besides asking whether the value of the derivative function is positive or negative and whether it is large or small, we can also ask “how is the derivative changing?”

Because the derivative, $y = f'(x)$, is itself a function, we can consider taking its derivative — the derivative of the derivative — and ask “what does the derivative of the derivative tell us about how the original function behaves?” We start with an investigation of a moving object.

Preview Activity 1.6.1

The position of a car driving along a straight road at time t in minutes is given by the function $y = s(t)$ that is pictured in Figure 1.6.2. The car's position function has units measured in thousands of feet. For instance, the point $(2, 4)$ on the graph indicates that after 2 minutes, the car has traveled 4000 feet.

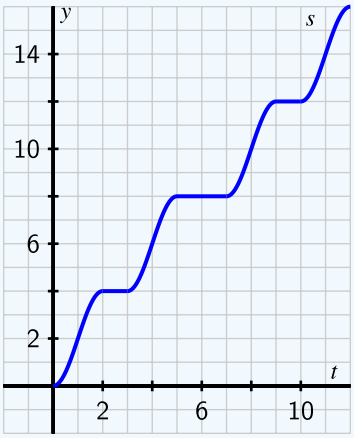


Figure 1.6.2. The graph of $y = s(t)$, the position of the car (measured in thousands of feet from its starting location) at time t in minutes.

1. In everyday language, describe the behavior of the car over the provided time interval. In particular, you should carefully discuss what is happening on each of the time intervals $[0, 1]$, $[1, 2]$, $[2, 3]$, $[3, 4]$, and $[4, 5]$, plus provide commentary overall on what the car is doing on the interval $[0, 12]$.
2. On the lefthand axes provided in Figure 1.6.3, sketch a careful, accurate graph of $y = s'(t)$.
3. What is the meaning of the function $y = s'(t)$ in the context of the given problem? What can we say about the car's behavior when $s'(t)$ is positive? when $s'(t)$ is zero? when $s'(t)$ is negative?
4. Rename the function you graphed in (b) to be called $y = v(t)$. Describe the behavior of v in words, using phrases like "v is increasing on the interval ..." and "v is constant on the interval ... ."
5. Sketch a graph of the function $y = v'(t)$ on the righthand axes provide in Figure 1.6.3. Write at least one sentence to explain how the behavior of $v'(t)$ is connected to the graph of $y = v(t)$.

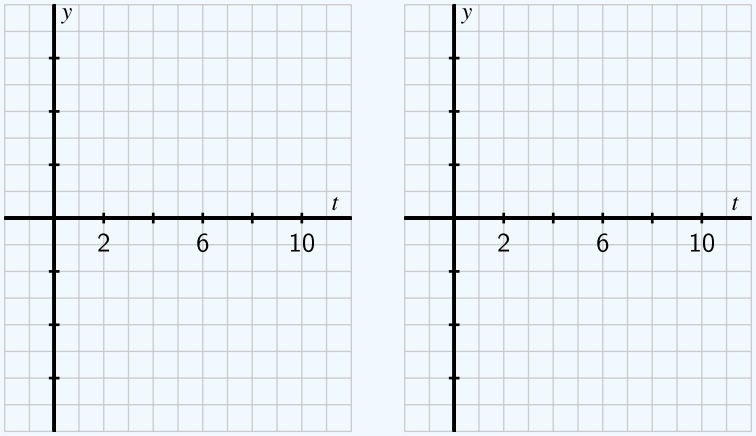


Figure 1.6.3. Axes for plotting $y = v(t) = s'(t)$ and $y = v'(t)$.

Increasing or decreasing

So far, we have used the words *increasing* and *decreasing* intuitively to describe a function's graph. Here we define these terms more formally.

Definition 1.6.1

Given a function $f(x)$ defined on the interval (a, b), we say that f *is increasing on* (a, b) provided that for all x, y in the interval (a, b), if $x < y$, then $f(x) < f(y)$. Similarly, we say that f *is decreasing on* (a, b) provided that for all x, y in the interval (a, b), if $x < y$, then $f(x) > f(y)$.

Simply put, an increasing function is one that is rising as we move from left to right along the graph, and a decreasing function is one that falls as the value of the input increases. If the function has a derivative, the sign of the derivative tells us whether the function is increasing or decreasing.

Let f be a function that is differentiable on an interval (a, b). It is possible to show that that if $f'(x) > 0$ for every x such that $a < x < b$, then f is increasing on (a, b); similarly, if $f'(x) < 0$ on (a, b), then f is decreasing on (a, b).

For example, the function pictured in Figure 1.6.4 is increasing on the entire interval $-2 < x < 0$, and decreasing on the interval $0 < x < 2$. Note that the value $x = 0$ is not included in either interval since at this location, the function is changing from increasing to decreasing.

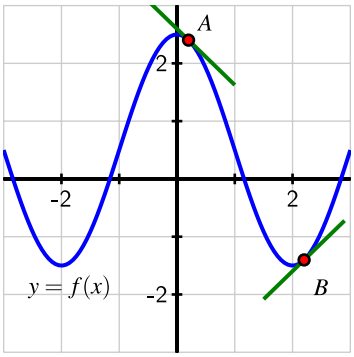


Figure 1.6.4 . A function that is decreasing on the intervals $-3 < x < -2$ and $0 < x < 2$ and increasing on $-2 < x < 0$ and $2 < x < 3$.

The Second Derivative

We are now accustomed to investigating the behavior of a function by examining its derivative. The derivative of a function f is a new function given by the rule

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

Because f' is itself a function, it is perfectly feasible for us to consider the derivative of the derivative, which is the new function $y = [f'(x)]'$. We call this resulting function *the second derivative* of $y = f(x)$, and denote the second derivative by $y = f''(x)$. Consequently, we will sometimes call f' "the first derivative" of f, rather than simply "the derivative" of f.

Definition 1.6.2

The *second derivative* is defined by the limit definition of the derivative of the first derivative. That is,

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

The meaning of the derivative function still holds, so when we compute $y = f''(x)$, this new function measures slopes of tangent lines to the curve $y = f'(x)$, as well as the instantaneous rate of change of $y = f'(x)$. In other words, just as the first derivative measures the rate at which the original function changes, the second derivative measures the rate at which the first derivative changes. The second derivative will help us understand how the rate of change of the original function is itself changing.

Concavity

In addition to asking *whether* a function is increasing or decreasing, it is also natural to inquire *how* a function is increasing or decreasing. There are three basic behaviors that an increasing function can demonstrate on an interval, as pictured in Figure $1.6.5$: the function can increase more and more rapidly, it can increase at the same rate, or it can increase in a way that is slowing down. Fundamentally, we are beginning to think about how a particular curve bends, with the natural comparison being made to lines, which don't bend at all. More than this, we want to understand how the bend in a function's graph is tied to behavior characterized by the first derivative of the function.

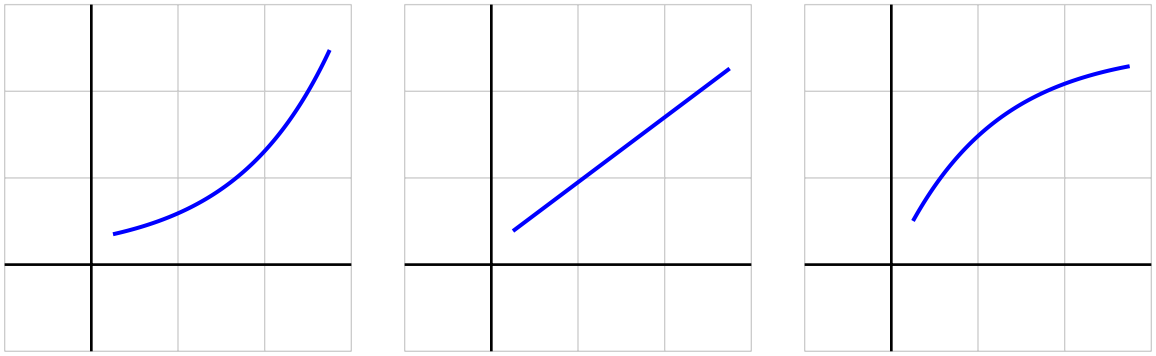


Figure $1.6.5$. Three functions that are all increasing, but doing so at an increasing rate, at a constant rate, and at a decreasing rate, respectively.

On the leftmost curve in Figure $1.6.5$, draw a sequence of tangent lines to the curve. As we move from left to right, the slopes of those tangent lines will increase. Therefore, the rate of change of the pictured function is increasing, and this explains why we say this function is *increasing at an increasing rate*. For the rightmost graph in Figure $1.6.5$, observe that as x increases, the function increases, but the slopes of the tangent lines decrease. This function is *increasing at a decreasing rate*.

Similar options hold for how a function can decrease. Here we must be extra careful with our language, because decreasing functions involve negative slopes. Negative numbers present an interesting tension between common language and mathematical language. For example, it can be tempting to say that "-100 is bigger than -2." But we must remember that "greater than" describes how numbers lie on a number line: $x > y$ provided that x lies to the right of y. So of course, -100 is less than -2. Informally, it might be helpful to say that "-100 is more negative than -2." When a function's values are negative, and those values get more negative as the input increases, the function must be decreasing.

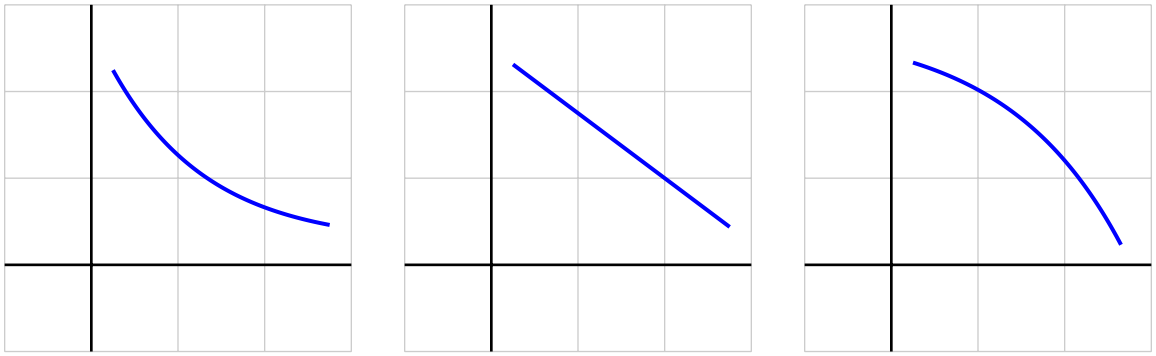


Figure $1.6.6$. From left to right, three functions that are all decreasing, but doing so in different ways.

Now consider the three graphs shown in Figure \({1{.}6{.}}\eatSpaces{16} \endSpaces}\ . Clearly the middle graph depicts a function decreasing at a constant rate. Now, on the first curve, draw a sequence of tangent lines. We see that the slopes of these lines get less and less negative as we move from left to right. That means that the values of the first derivative, while all negative, are increasing, and thus we say that the leftmost curve is *decreasing at an increasing rate*.

This leaves only the rightmost curve in Figure 1.6.6 to consider. For that function, the slopes of the tangent lines are negative throughout the pictured interval, but as we move from left to right, the slopes get more and more negative. Hence the slope of the curve is decreasing, and we say that the function is *decreasing at a decreasing rate*.

We now introduce the notion of *concavity* which provides simpler language to describe these behaviors. When a curve opens upward on a given interval, like the parabola $y = x^2$ or the exponential growth function $y = e^x$, we say that the curve is *concave up* on that interval. Likewise, when a curve opens down, like the parabola $y = -x^2$ or the opposite of the exponential function $y = -e^x$, we say that the function is *concave down*. Concavity is linked to both the first and second derivatives of the function.

In Figure 1.6.7, we see two functions and a sequence of tangent lines to each. On the lefthand plot, where the function is concave up, observe that the tangent lines always lie below the curve itself, and the slopes of the tangent lines are increasing as we move from left to right. In other words, the function f is concave up on the interval shown because its derivative, f', is increasing on that interval. Similarly, on the righthand plot in Figure 1.6.7, where the function shown is concave down, we see that the tangent lines alway lie above the curve, and the slopes of the tangent lines are decreasing as we move from left to right. The fact that its derivative, f', is decreasing makes f concave down on the interval.

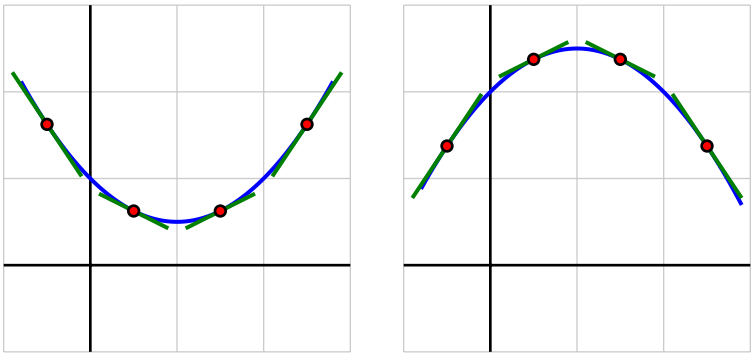


Figure 1.6.7. At left, a function that is concave up; at right, one that is concave down.

We state these most recent observations formally as the definitions of the terms *concave up* and *concave down*.

Definition 1.6.3

Let f be a differentiable function on an interval (a, b). Then f is ***concave up*** on (a, b) if and only if f' is increasing on (a, b); f is ***concave down*** on (a, b) if and only if f' is decreasing on (a, b).

Activity 1.6.2

The position of a car driving along a straight road at time t in minutes is given by the function $y = s(t)$ that is pictured in Figure 1.6.8. The car's position function has units measured in thousands of feet. Remember that you worked with this function and sketched graphs of $y = v(t) = s'(t)$ and $y = v'(t)$ in Preview Activity 1.6.1.

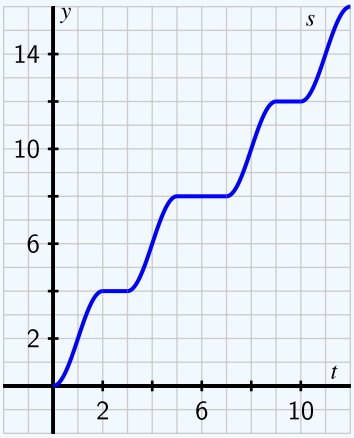


Figure 1.6.8. The graph of $y = s(t)$, the position of the car (measured in thousands of feet from its starting location) at time t in minutes.

1. On what intervals is the position function $y = s(t)$ increasing? decreasing? Why?
2. On which intervals is the velocity function $y = v(t) = s'(t)$ increasing? decreasing? neither? Why?
3. *Acceleration* is defined to be the instantaneous rate of change of velocity, as the acceleration of an object measures the rate at which the velocity of the object is changing. Say that the car's acceleration function is named $a(t)$. How is $a(t)$ computed from $v(t)$? How is $a(t)$ computed from $s(t)$? Explain.
4. What can you say about s'' whenever s' is increasing? Why?
5. Using only the words *increasing*, *decreasing*, *constant*, *concave up*, *concave down*, and *linear*, complete the following sentences. For the position function s with velocity v and acceleration a,
   - on an interval where v is positive, s is .
   - on an interval where v is negative, s is .
   - on an interval where v is zero, s is .
   - on an interval where a is positive, v is .
   - on an interval where a is negative, v is .
   - on an interval where a is zero, v is .
   - on an interval where a is positive, s is .
   - on an interval where a is negative, s is .
   - on an interval where a is zero, s is .

Exploring the context of position, velocity, and acceleration is an excellent way to understand how a function, its first derivative, and its second derivative are related to one another. In Activity 1.6.2 we can replace s, v, and a with an arbitrary function f and its derivatives f' and f'', and essentially all the same observations hold. In particular, note that the following are equivalent: on an interval where the graph of f is concave up, f' is increasing and f'' is positive. Likewise, on an interval where the graph of f is concave down, f' is decreasing and f'' is negative.

Activity 1.6.3

A potato is placed in an oven, and the potato's temperature F (in degrees Fahrenheit) at various points in time is taken and recorded in the following table. Time t is measured in minutes. In Activity 1.6.2, we computed approximations to $F'(30)$ and $F'(60)$ using central differences. Those values and more are provided in the second table below, along with several others computed in the same way.

Table 1.6.1. Select values of $F(t)$.

t	$F(t)$
0	70
15	180.5

30	251
45	296
60	324.5
75	342.8
90	354.5

Table 1.6.2. *Select values of* $F'(t)$.

t	$F'(t)$
0	NA
15	6.03
30	3.85
45	2.45
60	1.56
75	1.00
90	NA

1. What are the units on the values of $F'(t)$?
2. Use a central difference to estimate the value of $F''(30)$.
3. What is the meaning of the value of $F''(30)$ that you have computed in (b) in terms of the potato's temperature? Write several careful sentences that discuss, with appropriate units, the values of $F(30)$, $F'(30)$, and $F''(30)$, and explain the overall behavior of the potato's temperature at this point in time.
4. Overall, is the potato's temperature increasing at an increasing rate, increasing at a constant rate, or increasing at a decreasing rate? Why?

Activity 1.6.4

This activity builds on our experience and understanding of how to sketch the graph of f' given the graph of f.

In Figure 1.6.9, given the respective graphs of two different functions f, sketch the corresponding graph of f' on the first axes below, and then sketch f'' on the second set of axes. In addition, for each, write several careful sentences in the spirit of those in Activity 1.6.2 that connect the behaviors of f, f', and f''. For instance, write something such as

> f' is on the interval , which is connected to the fact that f is on the same interval , and f'' is on the interval.

but of course with the blanks filled in. Throughout, view the scale of the grid for the graph of f as being 1×1, and assume the horizontal scale of the grid for the graph of f' is identical to that for f. If you need to adjust the vertical scale on the axes for the graph of f' or f'', you should label that accordingly.



Figure 1.6.9. Two given functions f, with axes provided for plotting f' and f'' below.

Summary

- A differentiable function f is increasing on an interval whenever its first derivative is positive, and decreasing whenever its first derivative is negative.
- By taking the derivative of the derivative of a function f, we arrive at the second derivative, f''. The second derivative measures the instantaneous rate of change of the first derivative. The sign of the second derivative tells us whether the slope of the tangent line to f is increasing or decreasing.
- A differentiable function is concave up whenever its first derivative is increasing (or equivalently whenever its second derivative is positive), and concave down whenever its first derivative is decreasing (or equivalently whenever its second derivative is negative). Examples of functions that are everywhere concave up are $y = x^2$ and $y = e^x$; examples of functions that are everywhere concave down are $y = -x^2$ and $y = -e^x$.
- The units on the second derivative are "units of output per unit of input per unit of input." They tell us how the value of the derivative function is changing in response to changes in the input. In other words, the second derivative tells us the rate of change of the rate of change of the original function.

This page titled 1.6: The Second Derivative is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.6: The Second Derivative by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

1.7: Limits, Continuity, and Differentiability

Motivating Questions

- What does it mean graphically to say that f has limit L as $x \to a$? How is this connected to having a left-hand limit at $x = a$ and having a right-hand limit at $x = a$?
- What does it mean to say that a function f is continuous at $x = a$? What role do limits play in determining whether or not a function is continuous at a point?
- What does it mean graphically to say that a function f is differentiable at $x = a$? How is this connected to the function being locally linear?
- How are the characteristics of a function having a limit, being continuous, and being differentiable at a given point related to one another?

In Section 1.2, we learned how limits can be used to study the trend of a function near a fixed input value. In this section, we aim to quantify how the function acts and how its values change near a particular point. If the function has a limit L at $x = a$, we will consider how the value of the function $f(a)$ is related to $\lim_{x\to a} f(x)$, and whether or not the function has a derivative $f'(a)$ at $x = a$.

Preview Activity 1.7.1

A function f defined on $-4 < x < 4$ is given by the graph in Figure 1.7.1. Use the graph to answer each of the following questions. Note: to the right of $x = 2$, the graph of f is exhibiting infinite oscillatory behavior similar to the function $\sin(\frac{\pi}{x})$ that we encountered in the key example early in Section 1.2.

1. For each of the values $a = -3, -2, -1, 0, 1, 2, 3$, determine whether or not $\lim_{x\to a} f(x)$ exists. If the function has a limit L at a given point, state the value of the limit using the notation $\lim_{x\to a} f(x) = L$. If the function does not have a limit at a given point, write a sentence to explain why.
2. For each of the values of a from part (a) where f has a limit, determine the value of $f(a)$ at each such point. In addition, for each such a value, does $f(a)$ have the same value as $\lim_{x\to a} f(x)$?
3. For each of the values $a = -3, -2, -1, 0, 1, 2, 3$, determine whether or not $f'(a)$ exists. In particular, based on the given graph, ask yourself if it is reasonable to say that f has a tangent line at $(a, f(a))$ for each of the given a-values. If so, visually estimate the slope of the tangent line to find the value of $f'(a)$.

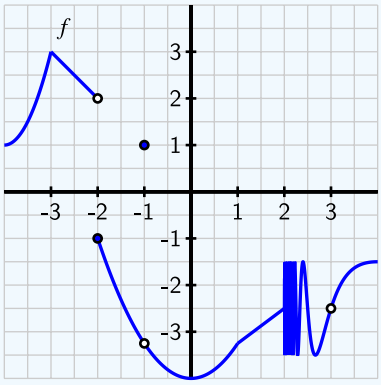


Figure 1.7.1. The graph of $y = f(x)$.

Having a limit at a point

In Section 1.2, we learned that f has limit L as x approaches a provided that we can make the value of $f(x)$ as close to L as we

like by taking x sufficiently close (but not equal to) a. If so, we write $\lim_{x\to a} f(x) = L$.

Essentially there are two behaviors that a function can exhibit near a point where it fails to have a limit. In Figure $1.7.2$, at left we see a function f whose graph shows a jump at $a = 1$. If we let x approach 1 from the left side, the value of f approaches 2, but if we let x approach 1 from the right, the value of f tends to 3. Because the value of f does not approach a single number as x gets arbitrarily close to 1 from both sides, we know that f does not have a limit at $a = 1$.

For such cases, we introduce the notion of *left* and *right* (or *one-sided*) limits.

Definition 1.7.1

We say that f *has limit* L_1 *as* x *approaches* a *from the left* and write

$$\lim_{x\to a^-} f(x) = L_1$$

provided that we can make the value of $f(x)$ as close to L_1 as we like by taking x sufficiently close to a while always having $x < a$. We call L_1 the left-hand limit of f as x approaches a. Similarly, we say L_2 *is the right-hand limit of* f *as* x *approaches* a and write

$$\lim_{x\to a^+} f(x) = L_2$$

provided that we can make the value of $f(x)$ as close to L_2 as we like by taking x sufficiently close to a while always having $x > a$.

In the graph of the function f in Figure $1.7.2$, we see that

$$\lim_{x\to 1^-} f(x) = 2 \quad \text{and} \quad \lim_{x\to 1^+} f(x) = 3.$$

Precisely because the left and right limits are not equal, the overall limit of f as $x \to 1$ fails to exist.

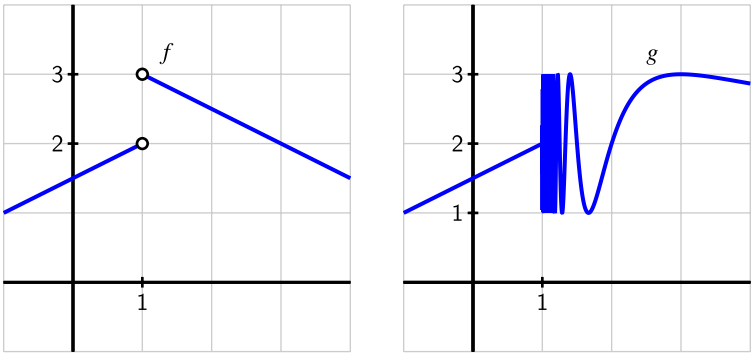


Figure $1.7.2$. Functions f and g that each fail to have a limit at $a = 1$.

For the function g pictured at right in Figure $1.7.2$, the function fails to have a limit at $a = 1$ for a different reason. While the function does not have a jump in its graph at $a = 1$, it is still not the case that g approaches a single value as x approaches 1. In particular, due to the infinitely oscillating behavior of g to the right of $a = 1$, we say that the right-hand limit of g as $x \to 1^+$ does not exist, and thus $\lim_{x\to 1} g(x)$ does not exist.

To summarize, if either a left- or right-hand limit fails to exist or if the left- and right-hand limits are not equal to each other, the overall limit does not exist.

Note

A function f has limit L as $x \to a$ if and only if

$$\lim_{x \to a^-} f(x) = L = \lim_{x \to a^+} f(x).$$

That is, a function has a limit at $x = a$ if and only if both the left- and right-hand limits at $x = a$ exist and have the same value.

In Preview Activity 1.7.1, the function f given in Figure 1.7.1fails to have a limit at only two values: at $a = -2$ (where the left- and right-hand limits are 2 and -1, respectively) and at $x = 2$, where $\lim_{x \to 2^+} f(x)$ does not exist). Note well that even at values such as $a = -1$ and $a = 0$ where there are holes in the graph, the limit still exists.

Activity 1.7.2

Consider a function that is piecewise-defined according to the formula

$$f(x) = \begin{cases} 3(x+2)+2 & \text{for } -3 < x < -2 \\ \frac{2}{3}(x+2)+1 & \text{for } -2 \le x < -1 \\ \frac{2}{3}(x+2)+1 & \text{for } -1 < x < 1 \\ 2 & \text{for } x = 1 \\ 4 - x & \text{for } x > 1 \end{cases}$$

Use the given formula to answer the following questions.

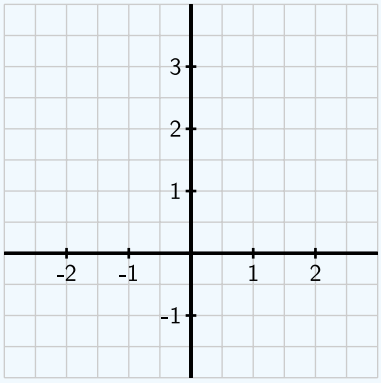


Figure 1.7.3. Axes for plotting the function $y = f(x)$ in Activity 1.7.2.

1. For each of the values $a = -2, -1, 0, 1, 2$, compute $f(a)$.
2. For each of the values $a = -2, -1, 0, 1, 2$, determine $\lim_{x \to a^-} f(x)$ and $\lim_{x \to a^+} f(x)$.
3. For each of the values $a = -2, -1, 0, 1, 2$, determine $\lim_{x \to a} f(x)$. If the limit fails to exist, explain why by discussing the left- and right-hand limits at the relevant a-value.
4. For which values of a is the following statement true?
$$\lim_{x \to a} f(x) \neq f(a)$$
5. On the axes provided in Figure 1.7.3, sketch an accurate, labeled graph of $y = f(x)$. Be sure to carefully use open circles (◦) and filled circles (•) to represent key points on the graph, as dictated by the piecewise formula.

Being continuous at a point

Intuitively, a function is continuous if we can draw its graph without ever lifting our pencil from the page. Alternatively, we might

say that the graph of a continuous function has no jumps or holes in it. In Figure 1.7.4 we consider three functions that have a limit at $a = 1$, and use them to make the idea of continuity more precise.

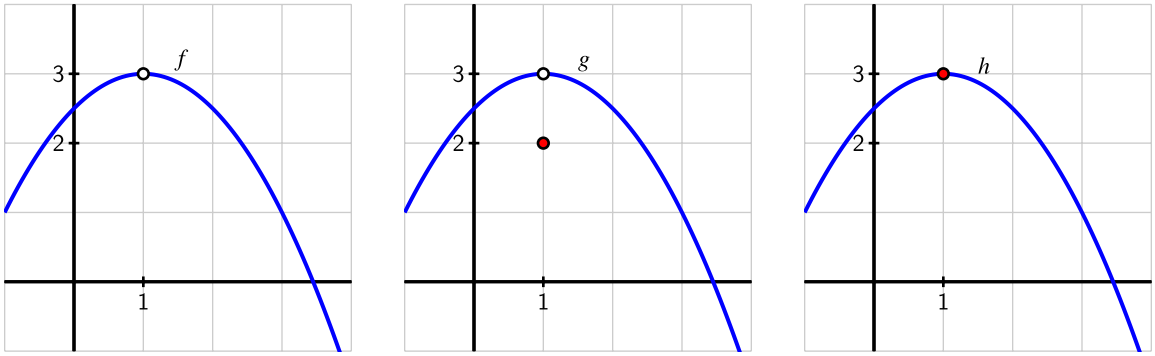


Figure 1.7.4. Functions f, g, and h that demonstrate subtly different behaviors at $a = 1$.

First consider the function in the left-most graph. Note that $f(1)$ is not defined, which leads to the resulting hole in the graph of f at $a = 1$. We will naturally say that f is *not continuous at* $a = 1$. For the function g, we observe that while $\lim_{x\to 1} g(x) = 3$, the value of $g(1) = 2$, and thus the limit does not equal the function value. Here, too, we will say that g is *not continuous*, even though the function is defined at $a = 1$. Finally, the function h appears to be the most well-behaved of all three, since at $a = 1$ its limit and its function value agree. That is,

$$\lim_{x\to 1} h(x) = 3 = h(1).$$

With no hole or jump in the graph of h at $a = 1$, we say that h is *continuous* there. More formally, we make the following definition.

Definition 1.7.2

A function f is *continuous at* $x = a$ provided that

a. f has a limit as $x \to a$,

b. f is defined at $x = a$, and

c. $\lim_{x\to a} f(x) = f(a)$.

Conditions (a) and (b) are technically contained implicitly in (c), but we state them explicitly to emphasize their individual importance. The definition says that a function is continuous at $x = a$ provided that its limit as $x \to a$ exists and equals its function value at $x = a$. If a function is continuous at every point in an interval $[a, b]$, we say the function is "continuous on $[a, b]$." If a function is continuous at every point in its domain, we simply say the function is "continuous." Thus, continuous functions are particularly nice: to evaluate the limit of a continuous function at a point, all we need to do is evaluate the function.

For example, consider $p(x) = x^2 - 2x + 3$. It can be proved that every polynomial is a continuous function at every real number, and thus if we would like to know $\lim_{x\to 2} p(x)$, we simply compute

$$\lim_{x\to 2}(x^2 - 2x + 3) = 2^2 - 2 \cdot 2 + 3 = 3.$$

This route of substituting an input value to evaluate a limit works whenever we know that the function being considered is continuous. Besides polynomial functions, all exponential functions and the sine and cosine functions are continuous at every point, as are many other familiar functions and combinations thereof.

Activity 1.7.3

This activity builds on your work in Preview Activity 1.7.1 , using the same function f as given by the graph that is repeated in Figure 1.7.5.

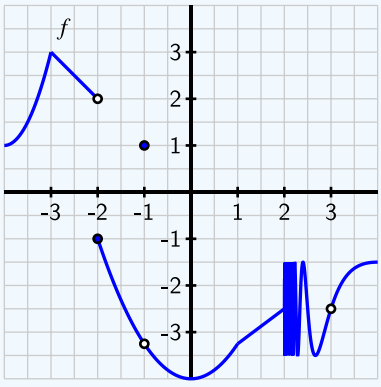


Figure 1.7.5. The graph of $y = f(x)$ for Activity 1.7.3.

1. At which values of a does $\lim_{x \to a} f(x)$ not exist?
2. At which values of a is $f(a)$ not defined?
3. At which values of a does f have a limit, but $\lim_{x \to a} f(x) \neq f(a)$?
4. State all values of a for which f is not continuous at $x = a$.
5. Which condition is stronger, and hence implies the other: f has a limit at $x = a$ or f is continuous at $x = a$? Explain, and hence complete the following sentence: "If f at $x = a$, then f at $x = a$," where you complete the blanks with *has a limit* and *is continuous*, using each phrase once.

Being differentiable at a point

We recall that a function f is said to be differentiable at $x = a$ if $f'(a)$ exists. Moreover, for $f'(a)$ to exist, we know that the function $y = f(x)$ must have a tangent line at the point $(a, f(a))$, since $f'(a)$ is precisely the slope of this line. In order to even ask if f has a tangent line at $(a, f(a))$, it is necessary that f be continuous at $x = a$: if f fails to have a limit at $x = a$, if $f(a)$ is not defined, or if $f(a)$ does not equal the value of $\lim_{x \to a} f(x)$, then it doesn't make sense to talk about a tangent line to the curve at this point.

Indeed, it can be proved formally that if a function f is differentiable at $x = a$, then it must be continuous at $x = a$. So, if f is not continuous at $x = a$, then it is automatically the case that f is not differentiable there. For example, in Figure 1.7.2, both f and g fail to be differentiable at $x = 1$ because neither function is continuous at $x = 1$. But can a function fail to be differentiable at a point where the function is continuous?

In Figure 1.7.6, the function has a sharp corner at a point. For the pictured function f, we observe that f is clearly continuous at $a = 1$, since $\lim_{x \to 1} f(x) = 1 = f(1)$.

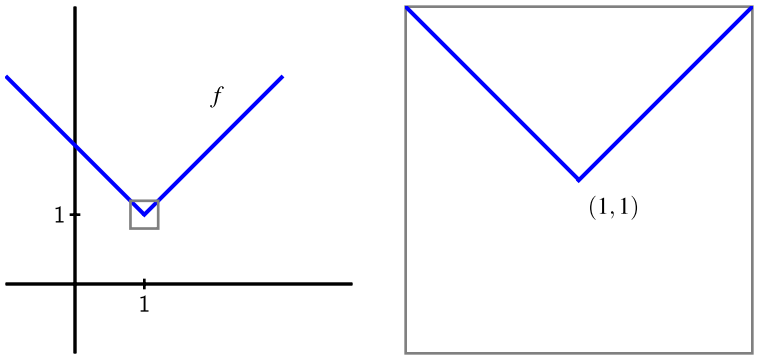


Figure 1.7.6. A function f that is continuous at $a = 1$ but not differentiable at $a = 1$; at right, we zoom in on the point $(1, 1)$ in a magnified version of the box in the left-hand plot.

But the function f in Figure 1.7.6 is not differentiable at $a = 1$ because $f'(1)$ fails to exist. One way to see this is to observe that $f'(x) = -1$ for every value of x that is less than 1, while $f'(x) = +1$ for every value of x that is greater than 1. That makes it seem that either $+1$ or -1 would be equally good candidates for the value of the derivative at $x = 1$. Alternately, we could use the limit definition of the derivative to attempt to compute $f'(1)$, and discover that the derivative does not exist. Finally, we can see visually that the function f in Figure 1.7.6 does not have a tangent line. When we zoom in on $(1, 1)$ on the graph of f, no matter how closely we examine the function, it will always look like a "V", and never like a single line, which tells us there is no possibility for a tangent line there.

If a function does have a tangent line at a given point, when we zoom in on the point of tangency, the function and the tangent line should appear essentially indistinguishable [1]. Conversely, if we zoom in on a point and the function looks like a single straight line, then the function should have a tangent line there, and thus be differentiable. Hence, a function that is differentiable at $x = a$ will, up close, look more and more like its tangent line at $(a, f(a))$. Therefore, we say that a function that is differentiable at $x = a$ is *locally linear*.

See, for instance, http://gvsu.edu/s/6J for an applet (due to David Austin, GVSU) where zooming in shows the increasing similarity between the tangent line and the curve.

To summarize the preceding discussion of differentiability and continuity, we make several important observations.

Note

- If f is differentiable at $x = a$, then f is continuous at $x = a$. Equivalently, if f fails to be continuous at $x = a$, then f will not be differentiable at $x = a$.
- A function can be continuous at a point, but not be differentiable there. In particular, a function f is not differentiable at $x = a$ if the graph has a sharp corner (or *cusp*) at the point $(a, f(a))$.
- If f is differentiable at $x = a$, then f is locally linear at $x = a$. That is, when a function is differentiable, it looks linear when viewed up close because it resembles its tangent line there.

Activity 1.7.4

In this activity, we explore two different functions and classify the points at which each is not differentiable. Let g be the function given by the rule $g(x) = |x|$, and let f be the function that we have previously explored in Preview Activity 1.7.1, whose graph is given again in Figure 1.7.7.

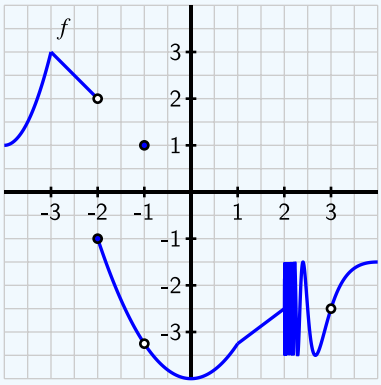


Figure 1.7.7. The graph of $y = f(x)$ for Activity 1.7.4 .

a. Reasoning visually, explain why g is differentiable at every point x such that $x \neq 0$.

b. Use the limit definition of the derivative to show that $g'(0) = \lim_{h \to 0} \frac{|h|}{h}$.

c. Explain why $g'(0)$ fails to exist by using small positive and negative values of h.

d. State all values of a for which f is not differentiable at $x = a$. For each, provide a reason for your conclusion.

e. True or false: if a function p is differentiable at $x = b$, then $\lim_{x \to b} p(x)$ must exist. Why?

Summary

- A function f has limit L as $x \to a$ if and only if f has a left-hand limit at $x = a$, f has a right-hand limit at $x = a$, and the left- and right-hand limits are equal. Visually, this means that there can be a hole in the graph at $x = a$, but the function must approach the same single value from either side of $x = a$.
- A function f is continuous at $x = a$ whenever $f(a)$ is defined, f has a limit as $x \to a$, and the value of the limit and the value of the function agree. This guarantees that there is not a hole or jump in the graph of f at $x = a$.
- A function f is differentiable at $x = a$ whenever $f'(a)$ exists, which means that f has a tangent line at $(a, f(a))$ and thus f is locally linear at $x = a$. Informally, this means that the function looks like a line when viewed up close at $(a, f(a))$ and that there is not a corner point or cusp at $(a, f(a))$.
- Of the three conditions discussed in this section (having a limit at $x = a$, being continuous at $x = a$, and being differentiable at $x = a$), the strongest condition is being differentiable, and the next strongest is being continuous. In particular, if f is differentiable at $x = a$, then f is also continuous at $x = a$, and if f is continuous at $x = a$, then f has a limit at $x = a$.

This page titled 1.7: Limits, Continuity, and Differentiability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.7: Limits, Continuity, and Differentiability by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

1.8: The Tangent Line Approximation

Motivating Questions

- What is the formula for the general tangent line approximation to a differentiable function $y = f(x)$ at the point $(a, f(a))$?
- What is the principle of local linearity and what is the local linearization of a differentiable function f at a point $(a, f(a))$?
- How does knowing just the tangent line approximation tell us information about the behavior of the original function itself near the point of approximation? How does knowing the second derivative's value at this point provide us additional knowledge of the original function's behavior?

Among all functions, linear functions are simplest. One of the powerful consequences of a function $y = f(x)$ being differentiable at a point $(a, f(a))$ is that, up close, the function $y = f(x)$ is locally linear and looks like its tangent line at that point. In certain circumstances, this allows us to approximate the original function f with a simpler function L that is linear: this can be advantageous when we have limited information about f or when f is computationally or algebraically complicated. We will explore all of these situations in what follows.

It is essential to recall that when f is differentiable at $x = a$, the value of $f'(a)$ provides the slope of the tangent line to $y = f(x)$ at the point $(a, f(a))$. If we know both a point on the line and the slope of the line we can find the equation of the tangent line and write the equation in point-slope form [1].

Recall that a line with slope m that passes through (x_0, y_0) has equation $y - y_0 = m(x - x_0)$, and this is the *point-slope form* of the equation.

Preview Activity 1.8.1

Consider the function $y = g(x) = -x^2 + 3x + 2$.

1. Use the limit definition of the derivative to compute a formula for $y = g'(x)$.
2. Determine the slope of the tangent line to $y = g(x)$ at the value $x = 2$.
3. Compute $g(2)$.
4. Find an equation for the tangent line to $y = g(x)$ at the point $(2, g(2))$. Write your result in point-slope form.
5. On the axes provided in Figure 1.8.1, sketch an accurate, labeled graph of $y = g(x)$ along with its tangent line at the point $(2, g(2))$.

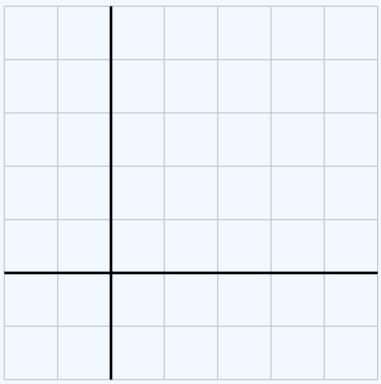


Figure 1.8.1. Axes for plotting $y = g(x)$ and its tangent line to the point $(2, g(2))$.

The tangent line

Given a function f that is differentiable at $x = a$, we know that we can determine the slope of the tangent line to $y = f(x)$ at $(a, f(a))$ by computing $f'(a)$. The equation of the resulting tangent line is given in point-slope form by

$$y - f(a) = f'(a)(x - a) \quad \text{or} \quad y = f'(a)(x - a) + f(a).$$

Note well: there is a major difference between $f(a)$ and $f(x)$ in this context. The former is a constant that results from using the given fixed value of a, while the latter is the general expression for the rule that defines the function. The same is true for $f'(a)$ and $f'(x)$: we must carefully distinguish between these expressions. Each time we find the tangent line, we need to evaluate the function and its derivative at a fixed a-value.

In Figure 1.8.2, we see the graph of a function f and its tangent line at the point $(a, f(a))$. Notice how when we zoom in we see the local linearity of f more clearly highlighted. The function and its tangent line are nearly indistinguishable up close. Local linearity can also be seen dynamically in the java applet at http://gvsu.edu/s/6J.

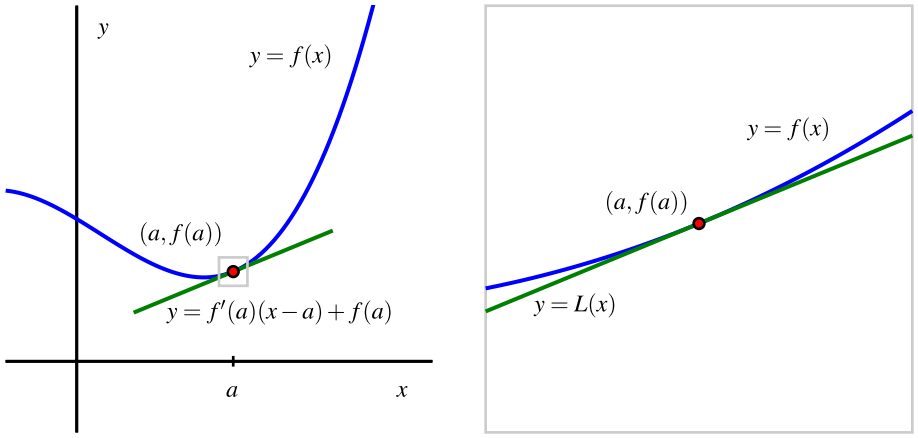


Figure 1.8.2. A function $y = f(x)$ and its tangent line at the point $(a, f(a))$: at left, from a distance, and at right, up close. At right, we label the tangent line function by $y = L(x)$ and observe that for x near a, $f(x) \approx L(x)$.

The local linearization

A slight change in perspective and notation will enable us to be more precise in discussing how the tangent line approximates f near $x = a$. By solving for y, we can write the equation for the tangent line as

$$y = f'(a)(x - a) + f(a)$$

This line is itself a function of x. Replacing the variable y with the expression $L(x)$, we call

$$L(x) = f'(a)(x - a) + f(a)$$

the *local linearization of* f at the point $(a, f(a))$. In this notation, $L(x)$ is nothing more than a new name for the tangent line. As we saw above, for x close to a, $f(x) \approx L(x)$.

Example 1.8.1

Suppose that a function $y = f(x)$ has its tangent line approximation given by $L(x) = 3 - 2(x - 1)$ at the point $(1, 3)$, but we do not know anything else about the function f. To estimate a value of $f(x)$ for x near 1, such as $f(1.2)$, we can use the fact that $f(1.2) \approx L(1.2)$ and hence

$$f(1.2) \approx L(1.2) = 3 - 2(1.2 - 1) = 3 - 2(0.2) = 2.6.$$

We emphasize that $y = L(x)$ is simply a new name for the tangent line function. Using this new notation and our observation that $L(x) \approx f(x)$ for x near a, it follows that we can write

$$f(x) \approx f(a) + f'(a)(x - a) \text{ for } x \text{ near } a.$$

Activity 1.8.2

Suppose it is known that for a given differentiable function $y = g(x)$, its local linearization at the point where $a = -1$ is given by $L(x) = -2 + 3(x + 1)$.

1. Compute the values of $L(-1)$ and $L'(-1)$.
2. What must be the values of $g(-1)$ and $g'(-1)$? Why?
3. Do you expect the value of $g(-1.03)$ to be greater than or less than the value of $g(-1)$? Why?
4. Use the local linearization to estimate the value of $g(-1.03)$.
5. Suppose that you also know that $g''(-1) = 2$. What does this tell you about the graph of $y = g(x)$ at $a = -1$?
6. For x near -1, sketch the graph of the local linearization $y = L(x)$ as well as a possible graph of $y = g(x)$ on the axes provided in Figure 1.8.3.

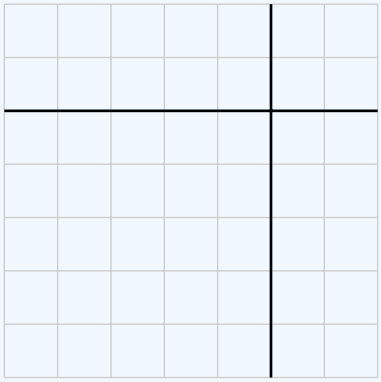


Figure 1.8.3. Axes for plotting $y = L(x)$ and $y = g(x)$.

From Activity 1.8.2, we see that the local linearization $y = L(x)$ is a linear function that shares two important values with the function $y = f(x)$ that it is derived from. In particular,

- because $L(x) = f(a) + f'(a)(x - a)$, it follows that $L(a) = f(a)$; and
- because L is a linear function, its derivative is its slope.

Hence, $L'(x) = f'(a)$ for every value of x, and specifically $L'(a) = f'(a)$. Therefore, we see that L is a linear function that has both the same value and the same slope as the function f at the point $(a, f(a))$.

Thus, if we know the linear approximation $y = L(x)$ for a function, we know the original function's value and its slope at the point of tangency. What remains unknown, however, is the shape of the function f at the point of tangency. There are essentially four possibilities, as shown in Figure 1.8.4.

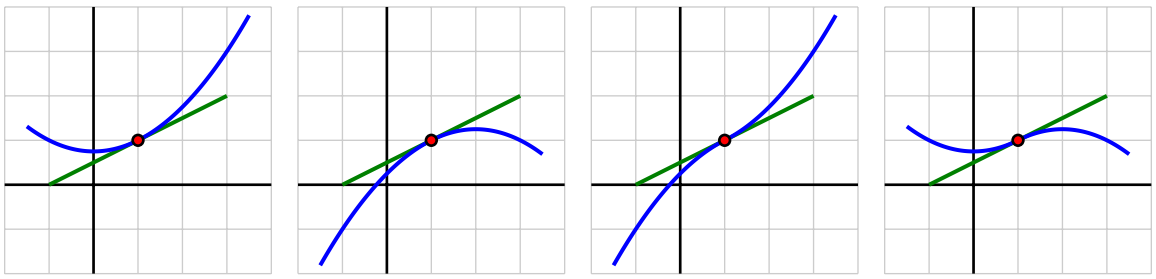


Figure 1.8.4. Four possible graphs for a nonlinear differentiable function and how it can be situated relative to its tangent line at a point.

These possible shapes result from the fact that there are three options for the value of the second derivative: either $f''(a) < 0$, $f''(a) = 0$, or $f''(a) > 0$.

- If $f''(a) > 0$, then we know the graph of f is concave up, and we see the first possibility on the left, where the tangent line lies entirely below the curve.
- If $f''(a) < 0$, then f is concave down and the tangent line lies above the curve, as shown in the second figure.
- If $f''(a) = 0$ and f'' changes sign at $x = a$, the concavity of the graph will change, and we will see either the third or fourth figure.[2].
- A fifth option (which is not very interesting) can occur if the function f itself is linear, so that $f(x) = L(x)$ for all values of x.

It is possible that $f''(a) = 0$ but f'' does not change sign at $x = a$, in which case the graph will look like one of the first two options.

The plots in Figure 1.8.4 highlight yet another important thing that we can learn from the concavity of the graph near the point of tangency: whether the tangent line lies above or below the curve itself. This is key because it tells us whether or not the tangent line approximation's values will be too large or too small in comparison to the true value of f. For instance, in the first situation in the leftmost plot in Figure 1.8.4 where $f''(a) > 0$, because the tangent line falls below the curve, we know that $L(x) \leq f(x)$ for all values of x near a.

Activity 1.8.3

This activity concerns a function $f(x)$ about which the following information is known:

- f is a differentiable function defined at every real number x
- \(\displaystyle f(2) = -1\)
- $y = f'(x)$ has its graph given in Figure 1.8.5

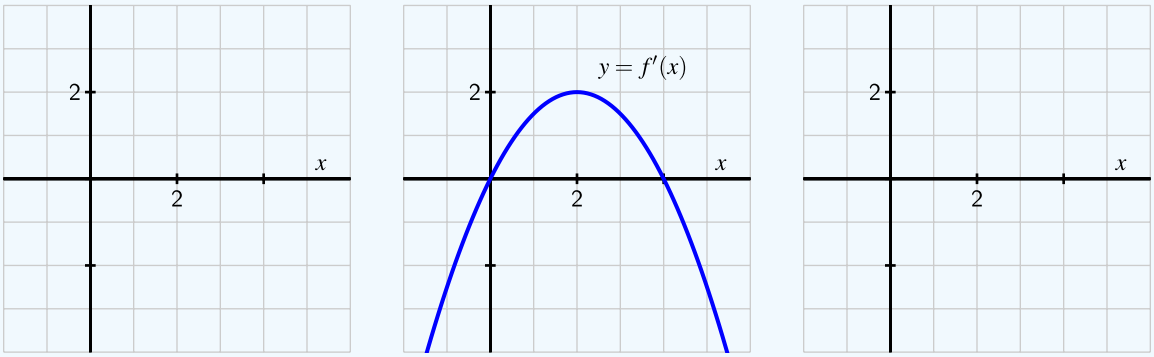


Figure 1.8.5. At center, a graph of $y = f'(x)$; at left, axes for plotting $y = f(x)$; at right, axes for plotting $y = f''(x)$.

Your task is to determine as much information as possible about f (especially near the value $a = 2$) by responding to the questions below.

1. Find a formula for the tangent line approximation, $L(x)$, to f at the point $(2, -1)$.
2. Use the tangent line approximation to estimate the value of $f(2.07)$. Show your work carefully and clearly.
3. Sketch a graph of $y = f''(x)$ on the righthand grid in Figure 1.8.5; label it appropriately.
4. Is the slope of the tangent line to $y = f(x)$ increasing, decreasing, or neither when $x = 2$? Explain.
5. Sketch a possible graph of $y = f(x)$ near $x = 2$ on the lefthand grid in Figure 1.8.5. Include a sketch of $y = L(x)$ (found in part (a)). Explain how you know the graph of $y = f(x)$ looks like you have drawn it.
6. Does your estimate in (b) over- or under-estimate the true value of $f(2.07)$? Why?

The idea that a differentiable function looks linear and can be well-approximated by a linear function is an important one that

finds wide application in calculus. For example, by approximating a function with its local linearization, it is possible to develop an effective algorithm to estimate the zeroes of a function. Local linearity also helps us to make further sense of certain challenging limits. For instance, we have seen that the limit

$$\lim_{x \to 0} \frac{\sin(x)}{x}$$

is indeterminate, because both its numerator and denominator tend to 0. While there is no algebra that we can do to simplify \(\frac{\sin(x)}{x}\text{,}\) it is straightforward to show that the linearization of $f(x) = \sin(x)$ at the point $(0, 0)$ is given by $L(x) = x$. Hence, for values of x near 0, \(\sin(x) \approx x\text{,}\) and therefore

$$\frac{\sin(x)}{x} \approx \frac{x}{x} = 1,$$

which makes plausible the fact that

$$\lim_{x \to 0} \frac{\sin(x)}{x} = 1.$$

Summary

- The tangent line to a differentiable function $y = f(x)$ at the point $(a, f(a))$ is given in point-slope form by the equation
$$y - f(a) = f'(a)(x - a).$$
- The principle of local linearity tells us that if we zoom in on a point where a function $y = f(x)$ is differentiable, the function will be indistinguishable from its tangent line. That is, a differentiable function looks linear when viewed up close. We rename the tangent line to be the function $y = L(x)$, where $L(x) = f(a) + f'(a)(x - a)$. Thus, $f(x) \approx L(x)$ for all x near $x = a$.
- If we know the tangent line approximation $L(x) = f(a) + f'(a)(x - a)$ to a function $y = f(x)$, then because $L(a) = f(a)$ and $L'(a) = f'(a)$, we also know the values of both the function and its derivative at the point where $x = a$. In other words, the linear approximation tells us the height and slope of the original function. If, in addition, we know the value of $f''(a)$, we then know whether the tangent line lies above or below the graph of $y = f(x)$, depending on the concavity of f.

This page titled 1.8: The Tangent Line Approximation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.8: The Tangent Line Approximation by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

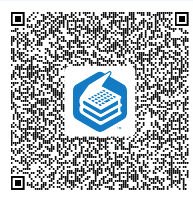
1.9: Understanding the Derivative (Exercises)

1.1: How do we Measure Velocity?

1. Average velocity from position

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

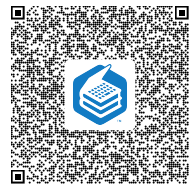
View this Interactive Content online



2. Rate of calorie consumption

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Average rate of change - quadratic function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Comparing average rate of change of two functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

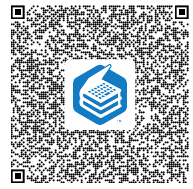
View this Interactive Content online



5. Matching a distance graph to velocity

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



6

A bungee jumper dives from a tower at time $t = 0$. Her height h (measured in feet) at time t (in seconds) is given by the graph in Figure 1.1.4. In this problem, you may base your answers on estimates from the graph or use the fact that the jumper's height function is given by $s(t) = 100\cos(0.75t) \cdot e^{-0.2t} + 100$.

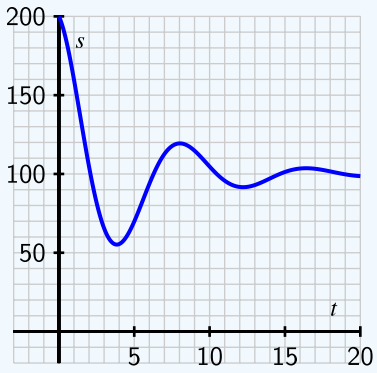


Figure 1.1.4. A bungee jumper's height function.

a. What is the change in vertical position of the bungee jumper between $t = 0$ and $t = 15$?
b. Estimate the jumper's average velocity on each of the following time intervals: $[0, 15]$, $[0, 2]$, $[1, 6]$, and $[8, 10]$. Include units on your answers.
c. On what time interval(s) do you think the bungee jumper achieves her greatest average velocity? Why?
d. Estimate the jumper's instantaneous velocity at $t = 5$. Show your work and explain your reasoning, and include units on your answer.
e. Among the average and instantaneous velocities you computed in earlier questions, which are positive and which are negative? What does negative velocity indicate?

7

A diver leaps from a 3 meter springboard. His feet leave the board at time $t = 0$, he reaches his maximum height of 4.5 m at $t = 1.1$ seconds, and enters the water at $t = 2.45$. Once in the water, the diver coasts to the bottom of the pool (depth 3.5 m), touches bottom at $t = 7$, rests for one second, and then pushes off the bottom. From there he coasts to the surface, and takes his first breath at $t = 13$.

a. a. Let $s(t)$ denote the function that gives the height of the diver's feet (in meters) above the water at time t. (Note that the “height” of the bottom of the pool is -3.5 meters.) Sketch a carefully labeled graph of $s(t)$ on the provided axes in Figure 1.1.5. Include scale and units on the vertical axis. Be as detailed as possible.

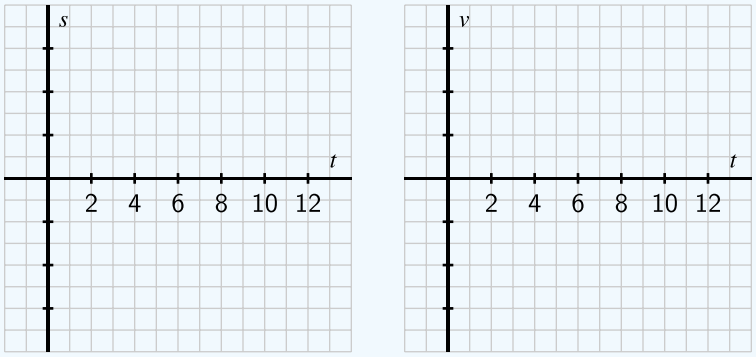


Figure 1.9.1: (left) Axes for plotting $s(t)$ in part (a). (right) Axes for plotting $v(t)$ in part (c).

a. Based on your graph in (a), what is the average velocity of the diver between $t = 2.45$ and $t = 7$? Is his average velocity the same on every time interval within $[2.45, 7]$?
b. Let the function $v(t)$ represent the *instantaneous vertical velocity* of the diver at time t (i.e. the speed at which the height function $s(t)$ is changing; note that velocity in the upward direction is positive, while the velocity of a falling object is negative). Based on your understanding of the diver's behavior, as well as your graph of the position function, sketch a carefully labeled graph of $v(t)$ on the axes provided in Figure 1.1.6. Include scale and units on the vertical axis. Write several sentences that explain how you constructed your graph, discussing when you expect $v(t)$ to be zero, positive, negative, relatively large, and relatively small.
c. Is there a connection between the two graphs that you can describe? What can you say about the velocity graph when the height function is increasing? decreasing? Make as many observations as you can.

8

According to the U.S. census, the population of the city of Grand Rapids, MI, was 181,843 in 1980; 189,126 in 1990; and 197,800 in 2000.

a. Between 1980 and 2000, by how many people did the population of Grand Rapids grow?
b. In an average year between 1980 and 2000, by how many people did the population of Grand Rapids grow?
c. Just like we can find the average velocity of a moving body by computing change in position over change in time, we can compute the average rate of change of any function f. In particular, the *average rate of change* of a function f over an interval $[a, b]$ is the quotient

$$\frac{f(b) - f(a)}{b - a}.$$

What does the quantity $\frac{f(b)-f(a)}{b-a}$ measure on the graph of $y = f(x)$ over the interval $[a, b]$?
d. Let $P(t)$ represent the population of Grand Rapids at time t, where t is measured in years from January 1, 1980. What is the average rate of change of P on the interval $t = 0$ to $t = 20$? What are the units on this quantity?
e. If we assume the population of Grand Rapids is growing at a rate of approximately 4% per decade, we can model the population function with the formula

$$P(t) = 181843(1.04)^{t/10}.$$

Use this formula to compute the average rate of change of the population on the intervals $[5, 10]$, $[5, 9]$, $[5, 8]$, $[5, 7]$, and $[5, 6]$.
f. How fast do you think the population of Grand Rapids was changing on January 1, 1985? Said differently, at what rate do you think people were being added to the population of Grand Rapids as of January 1, 1985? How many additional people should the city have expected in the following year? Why?

1.2: The Notion of Limit

Exercise 1: Limits on a piecewise graph

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Exercise 2: Estimating a limit numerically

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Exercise 3: Limits for a piecewise formula

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Exercise 4: Evaluating a limit algebraically

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Exercise 5

Consider the function whose formula is $f(x) = \frac{16-x^4}{x^2-4}$.

1. What is the domain of f?
2. Use a sequence of values of x near $a = 2$ to estimate the value of $\lim_{x\to 2} f(x)$, if you think the limit exists. If you think the limit doesn't exist, explain why.
3. Use algebra to simplify the expression $\frac{16-x^4}{x^2-4}$ and hence work to evaluate $\lim_{x\to 2} f(x)$ exactly, if it exists, or to explain how your work shows the limit fails to exist. Discuss how your findings compare to your results in (b).
4. True or false: $f(2) = -8$. Why?
5. True or false: $\frac{16-x^4}{x^2-4} = -4 - x^2$. Why? How is this equality connected to your work above with the function f?
6. Based on all of your work above, construct an accurate, labeled graph of $y = f(x)$ on the interval $[1, 3]$, and write a sentence that explains what you now know about $\lim_{x\to 2} \frac{16-x^4}{x^2-4}$.

Exercise 6

Let $g(x) = -\frac{|x+3|}{x+3}$.

1. What is the domain of g?
2. Use a sequence of values near $a = -3$ to estimate the value of $\lim_{x\to -3} g(x)$, if you think the limit exists. If you think the limit doesn't exist, explain why.
3. Use algebra to simplify the expression $\frac{|x+3|}{x+3}$ and hence work to evaluate $\lim_{x\to -3} g(x)$ exactly, if it exists, or to explain how your work shows the limit fails to exist. Discuss how your findings compare to your results in (b). (Hint: $|a| = a$ whenever $a \geq 0$, but $|a| = -a$ whenever $a < 0$.)
4. True or false: $g(-3) = -1$. Why?
5. True or false: $-\frac{|x+3|}{x+3} = -1$. Why? How is this equality connected to your work above with the function g?
6. Based on all of your work above, construct an accurate, labeled graph of $y = g(x)$ on the interval $[-4, -2]$, and write a sentence that explains what you now know about $\lim_{x\to -3} g(x)$.

Exercise 7

For each of the following prompts, sketch a graph on the provided axes of a function that has the stated properties.

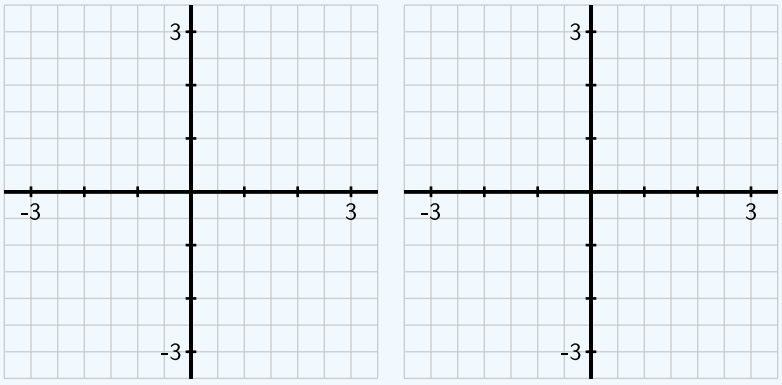


Figure 1.2.12. Axes for plotting $y = f(x)$ in (a) and $y = g(x)$ in (b).

1. $y = f(x)$ such that
   - $f(-2) = 2$ and $\lim_{x\to -2} f(x) = 1$
   - $f(-1) = 3$ and $\lim_{x\to -1} f(x) = 3$
   - $f(1)$ is not defined and $\lim_{x\to 1} f(x) = 0$
   - $f(2) = 1$ and $\lim_{x\to 2} f(x)$ does not exist.
2. $y = g(x)$ such that
   - $g(-2) = 3$, $g(-1) = -1$, $g(1) = -2$, and $g(2) = 3$
   - At $x = -2, -1, 1$ and 2, g has a limit, and its limit equals the value of the function at that point.
   - $g(0)$ is not defined and $\lim_{x\to 0} g(x)$ does not exist.

Exercise 8

A bungee jumper dives from a tower at time $t = 0$. Her height s in feet at time t in seconds is given by $s(t) = 100\cos(0.75t) \cdot e^{-0.2t} + 100$.

1. Write an expression for the average velocity of the bungee jumper on the interval $[1, 1 + h]$.
2. Use computing technology to estimate the value of the limit as $h \to 0$ of the quantity you found in (a).
3. What is the meaning of the value of the limit in (b)? What are its units?

1.3: The Derivative of a Function at a Point

1. Estimating derivative values graphically

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Tangent line to a curve

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Interpreting values and slopes from a graph

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Finding an exact derivative value algebraically

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Estimating a derivative from the limit definition

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6.

Consider the graph of $y = f(x)$ provided in Figure 1.3.12.

1. On the graph of $y = f(x)$, sketch and label the following quantities:
   - the secant line to $y = f(x)$ on the interval $[-3, -1]$ and the secant line to $y = f(x)$ on the interval $[0, 2]$.
   - the tangent line to $y = f(x)$ at $x = -3$ and the tangent line to $y = f(x)$ at $x = 0$.
2. What is the approximate value of the average rate of change of f on $[-3, -1]$? On $[0, 2]$? How are these values related to your work in (a)?
3. What is the approximate value of the instantaneous rate of change of f at $x = -3$? At $x = 0$? How are these values related to your work in (a)?

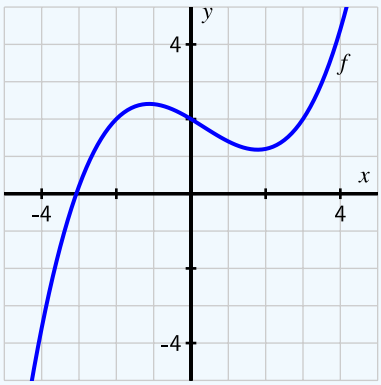


Figure 1.3.12. Plot of $y = f(x)$.

7.

For each of the following prompts, sketch a graph on the provided axes in Figure 1.3.13 of a function that has the stated properties.

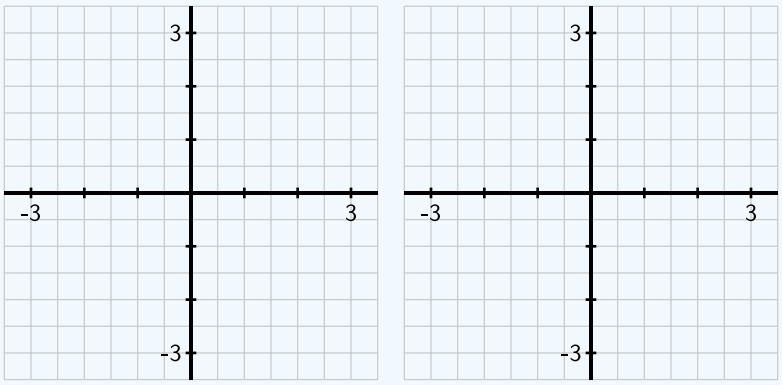


Figure 1.3.13. Axes for plotting $y = f(x)$ in (a) and $y = g(x)$ in (b).

1. $y = f(x)$ such that
   - the average rate of change of f on $[-3, 0]$ is -2 and the average rate of change of f on $[1, 3]$ is 0.5, and
   - the instantaneous rate of change of f at $x = -1$ is -1 and the instantaneous rate of change of f at $x = 2$ is 1.
2. $y = g(x)$ such that
   - $\frac{g(3)-g(-2)}{5} = 0$ and $\frac{g(1)-g(-1)}{2} = -1$, and
   - $g'(2) = 1$ and $g'(-1) = 0$

8.

Suppose that the population, P, of China (in billions) can be approximated by the function $P(t) = 1.15(1.014)^t$ where t is the number of years since the start of 1993.

1. According to the model, what was the total change in the population of China between January 1, 1993 and January 1, 2000? What will be the average rate of change of the population over this time period? Is this average rate of change greater or less than the instantaneous rate of change of the population on January 1, 2000? Explain and justify, being sure to include proper units on all your answers.
2. According to the model, what is the average rate of change of the population of China in the ten-year period starting on January 1, 2012?
3. Write an expression involving limits that, if evaluated, would give the exact instantaneous rate of change of the population on today's date. Then estimate the value of this limit (discuss how you chose to do so) and explain the meaning (including units) of the value you have found.
4. Find an equation for the tangent line to the function $y = P(t)$ at the point where the t-value is given by today's date.

9.

The goal of this problem is to compute the value of the derivative at a point for several different functions, where for each one we do so in three different ways, and then to compare the results to see that each produces the same value.

For each of the following functions, use the limit definition of the derivative to compute the value of $f'(a)$ using three different approaches: strive to use the algebraic approach first (to compute the limit exactly), then test your result using numerical evidence (with small values of h), and finally plot the graph of $y = f(x)$ near $(a, f(a))$ along with the appropriate tangent line to estimate the value of $f'(a)$ visually. Compare your findings among all three approaches; if you are unable to complete the algebraic approach, still work numerically and graphically.

1. $f(x) = x^2 - 3x$, $a = 2$
2. $f(x) = \frac{1}{x}$, $a = 1$
3. $f(x) = \sqrt{x}$, $a = 1$
4. $f(x) = 2 - |x - 1|$, $a = 1$
5. $f(x) = \sin(x)$, $a = \frac{\pi}{2}$

1.4: The Derivative Function

1. The derivative function graphically

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Applying the limit definition of the derivative

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Sketching the derivative

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Comparing function and derivative values

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Limit definition of the derivative for a rational function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6

Let f be a function with the following properties: f is differentiable at every value of x (that is, f has a derivative at every point), $f(-2) = 1$, and $f'(-2) = -2$, $f'(-1) = -1$, $f'(0) = 0$, $f'(1) = 1$, and $f'(2) = 2$.

a. On the axes provided at left in Figure 1.4.3, sketch a possible graph of $y = f(x)$. Explain why your graph meets the stated criteria.

b. Conjecture a formula for the function $y = f(x)$. Use the limit definition of the derivative to determine the corresponding formula for $y = f'(x)$. Discuss both graphical and algebraic evidence for whether or not your conjecture is correct.

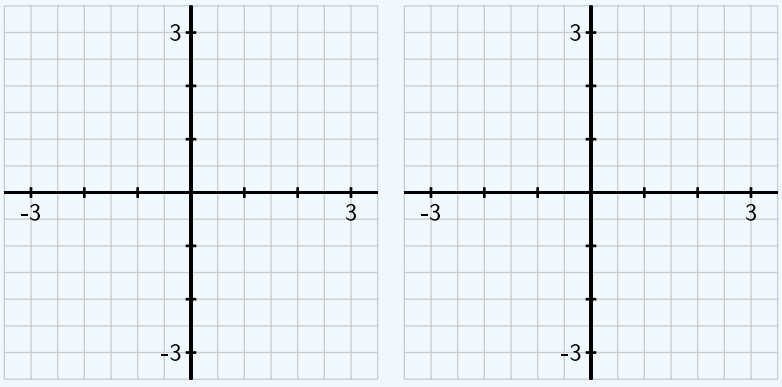


Figure 1.4.3. Axes for plotting $y = f(x)$ in (a) and $y = f'(x)$ in (b).

7.

Consider the function $g(x) = x^2 - x + 3$.

1. Use the limit definition of the derivative to determine a formula for $g'(x)$.
2. Use a graphing utility to plot both $y = g(x)$ and your result for $y = g'(x)$; does your formula for $g'(x)$ generate the graph you expected?
3. Use the limit definition of the derivative to find a formula for $p'(x)$ where $p(x) = 5x^2 - 4x + 12$.
4. Compare and contrast the formulas for $g'(x)$ and $p'(x)$ you have found. How do the constants 5, 4, 12, and 3 affect the results?

8.

Let g be a continuous function (that is, one with no jumps or holes in the graph) and suppose that a graph of $y = g'(x)$ is given by the graph on the right in Figure 1.4.4.

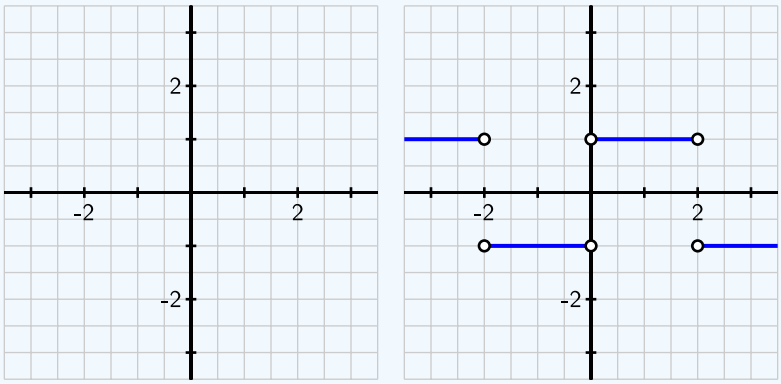


Figure 1.4.4. Axes for plotting $y = g(x)$ and, at right, the graph of $y = g'(x)$.

1. Observe that for every value of x that satisfies $0 < x < 2$, the value of $g'(x)$ is constant. What does this tell you about the behavior of the graph of $y = g(x)$ on this interval?
2. On what intervals other than $0 < x < 2$ do you expect $y = g(x)$ to be a linear function? Why?
3. At which values of x is $g'(x)$ not defined? What behavior does this lead you to expect to see in the graph of $y = g(x)$?
4. Suppose that $g(0) = 1$. On the axes provided at left in Figure 1.4.4, sketch an accurate graph of $y = g(x)$.

9

For each graph that provides an original function $y = f(x)$ in Figure 1.4.5, your task is to sketch an approximate graph of its derivative function, $y = f'(x)$, on the axes immediately below. View the scale of the grid for the graph of f as being 1×1, and assume the horizontal scale of the grid for the graph of f' is identical to that for f. If you need to adjust the vertical scale on the axes for the graph of f', you should label that accordingly.

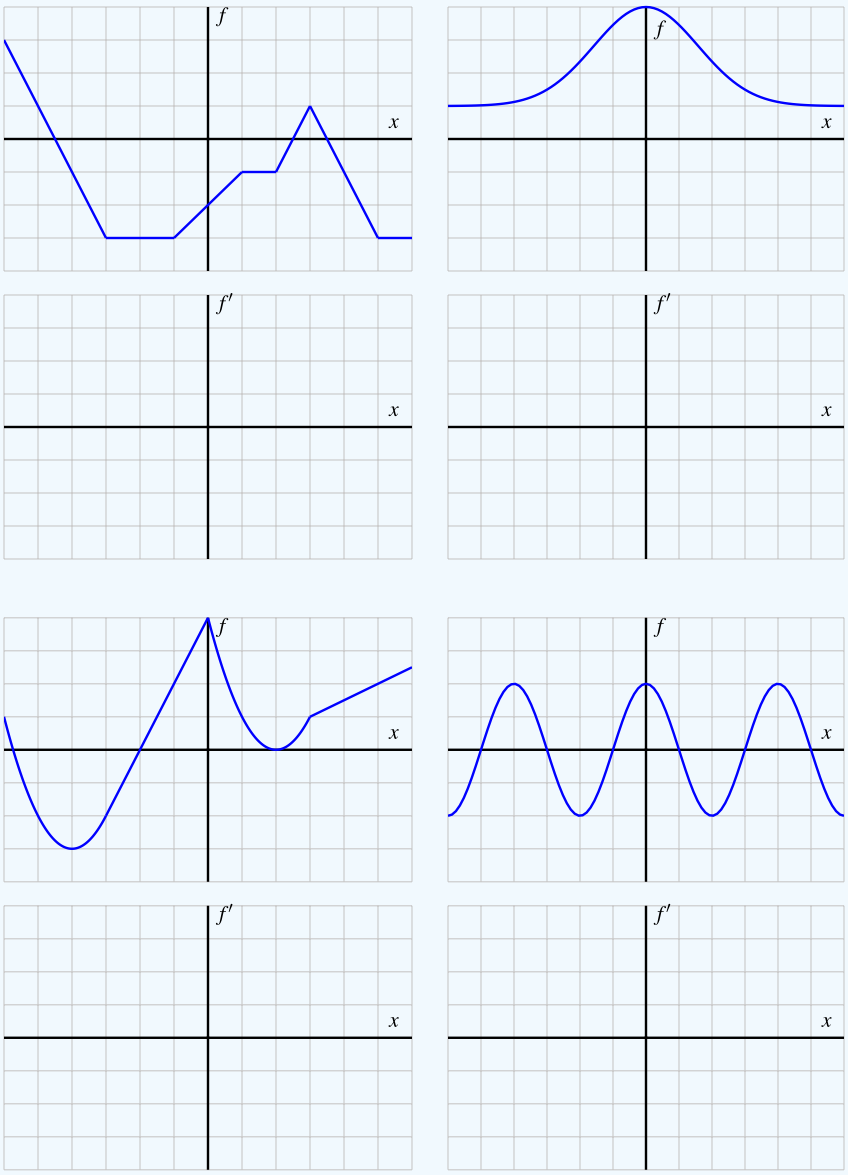


Figure 1.4.5. Graphs of $y = f(x)$ and grids for plotting the corresponding graph of $y = f'(x)$.

1.5: Interpretating, Estimating, and Using the Derivative

1. A cooling cup of coffee

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. A cost function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Weight as a function of calories

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Displacement and velocity

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5

A cup of coffee has its temperature F (in degrees Fahrenheit) at time t given by the function $F(t) = 75 + 110e^{-0.05t}$, where time is measured in minutes.

1. Use a central difference with $h = 0.01$ to estimate the value of $F'(10)$.
2. What are the units on the value of $F'(10)$ that you computed in (a)? What is the practical meaning of the value of $F'(10)$?
3. Which do you expect to be greater: $F'(10)$ or $F'(20)$? Why?
4. Write a sentence that describes the behavior of the function $y = F'(t)$ on the time interval $0 \leq t \leq 30$. How do you think its graph will look? Why?

6.

The temperature change T (in Fahrenheit degrees), in a patient, that is generated by a dose q (in milliliters), of a drug, is given by the function $T = f(q)$.

1. What does it mean to say $f(50) = 0.75$? Write a complete sentence to explain, using correct units.
2. A person's sensitivity, s, to the drug is defined by the function $s(q) = f'(q)$. What are the units of sensitivity?
3. Suppose that $f'(50) = -0.02$. Write a complete sentence to explain the meaning of this value. Include in your response the information given in (a).

7.

The velocity of a ball that has been tossed vertically in the air is given by $v(t) = 16 - 32t$, where v is measured in feet per second, and t is measured in seconds. The ball is in the air from $t = 0$ until $t = 2$.

1. When is the ball's velocity greatest?
2. Determine the value of $v'(1)$. Justify your thinking.
3. What are the units on the value of $v'(1)$? What does this value and the corresponding units tell you about the behavior of the ball at time $t = 1$?
4. What is the physical meaning of the function $v'(t)$?

8.

The value, V, of a particular automobile (in dollars) depends on the number of miles, m, the car has been driven, according to the function $V = h(m)$.

1. Suppose that $h(40000) = 15500$ and $h(55000) = 13200$. What is the average rate of change of h on the interval $[40000, 55000]$, and what are the units on this value?
2. In addition to the information given in (a), say that $h(70000) = 11100$. Determine the best possible estimate of $h'(55000)$ and write one sentence to explain the meaning of your result, including units on your answer.
3. Which value do you expect to be greater: $h'(30000)$ or $h'(80000)$? Why?
4. Write a sentence to describe the long-term behavior of the function $V = h(m)$, plus another sentence to describe the long-term behavior of $h'(m)$. Provide your discussion in practical terms regarding the value of the car and the rate at which that value is changing.

1.6: The Second Derivative

1. Comparing f, f', f'' values

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Signs of f, f', f'' values

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Acceleration from velocity

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Rates of change of stock values

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Interpreting a graph of f'

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6

Suppose that $y = f(x)$ is a twice-differentiable function such that f'' is continuous for which the following information is known: $f(2) = -3$, $f'(2) = 1.5$, $f''(2) = -0.25$.

1. Is f increasing or decreasing near $x = 2$? Is f concave up or concave down near $x = 2$?
2. Do you expect $f(2.1)$ to be greater than -3, equal to -3, or less than -3? Why?
3. Do you expect $f'(2.1)$ to be greater than 1.5, equal to 1.5, or less than 1.5? Why?
4. Sketch a graph of $y = f(x)$ near $(2, f(2))$ and include a graph of the tangent line.

7.

For a certain function $y = g(x)$, its derivative is given by the function pictured in Figure 1.6.15.

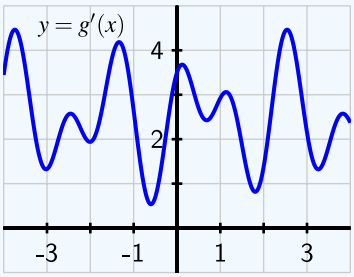


Figure 1.6.15. The graph of $y = g'(x)$.

1. What is the approximate slope of the tangent line to $y = g(x)$ at the point $(2, g(2))$?
2. How many real number solutions can there be to the equation $g(x) = 0$? Justify your conclusion fully and carefully by explaining what you know about how the graph of g must behave based on the given graph of g'.
3. On the interval $-3 < x < 3$, how many times does the concavity of g change? Why?
4. Use the provided graph to estimate the value of $g''(2)$.

8

A bungee jumper's height h (in feet) at time t (in seconds) is given in part by the table:

t	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$h(t)$	200	184.2	159.9	131.9	104.7	81.8	65.5	56.8	55.5	60.4	69.8

t	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0
$h(t)$	81.6	93.7	104.4	112.6	117.7	119.4	118.2	114.8	110.0	104.7

1. Use the given data to estimate $h'(4.5)$, $h'(5)$, and $h'(5.5)$. At which of these times is the bungee jumper rising most rapidly?
2. Use the given data and your work in (a) to estimate $h''(5)$.
3. What physical property of the bungee jumper does the value of $h''(5)$ measure? What are its units?
4. Based on the data, on what approximate time intervals is the function $y = h(t)$ concave down? What is happening to the velocity of the bungee jumper on these time intervals?

9

For each prompt that follows, sketch a possible graph of a function on the interval $-3 < x < 3$ that satisfies the stated properties.

1. $y = f(x)$ such that f is increasing on $-3 < x < 3$, concave up on $-3 < x < 0$, and concave down on $0 < x < 3$.
2. $y = g(x)$ such that g is increasing on $-3 < x < 3$, concave down on $-3 < x < 0$, and concave up on $0 < x < 3$.
3. $y = h(x)$ such that h is decreasing on $-3 < x < 3$, concave up on $-3 < x < -1$, neither concave up nor concave down on $-1 < x < 1$, and concave down on $1 < x < 3$.
4. $y = p(x)$ such that p is decreasing and concave down on $-3 < x < 0$ and is increasing and concave down on $0 < x < 3$.

1.7: Limits, Continuity, and Differentiability

1. Limit values of a piecewise graph

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Limit values of a piecewise formula

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Continuity and differentiability of a graph

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Continuity of a piecewise formula

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5.

Consider the graph of the function $y = p(x)$ that is provided in Figure 1.7.10. Assume that each portion of the graph of p is a straight line, as pictured.

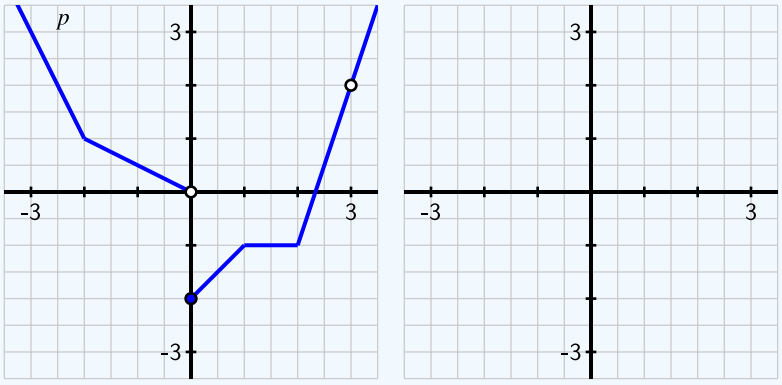


Figure 1.7.10. At left, the piecewise linear function $y = p(x)$. At right, axes for plotting $y = p'(x)$.

1. State all values of a for which $\lim_{x \to a} p(x)$ does not exist.
2. State all values of a for which p is not continuous at a.
3. State all values of a for which p is not differentiable at $x = a$.
4. On the axes provided in Figure 1.7.10, sketch an accurate graph of $y = p'(x)$.

6.

For each of the following prompts, give an example of a function that satisfies the stated criteria. A formula or a graph, with reasoning, is sufficient for each. If no such example is possible, explain why.

1. A function f that is continuous at $a = 2$ but not differentiable at $a = 2$.
2. A function g that is differentiable at $a = 3$ but does not have a limit at $a = 3$.
3. A function h that has a limit at $a = -2$, is defined at $a = -2$, but is not continuous at $a = -2$.
4. A function p that satisfies all of the following:
   - $p(-1) = 3$ and $\lim_{x\to -1} p(x) = 2$
   - $p(0) = 1$ and $p'(0) = 0$
   - $\lim_{x\to 1} p(x) = p(1)$ and $p'(1)$ does not exist

7.

Let $h(x)$ be a function whose derivative $y = h'(x)$ is given by the graph on the right in Figure 1.7.11.

1. Based on the graph of $y = h'(x)$, what can you say about the behavior of the function $y = h(x)$?
2. At which values of x is $y = h'(x)$ not defined? What behavior does this lead you to expect to see in the graph of $y = h(x)$?
3. Is it possible for $y = h(x)$ to have points where h is not continuous? Explain your answer.
4. On the axes provided at left, sketch at least two distinct graphs that are possible functions $y = h(x)$ that each have a derivative $y = h'(x)$ that matches the provided graph at right. Explain why there are multiple possibilities for $y = h(x)$.

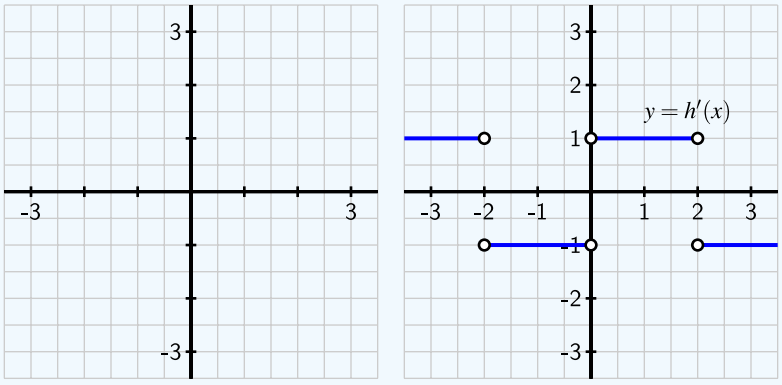


Figure 1.7.11. Axes for plotting $y = h(x)$ and, at right, the graph of $y = h'(x)$.

8.

Consider the function $g(x) = \sqrt{|x|}$.

1. Use a graph to explain visually why g is not differentiable at $x = 0$.
2. Use the limit definition of the derivative to show that
$$g'(0) = \lim_{h\to 0} \frac{\sqrt{|h|}}{h}.$$
3. Investigate the value of $g'(0)$ by estimating the limit in (b) using small positive and negative values of h. For instance, you might compute $\frac{\sqrt{|-0.01|}}{0.01}$. Be sure to use several different values of h (both positive and negative), including ones closer to 0 than 0.01. What do your results tell you about $g'(0)$?
4. Use your graph in (a) to sketch an approximate graph of $y = g'(x)$.

1.8: The Tangent Line Approximation

1. Approximating $\sqrt{x}$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Local linearization of a graph

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Estimating with the local linearization

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Predicting behavior from the local linearization

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5

A certain function $y = p(x)$ has its local linearization at $a = 3$ given by $L(x) = -2x + 5$.

1. What are the values of $p(3)$ and $p'(3)$? Why?
2. Estimate the value of $p(2.79)$.
3. Suppose that $p''(3) = 0$ and you know that $p''(x) < 0$ for $x < 3$. Is your estimate in (b) too large or too small?
4. Suppose that $p''(x) > 0$ for $x > 3$. Use this fact and the additional information above to sketch an accurate graph of $y = p(x)$ near $x = 3$. Include a sketch of $y = L(x)$ in your work.

6

A potato is placed in an oven, and the potato's temperature F (in degrees Fahrenheit) at various points in time is taken and recorded in the following table. Time t is measured in minutes.

Table 1.8.7. Temperature data for the potato.

t	$F(t)$
0	70
15	180.5
30	251
45	296
60	324.5
75	342.8
90	354.5

1. Use a central difference to estimate $F'(60)$. Use this estimate as needed in subsequent questions.
2. Find the local linearization $y = L(t)$ to the function $y = F(t)$ at the point where $a = 60$.
3. Determine an estimate for $F(63)$ by employing the local linearization.
4. Do you think your estimate in (c) is too large or too small? Why?

7

An object moving along a straight line path has a differentiable position function $y = s(t)$; $s(t)$ measures the object's position relative to the origin at time t. It is known that at time $t = 9$ seconds, the object's position is $s(9) = 4$ feet (i.e., 4 feet to the right of the origin). Furthermore, the object's instantaneous velocity at $t = 9$ is -1.2 feet per second, and its acceleration at the same instant is 0.08 feet per second per second.

1. Use local linearity to estimate the position of the object at $t = 9.34$.
2. Is your estimate likely too large or too small? Why?
3. In everyday language, describe the behavior of the moving object at $t = 9$. Is it moving toward the origin or away from it? Is its velocity increasing or decreasing?

8

For a certain function f, its derivative is known to be $f'(x) = (x - 1)e^{-x^2}$. Note that you do not know a formula for $y = f(x)$.

1. At what x-value(s) is $f'(x) = 0$? Justify your answer algebraically, but include a graph of f' to support your conclusion.
2. Reasoning graphically, for what intervals of x-values is $f''(x) > 0$? What does this tell you about the behavior of the original function f? Explain.
3. Assuming that $f(2) = -3$, estimate the value of $f(1.88)$ by finding and using the tangent line approximation to f at $x = 2$. Is your estimate larger or smaller than the true value of $f(1.88)$? Justify your answer.

This page titled 1.9: Understanding the Derivative (Exercises) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.E: Understanding the Derivative (Exercises) has no license indicated.
- 1.1: How do we Measure Velocity? by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 1.2: The Notion of Limit by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 1.3: The Derivative of a Function at a Point by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 1.4: The Derivative Function by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 1.5: Interpretating, Estimating, and Using the Derivative by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 1.6: The Second Derivative by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 1.7: Limits, Continuity, and Differentiability by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 1.8: The Tangent Line Approximation by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

2: COMPUTING DERIVATIVES

Throughout Chapter 2, we will be working to develop shortcut derivative rules that will help us to bypass the limit definition of the derivative in order to quickly determine the formula for $f'(x)$ when we are given a formula for $f(x)$.

2.1: Elementary Derivative Rules

2.2: The Sine and Cosine Function

2.3: The Product and Quotient Rules

2.4: Derivatives of Other Trigonometric Functions

2.5: The Chain Rule

2.6: Derivatives of Inverse Functions

2.7: Derivatives of Functions Given Implicitly

2.8: Using Derivatives to Evaluate Limits

2.9: Computing Derivatives (Exercises)

This page titled 2: Computing Derivatives is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 2: Computing Derivatives by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

2.1: Elementary Derivative Rules

Motivating Questions

- What are alternate notations for the derivative?
- How can we use the algebraic structure of a function $f(x)$ to compute a formula for $f'(x)$?
- What is the derivative of a power function of the form $f(x) = x^n$? What is the derivative of an exponential function of form $f(x) = a^x$?
- If we know the derivative of $y = f(x)$, what is the derivative of $y = kf(x)$, where k is a constant?
- If we know the derivatives of $y = f(x)$ and $y = g(x)$, how do we compute the derivative of $y = f(x) + g(x)$?

In Chapter 1, we developed the concept of the derivative of a function. We now know that the derivative f' of a function f measures the instantaneous rate of change of f with respect to x. The derivative also tells us the slope of the tangent line to $y = f(x)$ at any given value of x. So far, we have focused on interpreting the derivative graphically or, in the context of a physical setting, as a meaningful rate of change. To calculate the value of the derivative at a specific point, we have relied on the limit definition of the derivative,

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

In this chapter, we investigate how the limit definition of the derivative leads to interesting patterns and rules that enable us to find a formula for $f'(x)$ quickly, *without* using the limit definition directly. For example, we would like to apply shortcuts to differentiate a function such as $g(x) = 4x^7 - \sin(x) + 3e^x$

Preview Activity 2.1.1

Functions of the form $f(x) = x^n$, where $n = 1, 2, 3, \ldots$, are often called ***power functions***. The first two questions below revisit work we did earlier in Chapter 1, and the following questions extend those ideas to higher powers of x.

1. Use the limit definition of the derivative to find $f'(x)$ for $f(x) = x^2$.
2. Use the limit definition of the derivative to find $f'(x)$ for $f(x) = x^3$.
3. Use the limit definition of the derivative to find $f'(x)$ for $f(x) = x^4$. (Hint: $(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$. Apply this rule to $(x+h)^4$ within the limit definition.)
4. Based on your work in (a), (b), and (c), what do you conjecture is the derivative of $f(x) = x^5$? Of $f(x) = x^{13}$?
5. Conjecture a formula for the derivative of $f(x) = x^n$ that holds for any positive integer n. That is, given $f(x) = x^n$ where n is a positive integer, what do you think is the formula for $f'(x)$?

Some Key Notation

In addition to our usual f' notation, there are other ways to denote the derivative of a function, as well as the instruction to take the derivative. If we are thinking about the relationship between y and x, we sometimes denote the derivative of y with respect to x by the symbol

$$\frac{dy}{dx}$$

which we read “dee-y dee-x.” For example, if $y = x^2$, we'll write that the derivative is $\frac{dy}{dx} = 2x$. This notation comes from the fact that the derivative is related to the slope of a line, and slope is measured by $\frac{\Delta y}{\Delta x}$. Note that while we read $\frac{\Delta y}{\Delta x}$ as “change in y over change in x,” we view $\frac{dy}{dx}$ as a single symbol, not a quotient of two quantities.

We use a variant of this notation as the instruction to take the derivative. In particular,

$$\frac{d}{dx}[\square]$$

means “take the derivative of the quantity in □ with respect to x.” For example, we may write $\frac{d}{dx}[x^2] = 2x$.

It is important to note that the independent variable can be different from x. If we have $f(z) = z^2$, we then write $f'(z) = 2z$. Similarly, if $y = t^2$, we say $\frac{dy}{dt} = 2t$. And it is also true that $\frac{d}{dq}[q^2] = 2q$. This notation may also be used for second derivatives: $f''(z) = \frac{d}{dz}\left[\frac{df}{dz}\right] = \frac{d^2f}{dz^2}$.

In what follows, we'll build a repertoire of functions for which we can quickly compute the derivative.

Constant, Power, and Exponential Functions

So far, we know the derivative formula for two important classes of functions: constant functions and power functions. If $f(x) = c$ is a constant function, its graph is a horizontal line with slope zero at every point. Thus, $\frac{d}{dx}[c] = 0$. We summarize this with the following rule.

Constant Functions

For any real number c, if $f(x) = c$, then $f'(x) = 0$.

Example 2.1.1

If $f(x) = 7$, then $f'(x) = 0$. Similarly, $\frac{d}{dx}[\sqrt{3}] = 0$.

In your work in Preview Activity 2.1.1 , you conjectured that for any positive integer n, if $f(x) = x^n$, then $f'(x) = nx^{n-1}$. This rule can be formally proved for any positive integer n, and also for any nonzero real number (positive or negative).

Power Functions

For any nonzero real number n, if $f(x) = x^n$, then $f'(x) = nx^{n-1}$.

Example 2.1.2

Using the rule for power functions, we can compute the following derivatives. If $g(z) = z^{-3}$, then $g'(z) = -3z^{-4}$. Similarly, if $h(t) = t^{7/5}$, then $\frac{dh}{dt} = \frac{7}{5}t^{2/5}$, and $\frac{d}{dq}[q^\pi] = \pi q^{\pi-1}$.

It will be instructive to have a derivative formula for one more type of basic function. For now, we simply state this rule without explanation or justification; we will explore why this rule is true in one of the exercises. And we will encounter graphical reasoning for why the rule is plausible in Preview Activity 2.2.1.

Exponential Functions

For any positive real number a, if $f(x) = a^x$, then $f'(x) = a^x \ln(a)$.

Example 2.1.3

If $f(x) = 2^x$, then $f'(x) = 2^x \ln(2)$. Similarly, for $p(t) = 10^t$, $p'(t) = 10^t \ln(10)$. It is especially important to note that when $a = e$, where e is the base of the natural logarithm function, we have that

$$\frac{d}{dx}[e^x] = e^x \ln(e) = e^x$$

since $\ln(e) = 1$. This is an extremely important property of the function e^x: its derivative function is itself!

Note carefully the distinction between power functions and exponential functions: in power functions, the variable is in the base, as in x^2, while in exponential functions, the variable is in the power, as in 2^x. As we can see from the rules, this makes a big difference in the form of the derivative.

Activity 2.1.2

Use the three rules above to determine the derivative of each of the following functions. For each, state your answer using full and proper notation, labeling the derivative with its name. For example, if you are given a function $h(z)$, you should write "$h'(z) =$" or "$\frac{dh}{dz} =$" as part of your response.

1. $f(t) = \pi$
2. $g(z) = 7^z$
3. $h(w) = w^{3/4}$
4. $p(x) = 3^{1/2}$
5. $r(t) = (\sqrt{2})^t$
6. $s(q) = q^{-1}$
7. $m(t) = \dfrac{1}{t^3}$

Constant Multiples and Sums of Functions

Next we will learn how to compute the derivative of a function constructed as an algebraic combination of basic functions. For instance, we'd like to be able to take the derivative of a polynomial function such as

$$p(t) = 3t^5 - 7t^4 + t^2 - 9,$$

which is a sum of constant multiples of powers of t. To that end, we develop two new rules: the Constant Multiple Rule and the Sum Rule.

How is the derivative of $y = kf(x)$ related to the derivative of $y = f(x)$? Recall that when we multiply a function by a constant k, we vertically stretch the graph by a factor of $|k|$ (and reflect the graph across $y = 0$ if $k < 0$). This vertical stretch affects the slope of the graph, so the slope of the function $y = kf(x)$ is k times as steep as the slope of $y = f(x)$. Thus, when we multiply a function by a factor of k, we change the value of its derivative by a factor of k as well.[1],

The Constant Multiple Rule can be formally proved as a consequence of properties of limits, using the limit definition of the derivative.

Constant Multiple Rule

For any real number k, if $f(x)$ is a differentiable function with derivative $f'(x)$, then $\frac{d}{dx}[kf(x)] = kf'(x)$.

In words, this rule says that "the derivative of a constant times a function is the constant times the derivative of the function."

Example 2.1.4

If $g(t) = 3 \cdot 5^t$, we have $g'(t) = 3 \cdot 5^t \ln(5)$. Similarly, $\frac{d}{dz}[5z^{-2}] = 5(-2z^{-3})$.

Next we examine a sum of two functions. If we have $y = f(x)$ and $y = g(x)$, we can compute a new function $y = (f + g)(x)$ by

adding the outputs of the two functions: $(f+g)(x) = f(x) + g(x)$. Not only is the value of the new function the sum of the values of the two known functions, but the slope of the new function is the sum of the slopes of the known functions. Therefore[2], we arrive at the following Sum Rule for derivatives:

Like the Constant Multiple Rule, the Sum Rule can be formally proved as a consequence of properties of limits, using the limit definition of the derivative.

The Sum Rule

If $f(x)$ and $g(x)$ are differentiable functions with derivatives $f'(x)$ and $g'(x)$ respectively, then $\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$.

In words, the Sum Rule tells us that "the derivative of a sum is the sum of the derivatives." It also tells us that a sum of two differentiable functions is also differentiable. Furthermore, because we can view the difference function $y = (f-g)(x) = f(x) - g(x)$ as $y = f(x) + (-1 \cdot g(x))$, the Sum Rule and Constant Multiple Rules together tell us that $\frac{d}{dx}[f(x) + (-1 \cdot g(x))] = f'(x) - g'(x)$, or that "the derivative of a difference is the difference of the derivatives." We can now compute derivatives of sums and differences of elementary functions.

Example 2.1.5

Using the sum rule, $\frac{d}{dw}(2^w + w^2) = 2^w \ln(2) + 2w$. Using both the sum and constant multiple rules, if $h(q) = 3q^6 - 4q^{-3}$, then $h'(q) = 3(6q^5) - 4(-3q^{-4}) = 18q^5 + 12q^{-4}$.

Activity 2.1.3

Use only the rules for constant, power, and exponential functions, together with the Constant Multiple and Sum Rules, to compute the derivative of each function below with respect to the given independent variable. Note well that we do not yet know any rules for how to differentiate the product or quotient of functions. This means that you may have to do some algebra first on the functions below before you can actually use existing rules to compute the desired derivative formula. In each case, label the derivative you calculate with its name using proper notation such as $f'(x)$, $h'(z)$, dr/dt, etc.

1. $f(x) = x^{5/3} - x^4 + 2^x$
2. $g(x) = 14e^x + 3x^5 - x$
3. $h(z) = \sqrt{z} + \dfrac{1}{z^4} + 5^z$
4. $r(t) = \sqrt{53}\ t^7 - \pi e^t + e^4$
5. $s(y) = (y^2+1)(y^2-1)$
6. $q(x) = \dfrac{x^3 - x + 2}{x}$
7. $p(a) = 3a^4 - 2a^3 + 7a^2 - a + 12$

In the same way that we have shortcut rules to help us find derivatives, we introduce some language that is simpler and shorter. Often, rather than say "take the derivative of f," we'll instead say simply "differentiate f." Similarly, if the derivative exists at a point, we say "f is differentiable at that point," or that f can be differentiated.

As we work with the algebraic structure of functions, it is important to develop a big picture view of what we are doing. Here, we make several general observations based on the rules we have so far.

- The derivative of any polynomial function will be another polynomial function, and that the degree of the derivative is one less than the degree of the original function. For instance, if $p(t) = 7t^5 - 4t^3 + 8t$, p is a degree 5 polynomial, and its derivative, $p'(t) = 35t^4 - 12t^2 + 8$, is a degree 4 polynomial.
- The derivative of any exponential function is another exponential function: for example, if $g(z) = 7 \cdot 2^z$, then $g'(z) = 7 \cdot 2^z \ln(2)$, which is also exponential.
- We should not lose sight of the fact that all of the meaning of the derivative that we developed in Chapter 1 still holds. The derivative measures the instantaneous rate of change of the original function, as well as the slope of the tangent line at any selected point on the curve.

Activity 2.1.4

Each of the following questions asks you to use derivatives to answer key questions about functions. Be sure to think carefully about each question and to use proper notation in your responses.

1. Find the slope of the tangent line to $h(z) = \sqrt{z} + \frac{1}{z}$ at the point where $z = 4$.
2. A population of cells is growing in such a way that its total number in millions is given by the function $P(t) = 2(1.37)^t + 32$, where t is measured in days.
   1. Determine the instantaneous rate at which the population is growing on day 4, and include units on your answer.
   2. Is the population growing at an increasing rate or growing at a decreasing rate on day 4? Explain.
3. Find an equation for the tangent line to the curve $p(a) = 3a^4 - 2a^3 + 7a^2 - a + 12$ at the point where $a = -1$.
4. What is the difference between being asked to find the *slope* of the tangent line (asked in (a)) and the *equation* of the tangent line (asked in (c))?

Summary

- Given a differentiable function $y = f(x)$, we can express the derivative of f in several different notations: $f'(x)$, $\frac{df}{dx}$, $\frac{dy}{dx}$, and $\frac{d}{dx}[f(x)]$.
- The limit definition of the derivative leads to patterns among certain families of functions that enable us to compute derivative formulas without resorting directly to the limit definition. For example, if f is a power function of the form $f(x) = x^n$, then $f'(x) = nx^{n-1}$ for any real number n other than 0. This is called the Rule for Power Functions.
- We have stated a rule for derivatives of exponential functions in the same spirit as the rule for power functions: for any positive real number a, if $f(x) = a^x$, then $f'(x) = a^x \ln(a)$.
- If we are given a constant multiple of a function whose derivative we know, or a sum of functions whose derivatives we know, the Constant Multiple and Sum Rules make it straightforward to compute the derivative of the overall function. More formally, if $f(x)$ and $g(x)$ are differentiable with derivatives $f'(x)$ and $g'(x)$ and a and b are constants, then

$$\frac{d}{dx}\left[af(x) + bg(x)\right] = af'(x) + bg'(x).$$

This page titled 2.1: Elementary Derivative Rules is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.1: Elementary Derivative Rules by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

2.2: The Sine and Cosine Function

Motivating Questions

- What is a graphical justification for why $\frac{d}{dx}[a^x] = a^x \ln(a)$?
- What do the graphs of $y = \sin(x)$ and $y = \cos(x)$ suggest as formulas for their respective derivatives?
- Once we know the derivatives of $\sin(x)$ and $\cos(x)$, how do previous derivative rules work when these functions are involved?

Throughout Chapter 2, we will develop shortcut derivative rules to help us bypass the limit definition and quickly compute $f'(x)$ from a formula for $f(x)$. In Section 2.1, we stated the rule for power functions,

$$\text{if } f(x) = x^n, \text{ then } f'(x) = nx^{n-1},$$

and the rule for exponential functions,

$$\text{if } a \text{ is a positive real number and } f(x) = a^x, \text{ then } f'(x) = a^x \ln(a).$$

Later in this section, we will use a graphical argument to conjecture derivative formulas for the sine and cosine functions.

Preview Activity 2.2.1

Consider the function $g(x) = 2^x$, which is graphed in Figure 2.2.1 .

1. At each of $x = -2, -1, 0, 1, 2$, use a straightedge to sketch an accurate tangent line to $y = g(x)$.
2. Use the provided grid to estimate the slope of the tangent line you drew at each point in (a).
3. Use the limit definition of the derivative to estimate $g'(0)$ by using small values of h, and compare the result to your visual estimate for the slope of the tangent line to $y = g(x)$ at $x = 0$ in (b).
4. Based on your work in (a), (b), and (c), sketch an accurate graph of $y = g'(x)$ on the axes adjacent to the graph of $y = g(x)$.
5. Write at least one sentence that explains why it is reasonable to think that $g'(x) = cg(x)$, where c is a constant. In addition, calculate $\ln(2)$, and then discuss how this value, combined with your work above, reasonably suggests that $g'(x) = 2^x \ln(2)$.

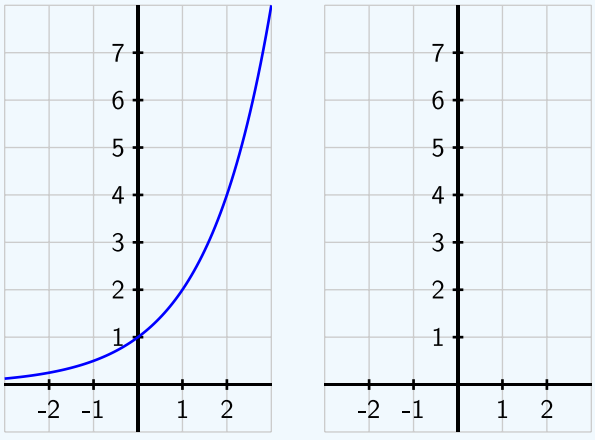


Figure 2.2.1. At left, the graph of $y = g(x) = 2^x$. At right, axes for plotting $y = g'(x)$.

The sine and cosine functions

The sine and cosine functions are among the most important functions in all of mathematics. Sometimes called the *circular* functions due to their definition on the unit circle, these periodic functions play a key role in modeling repeating phenomena such as tidal elevations, the behavior of an oscillating mass attached to a spring, or the location of a point on a bicycle tire. Like polynomial and exponential functions, the sine and cosine functions are considered basic functions, ones that are often used in building more complicated functions. As such, we would like to know formulas for $\frac{d}{dx}[\sin(x)]$ and $\frac{d}{dx}[\cos(x)]$, and the next two activities lead us to that end.

Activity 2.2.2

Consider the function $f(x) = \sin(x)$, which is graphed in Figure 2.2.2 below. Note carefully that the grid in the diagram does not have boxes that are 1×1, but rather approximately 1.57×1, as the horizontal scale of the grid is $\pi/2$ units per box.

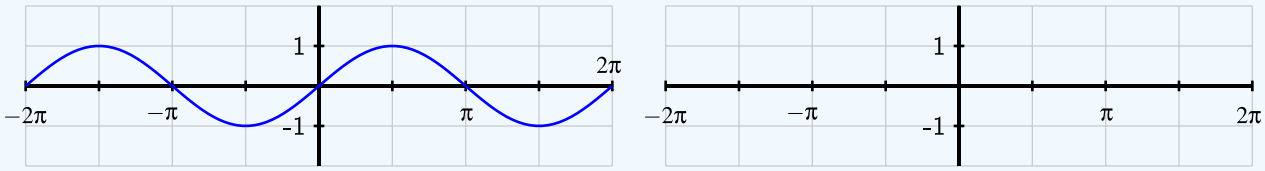


Figure 2.2.2 . At left, the graph of $y = f(x) = \sin(x)$.

1. At each of $x = -2\pi, -\frac{3\pi}{2}, -\pi, -\frac{\pi}{2}, 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}, 2\pi$, use a straightedge to sketch an accurate tangent line to $y = f(x)$.
2. Use the provided grid to estimate the slope of the tangent line you drew at each point. Pay careful attention to the scale of the grid.
3. Use the limit definition of the derivative to estimate $f'(0)$ by using small values of h, and compare the result to your visual estimate for the slope of the tangent line to $y = f(x)$ at $x = 0$ in (b). Using periodicity, what does this result suggest about $f'(2\pi)$? about $f'(-2\pi)$?
4. Based on your work in (a), (b), and (c), sketch an accurate graph of $y = f'(x)$ on the axes adjacent to the graph of $y = f(x)$.
5. What familiar function do you think is the derivative of $f(x) = \sin(x)$?

Activity 2.2.3

Consider the function $g(x) = \cos(x)$, which is graphed in Figure 2.2.3 below. Note carefully that the grid in the diagram does not have boxes that are 1×1, but rather approximately 1.57×1, as the horizontal scale of the grid is $\pi/2$ units per box.

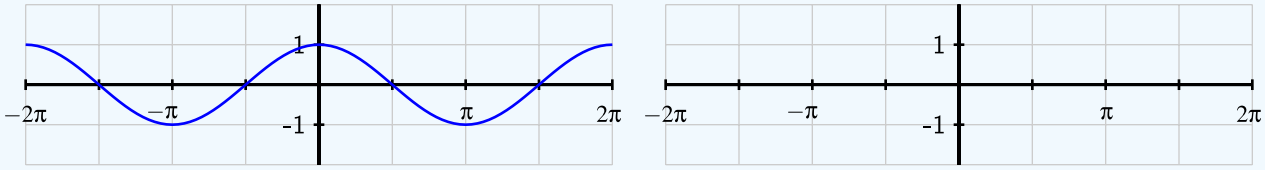


Figure 2.2.3 . At left, the graph of $y = g(x) = \cos(x)$.

1. At each of $x = -2\pi, -\frac{3\pi}{2}, -\pi, -\frac{\pi}{2}, 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}, 2\pi$, use a straightedge to sketch an accurate tangent line to $y = g(x)$.
2. Use the provided grid to estimate the slope of the tangent line you drew at each point. Again, note the scale of the axes and grid.
3. Use the limit definition of the derivative to estimate $g'(\frac{\pi}{2})$ by using small values of h, and compare the result to your visual estimate for the slope of the tangent line to $y = g(x)$ at $x = \frac{\pi}{2}$ in (b). Using periodicity, what does this result suggest about $g'(-\frac{3\pi}{2})$? can symmetry on the graph help you estimate other slopes easily?

4. Based on your work in (a), (b), and (c), sketch an accurate graph of $y = g'(x)$ on the axes adjacent to the graph of $y = g(x)$.
5. What familiar function do you think is the derivative of $g(x) = \cos(x)$?

The results of the two preceding activities suggest that the sine and cosine functions not only have beautiful connections such as the identities $\sin^2(x) + \cos^2(x) = 1$ and $\cos(x - \frac{\pi}{2}) = \sin(x)$, but that they are even further linked through calculus, as the derivative of each involves the other. The following rules summarize the results of the activities [1].

These two rules may be formally proved using the limit definition of the derivative and the expansion identities for $\sin(x + h)$ and $\cos(x + h)$.

Sine and Cosine Functions

For all real numbers x,

$$\frac{d}{dx}[\sin(x)] = \cos(x) \quad \text{and} \quad \frac{d}{dx}[\cos(x)] = -\sin(x).$$

We have now added the sine and cosine functions to our library of basic functions whose derivatives we know. The constant multiple and sum rules still hold, of course, as well as all of the inherent meaning of the derivative.

Activity 2.2.4

Answer each of the following questions. Where a derivative is requested, be sure to label the derivative function with its name using proper notation.

1. Determine the derivative of $h(t) = 3\cos(t) - 4\sin(t)$.
2. Find the exact slope of the tangent line to $y = f(x) = 2x + \frac{\sin(x)}{2}$ at the point where $x = \frac{\pi}{6}$.
3. Find the equation of the tangent line to $y = g(x) = x^2 + 2\cos(x)$ at the point where $x = \frac{\pi}{2}$.
4. Determine the derivative of $p(z) = z^4 + 4^z + 4\cos(z) - \sin(\frac{\pi}{2})$.
5. The function $P(t) = 24 + 8\sin(t)$ represents a population of a particular kind of animal that lives on a small island, where P is measured in hundreds and t is measured in decades since January 1, 2010. What is the instantaneous rate of change of P on January 1, 2030? What are the units of this quantity? Write a sentence in everyday language that explains how the population is behaving at this point in time.

Summary

- For an exponential function $f(x) = a^x$ $(a > 1)$, the graph of $f'(x)$ appears to be a scaled version of the original function. In particular, careful analysis of the graph of$f(x) = 2^x$, suggests that $\frac{d}{dx}[2^x] = 2^x \ln(2)$, which is a special case of the rule we stated in Section 2.1.
- By carefully analyzing the graphs of $y = \sin(x)$ and $y = \cos(x)$, and by using the limit definition of the derivative at select points, we found that $\frac{d}{dx}[\sin(x)] = \cos(x)$ and $\frac{d}{dx}[\cos(x)] = -\sin(x)$.
- We note that all previously encountered derivative rules still hold, but now may also be applied to functions involving the sine and cosine. All of the established meaning of the derivative applies to these trigonometric functions as well.

This page titled 2.2: The Sine and Cosine Function is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.2: The Sine and Cosine Function by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

2.3: The Product and Quotient Rules

Motivating Questions

- How does the algebraic structure of a function guide us in computing its derivative using shortcut rules?
- How do we compute the derivative of a product of two basic functions in terms of the derivatives of the basic functions?
- How do we compute the derivative of a quotient of two basic functions in terms of the derivatives of the basic functions?
- How do the product and quotient rules combine with the sum and constant multiple rules to expand the library of functions we can differentiate quickly?

So far, we can differentiate power functions (x^n), exponential functions (a^x), and the two fundamental trigonometric functions ($\sin(x)$ and $\cos(x)$). With the sum rule and constant multiple rules, we can also compute the derivative of combined functions.

Example 2.3.1

Differentiate

$$f(x) = 7x^{11} - 4 \cdot 9^x + \pi \sin(x) - \sqrt{3}\cos(x)$$

Because f is a sum of basic functions, we can now quickly say that $f'(x) = 77x^{10} - 4 \cdot 9^x \ln(9) + \pi \cos(x) + \sqrt{3}\sin(x)$.

What about a product or quotient of two basic functions, such as

$$p(z) = z^3 \cos(z),$$

or

$$q(t) = \frac{\sin(t)}{2^t}?$$

While the derivative of a sum is the sum of the derivatives, it turns out that the rules for computing derivatives of products and quotients are more complicated.

Preview Activity 2.3.1

Let f and g be the functions defined by $f(t) = 2t^2$ and $g(t) = t^3 + 4t$.

1. Determine $f'(t)$ and $g'(t)$.
2. Let $p(t) = 2t^2(t^3 + 4t)$ and observe that $p(t) = f(t) \cdot g(t)$. Rewrite the formula for p by distributing the $2t^2$ term. Then, compute $p'(t)$ using the sum and constant multiple rules.
3. True or false: $p'(t) = f'(t) \cdot g'(t)$.
4. Let $q(t) = \frac{t^3+4t}{2t^2}$ and observe that $q(t) = \frac{g(t)}{f(t)}$. Rewrite the formula for q by dividing each term in the numerator by the denominator and simplify to write q as a sum of constant multiples of powers of t. Then, compute $q'(t)$ using the sum and constant multiple rules.
5. True or false: $q'(t) = \frac{g'(t)}{f'(t)}$.

The product rule

As part (b) of Preview Activity 2.3.1 shows, it is not true in general that the derivative of a product of two functions is the product of the derivatives of those functions. To see why this is the case, we consider an example involving meaningful functions.

Say that an investor is regularly purchasing stock in a particular company. Let $N(t)$ represent the number of shares owned on day t, where $t = 0$ represents the first day on which shares were purchased. Let $S(t)$ give the value of one share of the stock on day t; note that the units on $S(t)$ are dollars per share. To compute the total value of the stock on day t, we take the product

$$V(t) = N(t)\,\text{shares} \cdot S(t)\,\text{dollars per share}.$$

Observe that over time, both the number of shares and the value of a given share will vary. The derivative $N'(t)$ measures the rate at which the number of shares is changing, while $S'(t)$ measures the rate at which the value per share is changing. How do these respective rates of change affect the rate of change of the total value function?

To help us understand the relationship among changes in N, S, and V, let's consider some specific data.

- Suppose that on day 100, the investor owns 520 shares of stock and the stock's current value is \$27.50 per share. This tells us that $N(100) = 520$ and $S(100) = 27.50$.
- On day 100, the investor purchases an additional 12 shares (so the number of shares held is rising at a rate of 12 shares per day).
- On that same day the price of the stock is rising at a rate of 0.75 dollars per share per day.

In calculus notation, the latter two facts tell us that $N'(100) = 12$ (shares per day) and $S'(100) = 0.75$ (dollars per share per day). At what rate is the value of the investor's total holdings changing on day 100?

Observe that the increase in total value comes from two sources: the growing number of shares, and the rising value of each share. If only the number of shares is increasing (and the value of each share is constant), the rate at which which total value would rise is the product of the current value of the shares and the rate at which the number of shares is changing. That is, the rate at which total value would change is given by

$$S(100) \cdot N'(100) = 27.50\,\frac{\text{dollars}}{\text{share}} \cdot 12\,\frac{\text{shares}}{\text{day}} = 330\,\frac{\text{dollars}}{\text{day}}.$$

Note particularly how the units make sense and show the rate at which the total value V is changing, measured in dollars per day.

If instead the number of shares is constant, but the value of each share is rising, the rate at which the total value would rise is the product of the number of shares and the rate of change of share value. The total value is rising at a rate of

$$N(100) \cdot S'(100) = 520\,\text{shares} \cdot 0.75\,\frac{\text{dollars per share}}{\text{day}} = 390\,\frac{\text{dollars}}{\text{day}}.$$

Of course, when both the number of shares and the value of each share are changing, we have to include both of these sources. In that case the rate at which the total value is rising is

$$V'(100) = S(100) \cdot N'(100) + N(100) \cdot S'(100) = 330 + 390 = 720\,\frac{\text{dollars}}{\text{day}}.$$

We expect the total value of the investor's holdings to rise by about \$720 on the 100th day.[1]

While this example highlights why the product rule is true, there are some subtle issues to recognize. For one, if the stock's value really does rise exactly \$0.75 on day 100, and the number of shares really rises by 12 on day 100, then we'd expect that $V(101) = N(101) \cdot S(101) = 532 \cdot 28.25 = 15029$. If, as noted above, we expect the total value to rise by \$720, then with $V(100) = N(100) \cdot S(100) = 520 \cdot 27.50 = 14300$, then it seems we should find that $V(101) = V(100) + 720 = 15020$. Why do the two results differ by 9? One way to understand why this difference occurs is to recognize that $N'(100) = 12$ represents an *instantaneous* rate of change, while our (informal) discussion has also thought of this number as the total change in the number of shares over the course of a single day. The formal proof of the product rule reconciles this issue by taking the limit as the change in the input tends to zero.

Next, we expand our perspective from the specific example above to the more general and abstract setting of a product p of two differentiable functions, f and g. If $P(x) = f(x) \cdot g(x)$, our work above suggests that $P'(x) = f(x)g'(x) + g(x)f'(x)$. Indeed, a formal proof using the limit definition of the derivative can be given to show that the following rule, called the *product rule*, holds in general.

Product Rule

If f and g are differentiable functions, then their product $P(x) = f(x) \cdot g(x)$ is also a differentiable function, and

$$P'(x) = f(x)g'(x) + g(x)f'(x).$$

In light of the earlier example involving shares of stock, the product rule also makes sense intuitively: the rate of change of P should take into account both how fast f and g are changing, as well as how large f and g are at the point of interest. In words the product rule says: if P is the product of two functions f (the first function) and g (the second), then "the derivative of P is the first times the derivative of the second, plus the second times the derivative of the first." It is often a helpful mental exercise to say this phrasing aloud when executing the product rule.

Example 2.3.2

If $P(z) = z^3 \cdot \cos(z)$, we can use the product rule to differentiate P. The first function is z^3 and the second function is $\cos(z)$. By the product rule, P' will be given by the first, z^3, times the derivative of the second, $-\sin(z)$, plus the second, $\cos(z)$, times the derivative of the first, $3z^2$. That is,

$$P'(z) = z^3(-\sin(z)) + \cos(z)3z^2 = -z^3\sin(z) + 3z^2\cos(z).$$

Activity 2.3.2

Use the product rule to answer each of the questions below. Throughout, be sure to carefully label any derivative you find by name. It is not necessary to algebraically simplify any of the derivatives you compute.

1. Let $m(w) = 3w^{17}4^w$. Find $m'(w)$.
2. Let $h(t) = (\sin(t) + \cos(t))t^4$. Find $h'(t)$.
3. Determine the slope of the tangent line to the curve $y = f(x)$ at the point where $a = 1$ if f is given by the rule $f(x) = e^x \sin(x)$.
4. Find the tangent line approximation $L(x)$ to the function $y = g(x)$ at the point where $a = -1$ if g is given by the rule $g(x) = (x^2 + x)2^x$.

The quotient rule

Because quotients and products are closely linked, we can use the product rule to understand how to take the derivative of a quotient. Let $Q(x)$ be defined by $Q(x) = f(x)/g(x)$, where f and g are both differentiable functions. It turns out that Q is differentiable everywhere that $g(x) \neq 0$. We would like a formula for Q' in terms of f, g, f', and g'. Multiplying both sides of the formula $Q = f/g$ by g, we observe that

$$f(x) = Q(x) \cdot g(x).$$

Now we can use the product rule to differentiate f.

$$f'(x) = Q(x)g'(x) + g(x)Q'(x).$$

We want to know a formula for Q', so we solve this equation for $Q'(x)$.

$$Q'(x)g(x) = f'(x) - Q(x)g'(x)$$

and dividing both sides by $g(x)$, we have

$$Q'(x) = \frac{f'(x) - Q(x)g'(x)}{g(x)}.$$

Finally, we recall that $Q(x) = \frac{f(x)}{g(x)}$. Substituting this expression in the preceding equation, we have

$$\begin{aligned} Q'(x) &= \frac{f'(x) - \frac{f(x)}{g(x)}g'(x)}{g(x)} \\ &= \frac{f'(x) - \frac{f(x)}{g(x)}g'(x)}{g(x)} \cdot \frac{g(x)}{g(x)} \\ &= \frac{g(x)f'(x) - f(x)g'(x)}{g(x)^2}. \end{aligned}$$

This calculation gives us the *quotient rule*.

Quotient Rule

If f and g are differentiable functions, then their quotient $Q(x) = \frac{f(x)}{g(x)}$ is also a differentiable function for all x where $g(x) \neq 0$ and

$$Q'(x) = \frac{g(x)f'(x) - f(x)g'(x)}{g(x)^2}.$$

As with the product rule, it can be helpful to think of the quotient rule verbally. If a function Q is the quotient of a top function f and a bottom function g, then Q' is given by "the bottom times the derivative of the top, minus the top times the derivative of the bottom, all over the bottom squared."

Example 2.3.3

If $Q(t) = \sin(t)/2^t$, we call $\sin(t)$ the top function and 2^t the bottom function. By the quotient rule, Q' is given by the bottom, 2^t, times the derivative of the top, $\cos(t)$, minus the top, $\sin(t)$, times the derivative of the bottom, $2^t \ln(2)$, all over the bottom squared, $(2^t)^2$. That is,

$$Q'(t) = \frac{2^t \cos(t) - \sin(t)2^t \ln(2)}{(2^t)^2}.$$

In this particular example, it is possible to simplify $Q'(t)$ by removing a factor of 2^t from both the numerator and denominator, so that

$$Q'(t) = \frac{\cos(t) - \sin(t)\ln(2)}{2^t}.$$

In general, we must be careful in doing any such simplification, as we don't want to execute the quotient rule correctly but then make an algebra error.

Activity 2.3.3

Use the quotient rule to answer each of the questions below. Throughout, be sure to carefully label any derivative you find by name. That is, if you're given a formula for $f(x)$, clearly label the formula you find for $f'(x)$. It is not necessary to algebraically simplify any of the derivatives you compute.

1. Let $r(z) = \frac{3^z}{z^4+1}$. Find $r'(z)$.
2. Let $v(t) = \frac{\sin(t)}{\cos(t)+t^2}$. Find $v'(t)$.
3. Determine the slope of the tangent line to the curve $R(x) = \dfrac{x^2 - 2x - 8}{x^2 - 9}$ at the point where $x = 0$.
4. When a camera flashes, the intensity I of light seen by the eye is given by the function

$$I(t) = \frac{100t}{e^t},$$

where I is measured in candles and t is measured in milliseconds. Compute $I'(0.5)$, $I'(2)$, and $I'(5)$; include appropriate units on each value; and discuss the meaning of each.

Combining rules

In order to apply the derivative shortcut rules correctly we must recognize the fundamental structure of a function.

Example 2.3.4

Determine the derivative of the function

$$f(x) = x\sin(x) + \frac{x^2}{\cos(x)+2}.$$

How do we decide which rules to apply? Our first task is to recognize the structure of the function. This function f is a sum of two slightly less complicated functions, so we can apply the sum rule[2] to get

$$\begin{aligned} f'(x) &= \frac{d}{dx}\left[x\sin(x) + \frac{x^2}{\cos(x)+2}\right] \\ &= \frac{d}{dx}\left[x\sin(x)\right] + \frac{d}{dx}\left[\frac{x^2}{\cos(x)+2}\right] \end{aligned}$$

When taking a derivative that involves the use of multiple derivative rules, it is often helpful to use the notation $\frac{d}{dx}[\ ]$ to wait to apply subsequent rules. This is demonstrated in each of the two examples presented here.

Now, the left-hand term above is a product, so the product rule is needed there, while the right-hand term is a quotient, so the quotient rule is required. Applying these rules respectively, we find that

$$\begin{aligned} f'(x) &= (x\cos(x) + \sin(x)) + \frac{(\cos(x)+2)2x - x^2(-\sin(x))}{(\cos(x)+2)^2} \\ &= x\cos(x) + \sin(x) + \frac{2x\cos(x) + 4x + x^2\sin(x)}{(\cos(x)+2)^2}. \end{aligned}$$

Example 2.3.5

Differentiate

$$s(y) = \frac{y\cdot 7^y}{y^2+1}.$$

The function s is a quotient of two simpler functions, so the quotient rule will be needed. To begin, we set up the quotient rule and use the notation $\frac{d}{dy}$ to indicate the derivatives of the numerator and denominator. Thus,

$$s'(y) = \frac{(y^2+1)\cdot\frac{d}{dy}\left[y\cdot 7^y\right] - y\cdot 7^y\cdot\frac{d}{dy}\left[y^2+1\right]}{(y^2+1)^2}.$$

Now, there remain two derivatives to calculate. The first one, $\frac{d}{dy}\left[y\cdot 7^y\right]$ calls for use of the product rule, while the second, $\frac{d}{dy}\left[y^2+1\right]$ needs only the sum rule. Applying these rules, we now have

$$s'(y) = \frac{(y^2+1)[y\cdot 7^y\ln(7) + 7^y\cdot 1] - y\cdot 7^y[2y]}{(y^2+1)^2}.$$

While some simplification is possible, we are content to leave $s'(y)$ in its current form.

Success in applying derivative rules begins with recognizing the structure of the function, followed by the careful and diligent application of the relevant derivative rules. The best way to become proficient at this process is to do a large number of examples.

Activity 2.3.4

Use relevant derivative rules to answer each of the questions below. Throughout, be sure to use proper notation and carefully label any derivative you find by name.

1. Let $f(r) = (5r^3 + \sin(r))(4^r - 2\cos(r))$. Find $f'(r)$.
2. Let $p(t) = \dfrac{\cos(t)}{t^6 \cdot 6^t}$. Find $p'(t)$.
3. Let $g(z) = 3z^7 e^z - 2z^2 \sin(z) + \frac{z}{z^2+1}$. Find $g'(z)$.
4. A moving particle has its position in feet at time t in seconds given by the function $s(t) = \frac{3\cos(t) - \sin(t)}{e^t}$. Find the particle's instantaneous velocity at the moment $t = 1$.
5. Suppose that $f(x)$ and $g(x)$ are differentiable functions and it is known that $f(3) = -2$, $f'(3) = 7$, $g(3) = 4$, and $g'(3) = -1$. If $p(x) = f(x) \cdot g(x)$ and $q(x) = \dfrac{f(x)}{g(x)}$, calculate $p'(3)$ and $q'(3)$.

As the algebraic complexity of the functions we are able to differentiate continues to increase, it is important to remember that all of the derivative's meaning continues to hold. Regardless of the structure of the function f, the value of $f'(a)$ tells us the instantaneous rate of change of f with respect to x at the moment $x = a$, as well as the slope of the tangent line to $y = f(x)$ at the point $(a, f(a))$.

Summary

- If a function is a sum, product, or quotient of simpler functions, then we can use the sum, product, or quotient rules to differentiate it in terms of the simpler functions and their derivatives.
- The product rule tells us that if P is a product of differentiable functions f and g according to the rule $P(x) = f(x)g(x)$, then
$$P'(x) = f(x)g'(x) + g(x)f'(x).$$
- The quotient rule tells us that if Q is a quotient of differentiable functions f and g according to the rule $Q(x) = \frac{f(x)}{g(x)}$, then
$$Q'(x) = \frac{g(x)f'(x) - f(x)g'(x)}{g(x)^2}.$$
- Along with the constant multiple and sum rules, the product and quotient rules enable us to compute the derivative of any function that consists of sums, constant multiples, products, and quotients of basic functions. For instance, if F has the form
$$F(x) = \frac{2a(x) - 5b(x)}{c(x) \cdot d(x)},$$
then F is a quotient, in which the numerator is a sum of constant multiples and the denominator is a product. This, the derivative of F can be found by applying the quotient rule and then using the sum and constant multiple rules to differentiate the numerator and the product rule to differentiate the denominator.

This page titled 2.3: The Product and Quotient Rules is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.3: The Product and Quotient Rules by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

2.4: Derivatives of Other Trigonometric Functions

Motivating Questions

- What are the derivatives of the tangent, cotangent, secant, and cosecant functions?
- How do the derivatives of $\tan(x)$, $\cot(x)$, $\sec(x)$, and $\csc(x)$ combine with other derivative rules we have developed to expand the library of functions we can quickly differentiate?

One of the powerful themes in trigonometry comes from a very simple idea: locating a point on the unit circle.

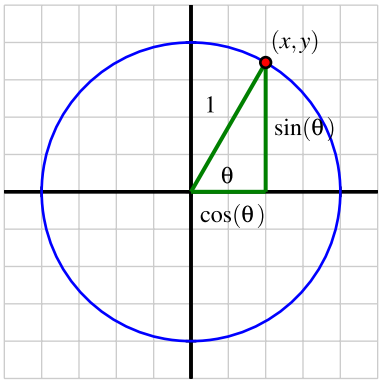


Figure 2.4.1. The unit circle and the definition of the sine and cosine functions.

Because each angle θ in standard position corresponds to one and only one point (x, y) on the unit circle, the x- and y-coordinates of this point are each functions of θ. In fact, this is the very definition of $\cos(\theta)$ and $\sin(\theta)$: $\cos(\theta)$ is the x-coordinate of the point on the unit circle corresponding to the angle θ, and $\sin(\theta)$ is the y-coordinate. From this simple definition, all of trigonometry is founded. For instance, the Fundamental Trigonometric Identity,

$$\sin^2(\theta) + \cos^2(\theta) = 1,$$

is a restatement of the Pythagorean Theorem, applied to the right triangle shown in Figure 2.4.1 .

There are four other trigonometric functions, each defined in terms of the sine and/or cosine functions.

- The tangent function is defined by $\tan(\theta) = \frac{\sin(\theta)}{\cos(\theta)}$;
- the cotangent function is its reciprocal: $\cot(\theta) = \frac{\cos(\theta)}{\sin(\theta)}$.
- The secant function is the reciprocal of the cosine function, $\sec(\theta) = \frac{1}{\cos(\theta)}$;
- and the cosecant function is the reciprocal of the sine function, $\csc(\theta) = \frac{1}{\sin(\theta)}$.

These six trigonometric functions together offer us a wide range of flexibility in problems involving right triangles.

Because we know the derivatives of the sine and cosine function, we can now develop shortcut differentiation rules for the tangent, cotangent, secant, and cosecant functions. In this section's preview activity, we work through the steps to find the derivative of $y = \tan(x)$.

Preview Activity 2.4.1

Consider the function $f(x) = \tan(x)$, and remember that $\tan(x) = \frac{\sin(x)}{\cos(x)}$.

1. What is the domain of f?
2. Use the quotient rule to show that one expression for $f'(x)$ is
$$f'(x) = \frac{\cos(x)\cos(x) + \sin(x)\sin(x)}{\cos^2(x)}.$$
3. What is the Fundamental Trigonometric Identity? How can this identity be used to find a simpler form for $f'(x)$?
4. Recall that $\sec(x) = \frac{1}{\cos(x)}$. How can we express $f'(x)$ in terms of the secant function?
5. For what values of x is $f'(x)$ defined? How does this set compare to the domain of f?

Derivatives of the cotangent, secant, and cosecant functions

In Preview Activity 2.4.1 , we found that the derivative of the tangent function can be expressed in several ways, but most simply in terms of the secant function. Next, we develop the derivative of the cotangent function.

Let $g(x) = \cot(x)$. To find $g'(x)$, we observe that $g(x) = \frac{\cos(x)}{\sin(x)}$ and apply the quotient rule. Hence

$$g'(x) = \frac{\sin(x)(-\sin(x)) - \cos(x)\cos(x)}{\sin^2(x)}$$
$$= -\frac{\sin^2(x) + \cos^2(x)}{\sin^2(x)}$$

By the Fundamental Trigonometric Identity, we see that $g'(x) = -\frac{1}{\sin^2(x)}$, and recalling that $\csc(x) = \frac{1}{\sin(x)}$, it follows that we can express g' by the rule

$$g'(x) = -\csc^2(x).$$

Note that neither g nor g' is defined when $\sin(x) = 0$, which occurs at every integer multiple of π. Hence we have the following rule.

Cotangent Function

For all real numbers x such that $x \neq k\pi$, where $k = 0, \pm 1, \pm 2, \ldots$,

$$\frac{d}{dx}[\cot(x)] = -\csc^2(x).$$

Notice that the derivative of the cotangent function is very similar to the derivative of the tangent function we discovered in Preview Activity 2.4.1.

Tangent Function

For all real numbers x such that $x \neq \frac{(2k+1)\pi}{2}$, where $k = \pm 1, \pm 2, \ldots$,

$$\frac{d}{dx}[\tan(x)] = \sec^2(x).$$

In the next two activities, we develop the rules for differentiating the secant and cosecant functions.

Activity 2.4.2

Let $h(x) = \sec(x)$ and recall that $\sec(x) = \frac{1}{\cos(x)}$.

1. What is the domain of h?
2. Use the quotient rule to develop a formula for $h'(x)$ that is expressed completely in terms of $\sin(x)$ and $\cos(x)$.
3. How can you use other relationships among trigonometric functions to write $h'(x)$ only in terms of $\tan(x)$ and $\sec(x)$?
4. What is the domain of h'? How does this compare to the domain of h?

Activity 2.4.3

Let $p(x) = \csc(x)$ and recall that $\csc(x) = \frac{1}{\sin(x)}$.

1. What is the domain of p?
2. Use the quotient rule to develop a formula for $p'(x)$ that is expressed completely in terms of $\sin(x)$ and $\cos(x)$.
3. How can you use other relationships among trigonometric functions to write $p'(x)$ only in terms of $\cot(x)$ and $\csc(x)$?
4. What is the domain of p'? How does this compare to the domain of p?

Using the quotient rule we have determined the derivatives of the tangent, cotangent, secant, and cosecant functions, expanding our overall library of functions we can differentiate. Observe that just as the derivative of any polynomial function is a polynomial, and the derivative of any exponential function is another exponential function, so it is that the derivative of any basic trigonometric function is another function that consists of basic trigonometric functions. This makes sense because all trigonometric functions are periodic, and hence their derivatives will be periodic, too.

The derivative retains all of its fundamental meaning as an instantaneous rate of change and as the slope of the tangent line to the function under consideration.

Activity 2.4.4

Answer each of the following questions. Where a derivative is requested, be sure to label the derivative function with its name using proper notation.

1. Let $f(x) = 5\sec(x) - 2\csc(x)$. Find the slope of the tangent line to f at the point where $x = \frac{\pi}{3}$.
2. Let $p(z) = z^2\sec(z) - z\cot(z)$. Find the instantaneous rate of change of p at the point where $z = \frac{\pi}{4}$.
3. Let $h(t) = \dfrac{\tan(t)}{t^2+1} - 2e^t\cos(t)$. Find $h'(t)$.
4. Let $g(r) = \dfrac{r\sec(r)}{5^r}$. Find $g'(r)$.
5. When a mass hangs from a spring and is set in motion, the object's position oscillates in a way that the size of the oscillations decrease. This is usually called a *damped oscillation*. Suppose that for a particular object, its displacement from equilibrium (where the object sits at rest) is modeled by the function

$$s(t) = \frac{15\sin(t)}{e^t}.$$

Assume that s is measured in inches and t in seconds. Sketch a graph of this function for $t \geq 0$ to see how it represents the situation described. Then compute ds/dt, state the units on this function, and explain what it tells you about the object's motion. Finally, compute and interpret $s'(2)$.

Summary

- The derivatives of the other four trigonometric functions are

$$\frac{d}{dx}[\tan(x)] = \sec^2(x), \quad \frac{d}{dx}[\cot(x)] = -\csc^2(x),$$
$$\frac{d}{dx}[\sec(x)] = \sec(x)\tan(x), \text{ and } \frac{d}{dx}[\csc(x)] = -\csc(x)\cot(x).$$

Each derivative exists and is defined on the same domain as the original function. For example, both the tangent function and its derivative are defined for all real numbers x such that $x \neq \frac{k\pi}{2}$, where $k = \pm 1, \pm 2, \ldots$.

- The four rules for the derivatives of the tangent, cotangent, secant, and cosecant can be used along with the rules for power functions, exponential functions, and the sine and cosine, as well as the sum, constant multiple, product, and quotient rules, to quickly differentiate a wide range of different functions.

This page titled 2.4: Derivatives of Other Trigonometric Functions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.4: Derivatives of Other Trigonometric Functions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

2.5: The Chain Rule

Motivating Questions

- What is a composite function and how do we recognize its structure algebraically?
- Given a composite function $C(x) = f(g(x))$ that is built from differentiable functions f and g, how do we compute $C'(x)$ in terms of f, g, f', and g'? What is the statement of the Chain Rule?

In addition to learning how to differentiate a variety of basic functions, we have also been developing our ability to use rules to differentiate certain algebraic combinations of them.

Example 2.5.1

State the rule(s) to find the derivative of each of the following combinations of $f(x) = \sin(x)$ and $g(x) = x^2$:

$$s(x) = 3x^2 - 5\sin(x),$$

$$p(x) = x^2\sin(x), \text{and}$$

$$q(x) = \frac{\sin(x)}{x^2}.$$

Answer available at the end of this section.

There is one more natural way to combine basic functions algebraically, and that is by *composing* them. For instance, let's consider the function

$$C(x) = \sin(x^2),$$

and observe that any input x passes through a *chain* of functions. In the process that defines the function $C(x)$, x is first squared, and then the sine of the result is taken. Using an arrow diagram,

$$x \longrightarrow x^2 \longrightarrow \sin(x^2).$$

In terms of the elementary functions f and g, we observe that x is the input for the function g, and the result is used as the input for f. We write

$$C(x) = f(g(x)) = \sin(x^2)$$

and say that C is the *composition* of f and g. We will refer to g, the function that is first applied to x, as the *inner* function, while f, the function that is applied to the result, is the *outer* function.

Given a composite function $C(x) = f(g(x))$ that is built from differentiable functions f and g, how do we compute $C'(x)$ in terms of f, g, f', and g'? In the same way that the rate of change of a product of two functions, $p(x) = f(x) \cdot g(x)$, depends on the behavior of both f and g, it makes sense intuitively that the rate of change of a composite function $C(x) = f(g(x))$ will also depend on some combination of f and g and their derivatives. The rule that describes how to compute C' in terms of f and g and their derivatives is called the *chain rule*.

But before we can learn what the chain rule says and why it works, we first need to be comfortable decomposing composite functions so that we can correctly identify the inner and outer functions, as we did in the example above with $C(x) = \sin(x^2)$.

Preview Activity 2.5.1

For each function given below, identify its fundamental algebraic structure. In particular, is the given function a sum, product, quotient, or composition of basic functions? If the function is a composition of basic functions, state a formula for the inner function g and the outer function f so that the overall composite function can be written in the form $f(g(x))$. If the function is a sum, product, or quotient of basic functions, use the appropriate rule to determine its derivative.

1. $h(x) = \tan(2^x)$
2. $p(x) = 2^x \tan(x)$
3. $r(x) = (\tan(x))^2$
4. $m(x) = e^{\tan(x)}$
5. $w(x) = \sqrt{x} + \tan(x)$
6. $z(x) = \sqrt{\tan(x)}$

The chain rule

Often a composite function cannot be written in an alternate algebraic form. For instance, the function $C(x) = \sin(x^2)$ cannot be expanded or otherwise rewritten, so it presents no alternate approaches to taking the derivative. But some composite functions can be expanded or simplified, and these provide a way to explore how the chain rule works.

Example 2.5.2

Let $f(x) = -4x + 7$ and $g(x) = 3x - 5$. Determine a formula for $C(x) = f(g(x))$ and compute $C'(x)$. How is C' related to f and g and their derivatives?

Answer available at the end of this section.

It may seem that Example 2.5.2 is too elementary to illustrate how to differentiate a composite function. Linear functions are the simplest of all functions, and composing linear functions yields another linear function. While this example does not illustrate the full complexity of a composition of nonlinear functions, at the same time we remember that any differentiable function is *locally* linear, and thus any function with a derivative behaves like a line when viewed up close. The fact that the derivatives of the linear functions f and g are multiplied to find the derivative of their composition turns out to be a key insight.

We now consider a composition involving a nonlinear function.

Example 2.5.3

Let $C(x) = \sin(2x)$. Use the double angle identity to rewrite C as a product of basic functions, and use the product rule to find C'. Rewrite C' in the simplest form possible.

Answer available at the end of this section.

In Example 2.5.3 , if we let $g(x) = 2x$ and $f(x) = \sin(x)$, we observe that $C(x) = f(g(x))$. Now, $g'(x) = 2$ and $f'(x) = \cos(x)$, so we can view the structure of $C'(x)$ as

$$C'(x) = 2\cos(2x) = g'(x)f'(g(x)).$$

In this example, as in the example involving linear functions, we see that the derivative of the composite function $C(x) = f(g(x))$ is found by multiplying the derivatives of f and g, but with f' evaluated at $g(x)$.

It makes sense intuitively that these two quantities are involved in the rate of change of a composite function: if we ask how fast C is changing at a given x value, it clearly matters how fast g is changing at x, as well as how fast f is changing at the value of $g(x)$. It turns out that this structure holds for all differentiable functions [1] as is stated in the Chain Rule.

Like other differentiation rules, the Chain Rule can be proved formally using the limit definition of the derivative.

Chain Rule

If g is differentiable at x and f is differentiable at $g(x)$, then the composite function C defined by $C(x) = f(g(x))$ is differentiable at x and

$$C'(x) = f'(g(x))g'(x).$$

As with the product and quotient rules, it is often helpful to think verbally about what the chain rule says: "If C is a composite function defined by an outer function f and an inner function g, then C' is given by the derivative of the outer function evaluated at the inner function, times the derivative of the inner function."

It is helpful to identify clearly the inner function g and outer function f, compute their derivatives individually, and then put all of the pieces together by the chain rule.

Example 2.5.4

Determine the derivative of the function

$$r(x) = (\tan(x))^2.$$

Answer available at the end of this section.

As a side note, we remark that $r(x)$ is usually written as $\tan^2(x)$. This is common notation for powers of trigonometric functions: $\cos^4(x)$, $\sin^5(x)$, and $\sec^2(x)$ are all composite functions, with the outer function a power function and the inner function a trigonometric one.

Activity 2.5.2

For each function given below, identify an inner function g and outer function f to write the function in the form $f(g(x))$. Determine $f'(x)$, $g'(x)$, and $f'(g(x))$, and then apply the chain rule to determine the derivative of the given function.

1. $h(x) = \cos(x^4)$
2. $p(x) = \sqrt{\tan(x)}$
3. $s(x) = 2^{\sin(x)}$
4. $z(x) = \cot^5(x)$
5. $m(x) = (\sec(x) + e^x)^9$

Using multiple rules simultaneously

The chain rule now joins the sum, constant multiple, product, and quotient rules in our collection of techniques for finding the derivative of a function through understanding its algebraic structure and the basic functions that constitute it. It takes practice to get comfortable applying multiple rules to differentiate a single function, but using proper notation and taking a few extra steps will help.

Example 2.5.5

Find a formula for the derivative of $h(t) = 3^{t^2+2t}\sec^4(t)$.

Answer available at the end of this section.

Activity 2.5.3

For each of the following functions, find the function's derivative. State the rule(s) you use, label relevant derivatives appropriately, and be sure to clearly identify your overall answer.

1. $p(r) = 4\sqrt{r^6 + 2e^r}$
2. $m(v) = \sin(v^2)\cos(v^3)$
3. $h(y) = \dfrac{\cos(10y)}{e^{4y} + 1}$
4. $s(z) = 2^{z^2 \sec(z)}$
5. $c(x) = \sin(e^{x^2})$

The chain rule now adds substantially to our ability to compute derivatives. Whether we are finding the equation of the tangent line to a curve, the instantaneous velocity of a moving particle, or the instantaneous rate of change of a certain quantity, if the function under consideration is a composition, the chain rule is indispensable.

Activity 2.5.4

Use known derivative rules, including the chain rule, as needed to answer each of the following questions.

1. Find an equation for the tangent line to the curve $y = \sqrt{e^x + 3}$ at the point where $x = 0$.
2. If $s(t) = \dfrac{1}{(t^2+1)^3}$ represents the position function of a particle moving horizontally along an axis at time t (where s is measured in inches and t in seconds), find the particle's instantaneous velocity at $t = 1$. Is the particle moving to the left or right at that instant?
3. At sea level, air pressure is 30 inches of mercury. At an altitude of h feet above sea level, the air pressure, P, in inches of mercury, is given by the function $P = 30e^{-0.0000323h}$. Compute dP/dh and explain what this derivative function tells you about air pressure, including a discussion of the units on dP/dh. In addition, determine how fast the air pressure is changing for a pilot of a small plane passing through an altitude of 1000 feet.
4. Suppose that $f(x)$ and $g(x)$ are differentiable functions and that the following information about them is known:

   Table 2.5.1 . Data for functions f and g.

x	$f(x)$	$f'(x)$	$g(x)$	$g'(x)$
-1	2	-5	-3	4
2	-3	4	-1	2

   If $C(x)$ is a function given by the formula $f(g(x))$, determine $C'(2)$. In addition, if $D(x)$ is the function $f(f(x))$, find $D'(-1)$.

The composite version of basic function rules

As we gain more experience with differentiation, we will become more comfortable in simply writing down the derivative without taking multiple steps. This is particularly simple when the inner function is linear, since the derivative of a linear function is a constant.

Example 2.5.6

Use the chain rule to differentiate each of the following composite functions whose inside function is linear:

$$\frac{d}{dx}\left[(5x+7)^{10}\right] = 10(5x+7)^9 \cdot 5,$$

$$\frac{d}{dx}\left[\tan(17x)\right] = 17\sec^2(17x), \text{ and}$$

$$\frac{d}{dx}\left[e^{-3x}\right] = -3e^{-3x}.$$

More generally, following is an excellent exercise for getting comfortable with the derivative rules. Write down a list of all the basic functions whose derivatives we know, and list the derivatives. Then write a composite function with the inner function being an unknown function $u(x)$ and the outer function being a basic function. Finally, write the chain rule for the composite function. The following example illustrates this for two different functions.

Example 2.5.7

To determine

$$\frac{d}{dx}[\sin(u(x))],$$

where u is a differentiable function of x, we use the chain rule with the sine function as the outer function. Applying the chain rule, we find that

$$\frac{d}{dx}[\sin(u(x))] = \cos(u(x)) \cdot u'(x).$$

This rule is analogous to the basic derivative rule that $\frac{d}{dx}[\sin(x)] = \cos(x)$.

Similarly, since $\frac{d}{dx}[a^x] = a^x \ln(a)$, it follows by the chain rule that

$$\frac{d}{dx}[a^{u(x)}] = a^{u(x)} \ln(a) \cdot u'(x).$$

This rule is analogous to the basic derivative rule that $\frac{d}{dx}[a^x] = a^x \ln(a)$.

Summary

- A composite function is one where the input variable x first passes through one function, and then the resulting output passes through another. For example, the function $h(x) = 2^{\sin(x)}$ is composite since $x \longrightarrow \sin(x) \longrightarrow 2^{\sin(x)}$.
- Given a composite function $C(x) = f(g(x))$ where f and g are differentiable functions, the chain rule tells us that

$$C'(x) = f'(g(x))g'(x).$$

Answers

Example 2.5.1

Finding s' uses the sum and constant multiple rules, because $s(x) = 3g(x) - 5f(x)$. Determining p' requires the product rule, because $p(x) = g(x) \cdot f(x)$. To calculate q' we use the quotient rule, because $q(x) = \frac{f(x)}{g(x)}$.

Example 2.5.2

By the rules given for f and g,

$$\begin{aligned} C(x) &= f(g(x)) \\ &= f(3x-5) \\ &= -4(3x-5)+7 \\ &= -12x+20+7 \\ &= -12x+27. \end{aligned}$$

Thus, $C'(x) = -12$. Noting that $f'(x) = -4$ and $g'(x) = 3$, we observe that C' appears to be the product of f' and g'.

Example 2.5.3

Using the double angle identity for the sine function, we write

$$C(x) = \sin(2x) = 2\sin(x)\cos(x).$$

Applying the product rule and simplifying, we find

$$C'(x) = 2\sin(x)(-\sin(x)) + \cos(x)(2\cos(x)) = 2(\cos^2(x) - \sin^2(x)).$$

Next, we recall that the double angle identity for the cosine,

$$\cos(2x) = \cos^2(x) - \sin^2(x).$$

Substituting this result into our expression for $C'(x)$, we now have that

$$C'(x) = 2\cos(2x).$$

Example 2.5.4

The function r is composite, with inner function $g(x) = \tan(x)$ and outer function $f(x) = x^2$. Organizing the key information involving f, g, and their derivatives, we have

$f(x) = x^2$		$g(x) = \tan(x)$
$f'(x) = 2x$		$g'(x) = \sec^2(x)$
$f'(g(x)) = 2\tan(x)$		

Applying the chain rule, we find that

$$r'(x) = f'(g(x))g'(x) = 2\tan(x)\sec^2(x).$$

Example 2.5.5

We first observe that h is the product of two functions: $h(t) = a(t) \cdot b(t)$, where $a(t) = 3^{t^2+2t}$ and $b(t) = \sec^4(t)$. We will need to use the product rule to differentiate h. And because a and b are composite functions, we will need the chain rule. We therefore begin by computing $a'(t)$ and $b'(t)$.

Writing $a(t) = f(g(t)) = 3^{t^2+2t}$, and finding the derivatives of f and g, we have

$f(t) = 3^t$		$g(t) = t^2 + 2t$
$f'(t) = 3^t\ln(3)$		$g'(t) = 2t + 2$
$f'(g(t)) = 3^{t^2+2t}\ln(3)$		

Thus, by the chain rule, it follows that $a'(t) = f'(g(t))g'(t) = 3^{t^2+2t}\ln(3)(2t+2)$.

Turning next to b, we write $b(t) = r(s(t)) = \sec^4(t)$ and find the derivatives of r and s.

$r(t) = t^4$		$s(t) = \sec(t)$
$r'(t) = 4t^3$		$s'(t) = \sec(t)\tan(t)$
$r'(s(t)) = 4\sec^3(t)$		

By the chain rule,

$$b'(t) = r'(s(t))s'(t) = 4\sec^3(t)\sec(t)\tan(t) = 4\sec^4(t)\tan(t).$$

Now we are finally ready to compute the derivative of the function h. Recalling that $h(t) = 3^{t^2+2t}\sec^4(t)$, by the product rule we have

$$h'(t) = 3^{t^2+2t}\frac{d}{dt}[\sec^4(t)] + \sec^4(t)\frac{d}{dt}[3^{t^2+2t}].$$

From our work above with a and b, we know the derivatives of 3^{t^2+2t} and $\sec^4(t)$, and therefore

$$h'(t) = 3^{t^2+2t}4\sec^4(t)\tan(t) + \sec^4(t)3^{t^2+2t}\ln(3)(2t+2).$$

This page titled 2.5: The Chain Rule is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.5: The Chain Rule by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

2.6: Derivatives of Inverse Functions

Motivating Questions

- What is the derivative of the natural logarithm function?
- What are the derivatives of the inverse trigonometric functions $\arcsin(x)$ and $\arctan(x)$?
- If g is the inverse of a differentiable function f, how is g' computed in terms of f, f', and g?

Much of mathematics centers on the notion of function. Indeed, throughout our study of calculus, we are investigating the behavior of functions, with particular emphasis on how fast the output of the function changes in response to changes in the input. Because each function represents a process, a natural question to ask is whether or not the particular process can be reversed. That is, if we know the output that results from the function, can we determine the input that led to it? And if we know how fast a particular process is changing, can we determine how fast the inverse process is changing?

One of the most important functions in all of mathematics is the natural exponential function $f(x) = e^x$. Its inverse, the natural logarithm, $g(x) = \ln(x)$, is similarly important. One of our goals in this section is to learn how to differentiate the logarithm function. First, we review some of the basic concepts surrounding functions and their inverses.

Preview Activity 2.6.1

The equation $y = \frac{5}{9}(x - 32)$ relates a temperature given in x degrees Fahrenheit to the corresponding temperature y measured in degrees Celsius.

1. Solve the equation $y = \frac{5}{9}(x - 32)$ for x to write x (Fahrenheit temperature) in terms of y (Celsius temperature).
2. Let $C(x) = \frac{5}{9}(x - 32)$ be the function that takes a Fahrenheit temperature as input and produces the Celsius temperature as output. In addition, let $F(y)$ be the function that converts a temperature given in y degrees Celsius to the temperature $F(y)$ measured in degrees Fahrenheit. Use your work in (a) to write a formula for $F(y)$.
3. Next consider the new function defined by $p(x) = F(C(x))$. Use the formulas for F and C to determine an expression for $p(x)$ and simplify this expression as much as possible. What do you observe?
4. Now, let $r(y) = C(F(y))$. Use the formulas for F and C to determine an expression for $r(y)$ and simplify this expression as much as possible. What do you observe?
5. What is the value of $C'(x)$? of $F'(y)$? How do these values appear to be related?

Basic facts about inverse functions

A function $f : A \to B$ is a rule that associates each element in the set A to one and only one element in the set B. We call A the *domain* of f and B the *codomain* of f. If there exists a function $g : B \to A$ such that $g(f(a)) = a$ for every possible choice of a in the set A and $f(g(b)) = b$ for every b in the set B, then we say that g is the *inverse* of f.

We often use the notation f^{-1} (read "f-inverse") to denote the inverse of f. The inverse function undoes the work of f. Indeed, if $y = f(x)$, then

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

Thus, the equations $y = f(x)$ and $x = f^{-1}(y)$ say the same thing. The only difference between the two equations is one of perspective — one is solved for x, while the other is solved for y.

Here we briefly remind ourselves of some key facts about inverse functions.

Note 2.6.1

For a function $f : A \to B$,

- f has an inverse if and only if f is one-to-one [1] and onto [2];
- provided f^{-1} exists, the domain of f^{-1} is the codomain of f, and the codomain of f^{-1} is the domain of f;
- $f^{-1}(f(x)) = x$ for every x in the domain of f and $f(f^{-1}(y)) = y$ for every y in the codomain of f;
- $y = f(x)$ if and only if $x = f^{-1}(y)$.

A function f is *one-to-one* provided that no two distinct inputs lead to the same output.

A function f is *onto* provided that every possible element of the codomain can be realized as an output of the function for some choice of input from the domain.

The last fact reveals a special relationship between the graphs of f and f^{-1}. If a point (x, y) that lies on the graph of $y = f(x)$, then it is also true that $x = f^{-1}(y)$, which means that the point (y, x) lies on the graph of f^{-1}. This shows us that the graphs of f and f^{-1} are the reflections of each other across the line $y = x$, because this reflection is precisely the geometric action that swaps the coordinates in an ordered pair. In Figure 2.6.2, we see this illustrated by the function $y = f(x) = 2^x$ and its inverse, with the points $(-1, \frac{1}{2})$ and $(\frac{1}{2}, -1)$ highlighting the reflection of the curves across $y = x$.

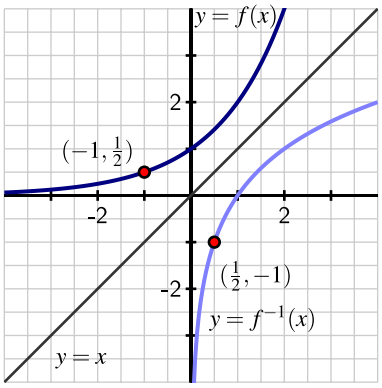


Figure 2.6.1. A graph of a function $y = f(x)$ along with its inverse, $y = f^{-1}(x)$.

To close our review of important facts about inverses, we recall that the natural exponential function $y = f(x) = e^x$ has an inverse function, namely the natural logarithm, $x = f^{-1}(y) = \ln(y)$. Thus, writing $y = e^x$ is interchangeable with $x = \ln(y)$, plus $\ln(e^x) = x$ for every real number x and $e^{\ln(y)} = y$ for every positive real number y.

The derivative of the natural logarithm function

In what follows, we find a formula for the derivative of $g(x) = \ln(x)$. To do so, we take advantage of the fact that we know the derivative of the natural exponential function, the inverse of g. In particular, we know that writing $g(x) = \ln(x)$ is equivalent to writing $e^{g(x)} = x$. Now we differentiate both sides of this equation and observe that

$$\frac{d}{dx}\left[e^{g(x)}\right] = \frac{d}{dx}[x].$$

The righthand side is simply 1; by applying the chain rule to the left side, we find that

$$e^{g(x)}g'(x) = 1.$$

Next we solve for $g'(x)$, to get

$$g'(x) = \frac{1}{e^{g(x)}}.$$

Finally, we recall that $g(x) = \ln(x)$, so $e^{g(x)} = e^{\ln(x)} = x$, and thus

$$g'(x) = \frac{1}{x}.$$

Natural Logarithm

For all positive real numbers x, $\frac{d}{dx}[\ln(x)] = \frac{1}{x}$.

This rule for the natural logarithm function now joins our list of basic derivative rules. Note that this rule applies only to positive values of x, as these are the only values for which $\ln(x)$ is defined.

Also notice that for the first time in our work, differentiating a basic function of a particular type has led to a function of a very different nature: the derivative of the natural logarithm is not another logarithm, nor even an exponential function, but rather a rational one.

Derivatives of logarithms may now be computed in concert with all of the rules known to date. For instance, if $f(t) = \ln(t^2 + 1)$, then by the chain rule, $f'(t) = \frac{1}{t^2+1} \cdot 2t$.

There are interesting connections between the graphs of $f(x) = e^x$ and $f^{-1}(x) = \ln(x)$.

In Figure 2.6.3, we are reminded that since the natural exponential function has the property that its derivative is itself, the slope of the tangent to $y = e^x$ is equal to the height of the curve at that point. For instance, at the point $A = (\ln(0.5), 0.5)$, the slope of the tangent line is $m_A = 0.5$, and at $B = (\ln(5), 5)$, the tangent line's slope is $m_B = 5$.

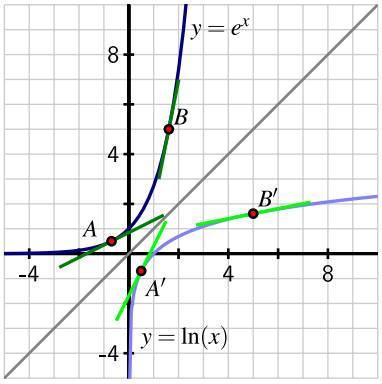


Figure 2.6.2. A graph of the function $y = e^x$ along with its inverse, $y = \ln(x)$, where both functions are viewed using the input variable x.

At the corresponding points A' and B' on the graph of the natural logarithm function (which come from reflecting A and B across the line $y = x$), we know that the slope of the tangent line is the reciprocal of the x-coordinate of the point (since $\frac{d}{dx}[\ln(x)] = \frac{1}{x}$). Thus, at $A' = (0.5, \ln(0.5))$, we have $m_{A'} = \frac{1}{0.5} = 2$, and at $B' = (5, \ln(5))$, $m_{B'} = \frac{1}{5}$.

In particular, we observe that $m_{A'} = \frac{1}{m_A}$ and $m_{B'} = \frac{1}{m_B}$. This is not a coincidence, but in fact holds for any curve $y = f(x)$ and its inverse, provided the inverse exists. This is due to the reflection across $y = x$. It changes the roles of x and y, thus reversing the rise and run, so the slope of the inverse function at the reflected point is the reciprocal of the slope of the original function.

Activity 2.6.2

For each function given below, find its derivative.

1. $h(x) = x^2 \ln(x)$
2. $p(t) = \dfrac{\ln(t)}{e^t + 1}$

3. $s(y) = \ln(\cos(y) + 2)$
4. $z(x) = \tan(\ln(x))$
5. $m(z) = \ln(\ln(z))$

Inverse trigonometric functions and their derivatives

Trigonometric functions are periodic, so they fail to be one-to-one, and thus do not have inverse functions. However, we can restrict the domain of each trigonometric function so that it is one-to-one on that domain.

For instance, consider the sine function on the domain $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Because no output of the sine function is repeated on this interval, the function is one-to-one and thus has an inverse. Thus, the function $f(x) = \sin(x)$ with $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and codomain $[-1, 1]$ has an inverse function f^{-1} such that

$$f^{-1} : [-1, 1] \to [-\frac{\pi}{2}, \frac{\pi}{2}].$$

We call f^{-1} the *arcsine* (or inverse sine) function and write $f^{-1}(y) = \arcsin(y)$. It is especially important to remember that

$$y = \sin(x) \quad \text{and} \quad x = \arcsin(y)$$

say the same thing. "The arcsine of y" means "the angle whose sine is y." For example, $\arcsin(\frac{1}{2}) = \frac{\pi}{6}$ means that $\frac{\pi}{6}$ is the angle whose sine is $\frac{1}{2}$, which is equivalent to writing $\sin(\frac{\pi}{6}) = \frac{1}{2}$.

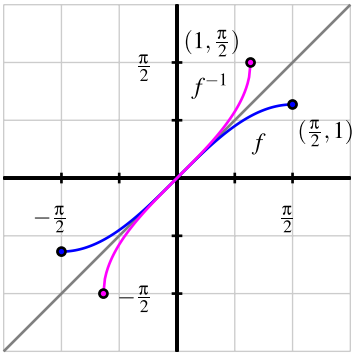


Figure 2.6.3. A graph of $f(x) = \sin(x)$ (in blue), restricted to the domain $[-\frac{\pi}{2}, \frac{\pi}{2}]$, along with its inverse, $f^{-1}(x) = \arcsin(x)$ (in magenta).

Next, we determine the derivative of the arcsine function. Letting $h(x) = \arcsin(x)$, our goal is to find $h'(x)$. Since $h(x)$ is the angle whose sine is x, it is equivalent to write

$$\sin(h(x)) = x.$$

Differentiating both sides of the previous equation, we have

$$\frac{d}{dx}[\sin(h(x))] = \frac{d}{dx}[x].$$

The righthand side is simply 1, and by applying the chain rule applied to the left side,

$$\cos(h(x))h'(x) = 1.$$

Solving for $h'(x)$, it follows that

$$h'(x) = \frac{1}{\cos(h(x))}.$$

Finally, we recall that $h(x) = \arcsin(x)$, so the denominator of $h'(x)$ is the function $\cos(\arcsin(x))$, or in other words, "the cosine of

the angle whose sine is x." A bit of right triangle trigonometry allows us to simplify this expression considerably.

Let's say that $\theta = \arcsin(x)$, so that θ is the angle whose sine is x. We can picture θ as an angle in a right triangle with hypotenuse 1 and a vertical leg of length x, as shown in Figure 2.6.5. The horizontal leg must be $\sqrt{1-x^2}$, by the Pythagorean Theorem.

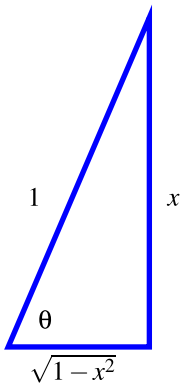


Figure 2.6.4. The right triangle that corresponds to the angle $\theta = \arcsin(x)$.

Now, because $\theta = \arcsin(x)$, the expression for $\cos(\arcsin(x))$ is equivalent to $\cos(\theta)$. From the figure, $\cos(\arcsin(x)) = \cos(\theta) = \sqrt{1-x^2}$.

Substituting this expression into our formula, $h'(x) = \frac{1}{\cos(\arcsin(x))}$, we have now shown that

$$h'(x) = \frac{1}{\sqrt{1-x^2}}.$$

Inverse sine

For all real numbers x such that $-1 < x < 1$,

$$\frac{d}{dx}[\arcsin(x)] = \frac{1}{\sqrt{1-x^2}}.$$

Activity 2.6.3

The following prompts in this activity will lead you to develop the derivative of the inverse tangent function.

1. Let $r(x) = \arctan(x)$. Use the relationship between the arctangent and tangent functions to rewrite this equation using only the tangent function.
2. Differentiate both sides of the equation you found in (a). Solve the resulting equation for $r'(x)$, writing $r'(x)$ as simply as possible in terms of a trigonometric function evaluated at $r(x)$.
3. Recall that $r(x) = \arctan(x)$. Update your expression for $r'(x)$ so that it only involves trigonometric functions and the independent variable x.
4. Introduce a right triangle with angle θ so that $\theta = \arctan(x)$. What are the three sides of the triangle?
5. In terms of only x and 1, what is the value of $\cos(\arctan(x))$?
6. Use the results of your work above to find an expression involving only 1 and x for $r'(x)$.

While derivatives for other inverse trigonometric functions can be established similarly, for now we limit ourselves to the arcsine and arctangent functions.

Activity 2.6.4

Determine the derivative of each of the following functions.

1. $f(x) = x^3 \arctan(x) + e^x \ln(x)$
2. $p(t) = 2^{t \arcsin(t)}$
3. $h(z) = (\arcsin(5z) + \arctan(4 - z))^{27}$
4. $s(y) = \cot(\arctan(y))$
5. $m(v) = \ln(\sin^2(v) + 1)$
6. $g(w) = \arctan\left(\dfrac{\ln(w)}{1 + w^2}\right)$

The link between the derivative of a function and the derivative of its inverse

In Figure 2.6.3, we saw an interesting relationship between the slopes of tangent lines to the natural exponential and natural logarithm functions at points reflected across the line $y = x$. In particular, we observed that at the point $(\ln(2), 2)$ on the graph of $f(x) = e^x$, the slope of the tangent line is $f'(\ln(2)) = 2$, while at the corresponding point $(2, \ln(2))$ on the graph of $f^{-1}(x) = \ln(x)$, the slope of the tangent line is $(f^{-1})'(2) = \frac{1}{2}$, which is the reciprocal of $f'(\ln(2))$.

That the two corresponding tangent lines have reciprocal slopes is not a coincidence. If f and g are differentiable inverse functions, then $y = f(x)$ if and only if $x = g(y)$, then$f(g(x)) = x$ for every x in the domain of f^{-1}. Differentiating both sides of this equation, we have

$$\frac{d}{dx}[f(g(x))] = \frac{d}{dx}[x],$$

and by the chain rule,

$$f'(g(x))g'(x) = 1.$$

Solving for $g'(x)$, we have $g'(x) = \frac{1}{f'(g(x))}$. Here we see that the slope of the tangent line to the inverse function g at the point $(x, g(x))$ is precisely the reciprocal of the slope of the tangent line to the original function f at the point $(g(x), f(g(x))) = (g(x), x)$.

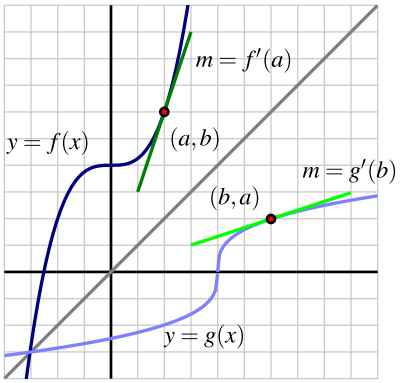


Figure 2.6.5. A graph of function $y = f(x)$ along with its inverse, $y = g(x) = f^{-1}(x)$. Observe that the slopes of the two tangent lines are reciprocals of one another.

To see this more clearly, consider the graph of the function $y = f(x)$ shown in Figure 2.6.6, along with its inverse $y = g(x)$. Given a point (a, b) that lies on the graph of f, we know that (b, a) lies on the graph of g; because $f(a) = b$ and $g(b) = a$. Now, applying the

rule that $g'(x) = 1/f'(g(x))$ to the value $x = b$, we have

$$g'(b) = \frac{1}{f'(g(b))} = \frac{1}{f'(a)},$$

which is precisely what we see in the figure: the slope of the tangent line to g at (b, a) is the reciprocal of the slope of the tangent line to f at (a, b), since these two lines are reflections of one another across the line $y = x$.

Derivative of an inverse function

Suppose that f is a differentiable function with inverse g and that (a, b) is a point that lies on the graph of f at which $f'(a) \neq 0$. Then

$$g'(b) = \frac{1}{f'(a)}.$$

More generally, for any x in the domain of g', we have $g'(x) = 1/f'(g(x))$.

The rules we derived for $\ln(x)$, $\arcsin(x)$, and $\arctan(x)$ are all just specific examples of this general property of the derivative of an inverse function. For example, with $g(x) = \ln(x)$ and $f(x) = e^x$, it follows that

$$g'(x) = \frac{1}{f'(g(x))} = \frac{1}{e^{\ln(x)}} = \frac{1}{x}.$$

Summary

- For all positive real numbers x, $\frac{d}{dx}[\ln(x)] = \frac{1}{x}$.
- For all real numbers x such that $-1 < x < 1$, $\frac{d}{dx}[\arcsin(x)] = \frac{1}{\sqrt{1-x^2}}$. In addition, for all real numbers x, $\frac{d}{dx}[\arctan(x)] = \frac{1}{1+x^2}$.
- If g is the inverse of a differentiable function f, then for any point x in the domain of g', $g'(x) = \frac{1}{f'(g(x))}$.

This page titled 2.6: Derivatives of Inverse Functions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.6: Derivatives of Inverse Functions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

2.7: Derivatives of Functions Given Implicitly

Motivating Questions

- What does it mean to say that a curve is an implicit function of x, rather than an explicit function of x?
- How does implicit differentiation enable us to find a formula for $\frac{dy}{dx}$ when y is an implicit function of x?
- In the context of an implicit curve, how can we use $\frac{dy}{dx}$ to answer important questions about the tangent line to the curve?

In all of our studies with derivatives so far, we have worked with functions whose formula is given explicitly in terms of x. But there are many interesting curves whose equations involving x and y are impossible to solve for y in terms of x.

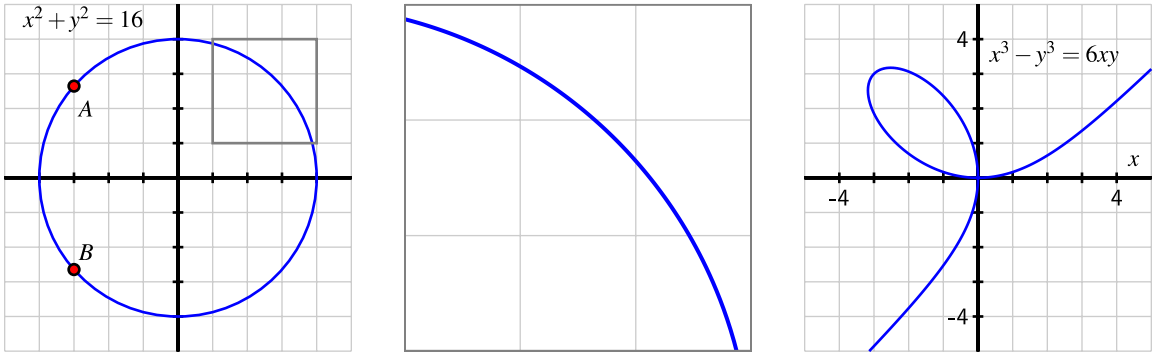


Figure 2.7.1 . At left, the circle given by $x^2 + y^2 = 16$. In the middle, the portion of the circle $x^2 + y^2 = 16$ that has been highlighted in the box at left. And at right, the curve given by $x^3 - y^3 = 6xy$.

Perhaps the simplest and most natural of all such curves are circles. Because of the circle's symmetry, for each x value strictly between the endpoints of the horizontal diameter, there are two corresponding y-values. For instance, in Figure 2.7.1 , we have labeled $A = (-3, \sqrt{7})$ and $B = (-3, -\sqrt{7})$, and these points demonstrate that the circle fails the vertical line test. Hence, it is impossible to represent the circle through a single function of the form $y = f(x)$. But portions of the circle can be represented explicitly as a function of x, such as the highlighted arc that is magnified in the center of Figure 2.7.1 Moreover, it is evident that the circle is locally linear, so we ought to be able to find a tangent line to the curve at every point. Thus, it makes sense to wonder if we can compute $\frac{dy}{dx}$ at any point on the circle, even though we cannot write y explicitly as a function of x.

We say that the equation $x^2 + y^2 = 16$ defines y *implicitly* as a function of x. The graph of the equation can be broken into pieces where each piece can be defined by an explicit function of x. For the circle, we could choose to take the top half as one function of x, namely $y = \sqrt{16 - x^2}$ and the bottom half as $y = -\sqrt{16 - x^2}$. The equation for the circle defines two *implicit functions* of x.

The righthand curve in Figure 2.7.1 is called a *lemniscate* and is just one of many fascinating possibilities for implicitly given curves.

How can we find an equation for $\frac{dy}{dx}$ without an explicit formula for y in terms of x? The following preview activity reminds us of some ways we can compute derivatives of functions in settings where the function's formula is not known.

Preview Activity 2.7.1

Let f be a differentiable function of x (whose formula is not known) and recall that $\frac{d}{dx}[f(x)]$ and $f'(x)$ are interchangeable notations. Determine each of the following derivatives of combinations of explicit functions of x, the unknown function f, and an arbitrary constant c.

1. $\frac{d}{dx}\left[x^2+f(x)\right]$
2. $\frac{d}{dx}\left[x^2 f(x)\right]$
3. $\frac{d}{dx}\left[c+x+f(x)^2\right]$
4. $\frac{d}{dx}\left[f(x^2)\right]$
5. $\frac{d}{dx}\left[xf(x)+f(cx)+cf(x)\right]$

Implicit Differentiation

We begin our exploration of implicit differentiation with the example of the circle given by $x^2+y^2=16$. How can we find a formula for $\frac{dy}{dx}$?

By viewing y as an *implicit* function of x, we think of y as some function whose formula $f(x)$ is unknown, but which we can differentiate. Just as y represents an unknown formula, so too its derivative with respect to x, $\frac{dy}{dx}$, will be (at least temporarily) unknown.

So we view y as an unknown differentiable function of x and differentiate both sides of the equation with respect to x.

$$\frac{d}{dx}\left[x^2+y^2\right]=\frac{d}{dx}\left[16\right].$$

On the right, the derivative of the constant 16 is 0, and on the left we can apply the sum rule, so it follows that

$$\frac{d}{dx}\left[x^2\right]+\frac{d}{dx}\left[y^2\right]=0.$$

Note carefully the different roles being played by x and y. Because x is the independent variable, $\frac{d}{dx}\left[x^2\right]=2x$. But y is the dependent variable and y is an implicit function of x. Recall Preview Activity $2.7.1$, where we computed $\frac{d}{dx}[f(x)^2]$. Computing $\frac{d}{dx}[y^2]$ is the same, and requires the chain rule, by which we find that $\frac{d}{dx}[y^2]=2y^1\frac{dy}{dx}$. We now have that

$$2x+2y\frac{dy}{dx}=0.$$

We solve this equation for $\frac{dy}{dx}$ by subtracting $2x$ from both sides and dividing by $2y$.

$$\frac{dy}{dx}=-\frac{2x}{2y}=-\frac{x}{y}.$$

There are several important things to observe about the result that $\frac{dy}{dx}=-\frac{x}{y}$. First, this expression for the derivative involves both x and y. This makes sense because there are two corresponding points on the circle for each value of x between -4 and 4, and the slope of the tangent line is different at each of these points.

Second, this formula is entirely consistent with our understanding of circles. The slope of the radius from the origin to the point (a,b) is $m_r=\frac{b}{a}$. The tangent line to the circle at (a,b) is perpendicular to the radius, and thus has slope $m_t=-\frac{a}{b}$, as shown in Figure $2.7.2$. In particular, the slope of the tangent line is zero at $(0,4)$ and $(0,-4)$, and is undefined at $(-4,0)$ and $(4,0)$. All of these values are consistent with the formula $\frac{dy}{dx}=-\frac{x}{y}$.

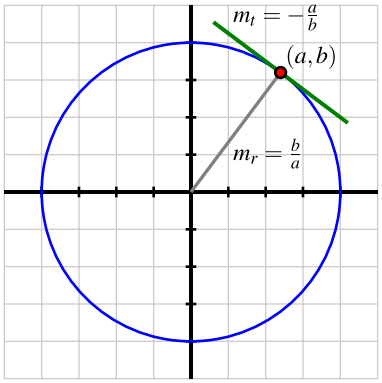


Figure 2.7.2 . The circle given by $x^2 + y^2 = 16$ with point (a, b) on the circle and the tangent line at that point, with labeled slopes of the radial line, m_r, and tangent line, m_t.

Example 2.7.1

For the curve given implicitly by $x^3 + y^2 - 2xy = 2$, shown in Figure 2.7.3 , find the slope of the tangent line at $(-1, 1)$.

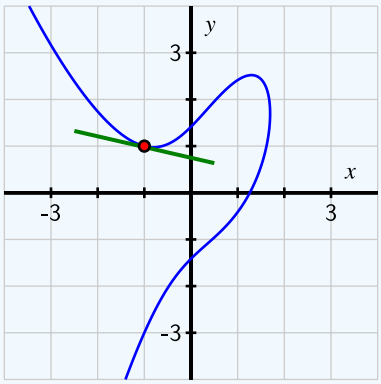


Figure 2.7.3 . The curve $x^3 + y^2 - 2xy = 2$.

Answer available at the end of this section.

Example 2.7.3 shows that it is possible when differentiating implicitly to have multiple terms involving $\frac{dy}{dx}$. We use addition and subtraction to collect all terms involving $\frac{dy}{dx}$ on one side of the equation, then factor to get a single term of $\frac{dy}{dx}$. Finally, we divide to solve for $\frac{dy}{dx}$.

We use the notation

$$\left.\frac{dy}{dx}\right|_{(a,b)}$$

to denote the evaluation of $\frac{dy}{dx}$ at the point (a, b). This is analogous to writing $f'(a)$ when f' depends on a single variable.

There is a big difference between writing $\frac{d}{dx}$ and $\frac{dy}{dx}$. For example,

$$\frac{d}{dx}[x^2 + y^2]$$

gives an instruction to take the derivative with respect to x of the quantity $x^2 + y^2$, presumably where y is a function of x. On the other hand,

$$\frac{dy}{dx}(x^2+y^2)$$

means the product of the derivative of y with respect to x with the quantity x^2+y^2. Understanding this notational subtlety is essential.

Activity 2.7.1

Consider the curve defined by the equation $x = y^5 - 5y^3 + 4y$, whose graph is pictured in Figure 2.7.4 .

1. Explain why it is not possible to express y as an explicit function of x.
2. Use implicit differentiation to find a formula for dy/dx.
3. Use your result from part (b) to find an equation of the line tangent to the graph of $x = y^5 - 5y^3 + 4y$ at the point $(0, 1)$.
4. Use your result from part (b) to determine all of the points at which the graph of $x = y^5 - 5y^3 + 4y$ has a vertical tangent line.

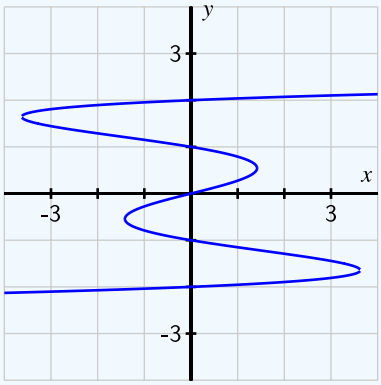


Figure 2.7.4 . The curve $x = y^5 - 5y^3 + 4y$.

It is natural to ask where the tangent line to a curve is vertical or horizontal. The slope of a horizontal tangent line must be zero, while the slope of a vertical tangent line is undefined. Often the formula for $\frac{dy}{dx}$ is expressed as a quotient of functions of x and y, say

$$\frac{dy}{dx} = \frac{p(x,y)}{q(x,y)}.$$

The tangent line is horizontal precisely when the numerator is zero and the denominator is nonzero, making the slope of the tangent line zero. If we can solve the equation $p(x, y) = 0$ for either x and y in terms of the other, we can substitute that expression into the original equation for the curve. This gives an equation in a single variable, and if we can solve that equation we can find the point(s) on the curve where $p(x, y) = 0$. At those points, the tangent line is horizontal.

Similarly, the tangent line is vertical whenever $q(x, y) = 0$ and $p(x, y) \neq 0$, making the slope undefined.

Activity 2.7.2

Consider the curve defined by the equation $y(y^2 - 1)(y - 2) = x(x - 1)(x - 2)$, whose graph is pictured in Figure 2.7.5 . Through implicit differentiation, it can be shown that

$$\frac{dy}{dx} = \frac{(x-1)(x-2)+x(x-2)+x(x-1)}{(y^2-1)(y-2)+2y^2(y-2)+y(y^2-1)}.$$

Use this fact to answer each of the following questions.

1. Determine all points (x, y) at which the tangent line to the curve is horizontal. (Use technology appropriately to find the needed zeros of the relevant polynomial function.)
2. Determine all points (x, y) at which the tangent line is vertical. (Use technology appropriately to find the needed zeros of the relevant polynomial function.)
3. Find the equation of the tangent line to the curve at one of the points where $x = 1$.

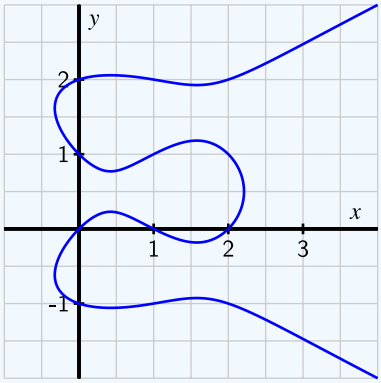


Figure 2.7.5 . $y(y^2 - 1)(y - 2) = x(x - 1)(x - 2)$.

Activity 2.7.3

For each of the following curves, use implicit differentiation to find dy/dx and determine the equation of the tangent line at the given point.

1. $x^3 - y^3 = 6xy$, $(-3, 3)$
2. $\sin(y) + y = x^3 + x$, $(0, 0)$
3. $3xe^{-xy} = y^2$, $(0.619061, 1)$

Summary

- In an equation involving x and y where portions of the graph can be defined by explicit functions of x, we say that y is an implicit function of x. A good example of such a curve is the unit circle.
- We use implicit differentiation to differentiate an implicitly defined function. We differentiate both sides of the equation with respect to x, treating y as a function of x by applying the chain rule. If possible, we subsequently solve for $\frac{dy}{dx}$ using algebra.
- While $\frac{dy}{dx}$ may now involve both the variables x and y, $\frac{dy}{dx}$ still gives the slope of the tangent line to the curve. It may be used to decide where the tangent line is horizontal ($\frac{dy}{dx} = 0$) or vertical ($\frac{dy}{dx}$ is undefined), or to find the equation of the tangent line at a particular point on the curve.

Answers

Example 2.7.1

We begin by differentiating the curve's equation implicitly. Taking the derivative of each side with respect to x,

$$\frac{d}{dx}\left[x^3 + y^2 - 2xy\right] = \frac{d}{dx}\left[2\right],$$

by the sum rule and the fact that the derivative of a constant is zero, we have

$$\frac{d}{dx}[x^3] + \frac{d}{dx}[y^2] - \frac{d}{dx}[2xy] = 0.$$

For the three derivatives we now must execute, the first uses the simple power rule, the second requires the chain rule (since y is an implicit function of x), and the third necessitates the product rule (again since y is a function of x). Applying these rules, we now find that

$$3x^2 + 2y\frac{dy}{dx} - [2x\frac{dy}{dx} + 2y] = 0.$$

We want to solve this equation for $\frac{dy}{dx}$. To do so, we first collect all of the terms involving $\frac{dy}{dx}$ on one side of the equation.

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

Then we factor the left side to isolate $\frac{dy}{dx}$.

$$\frac{dy}{dx}(2y - 2x) = 2y - 3x^2.$$

Finally, we divide both sides by $(2y - 2x)$ and conclude that

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

Note that the expression for $\frac{dy}{dx}$ depends on both x and y. To find the slope of the tangent line at $(-1, 1)$, we substitute the coordinates into the formula for $\frac{dy}{dx}$, using the notation

$$\left.\frac{dy}{dx}\right|_{(-1,1)} = \frac{2(1) - 3(-1)^2}{2(1) - 2(-1)} = -\frac{1}{4}.$$

This value matches our visual estimate of the slope of the tangent line shown in Figure 2.7.3 .

This page titled 2.7: Derivatives of Functions Given Implicitly is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.7: Derivatives of Functions Given Implicitly by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

2.8: Using Derivatives to Evaluate Limits

Motivating Questions

- How can derivatives be used to help us evaluate indeterminate limits of the form $\frac{0}{0}$?
- What does it mean to say that $\lim_{x\to\infty} f(x) = L$ and $\lim_{x\to a} f(x) = \infty$?
- How can derivatives assist us in evaluating indeterminate limits of the form $\frac{\infty}{\infty}$?

Because differential calculus is based on the definition of the derivative, and the definition of the derivative involves a limit, there is a sense in which all of calculus rests on limits. In addition, the limit involved in the definition of the derivative always generates the indeterminate form $\frac{0}{0}$. If f is a differentiable function, then in the definition

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

not only does $h \to 0$ in the denominator, but also $(f(x+h) - f(x)) \to 0$ in the numerator, since f is continuous. Remember, saying that a limit has an indeterminate form only means that we don't yet know its value and have more work to do: indeed, limits of the form $\frac{0}{0}$ can take on any value, as is evidenced by evaluating $f'(x)$ for varying values of x for a function such as $f'(x) = x^2$.

We have learned many techniques for evaluating the limits that result from the derivative definition, including a large number of shortcut rules. In this section, we turn the situation upside-down: instead of using limits to evaluate derivatives, we explore how to use derivatives to evaluate certain limits.

Preview Activity 2.8.1

Let h be the function given by $h(x) = \frac{x^5+x-2}{x^2-1}$.

1. What is the domain of h?
2. Explain why $\lim_{x\to 1} \dfrac{x^5+x-2}{x^2-1}$ results in an indeterminate form.
3. Next we will investigate the behavior of both the numerator and denominator of h near the point where $x = 1$. Let $f(x) = x^5 + x - 2$ and $g(x) = x^2 - 1$. Find the local linearizations of f and g at $a = 1$, and call these functions $L_f(x)$ and $L_g(x)$, respectively.
4. Explain why $h(x) \approx \frac{L_f(x)}{L_g(x)}$ for x near $a = 1$.
5. Using your work from (c) and (d), evaluate

   $$\lim_{x\to 1} \frac{L_f(x)}{L_g(x)}.$$

   What do you think your result tells us about $\lim_{x\to 1} h(x)$?
6. Investigate the function $h(x)$ graphically and numerically near $x = 1$. What do you think is the value of $\lim_{x\to 1} h(x)$?

2.8.1 Using derivatives to evaluate indeterminate limits of the form $\frac{0}{0}$.

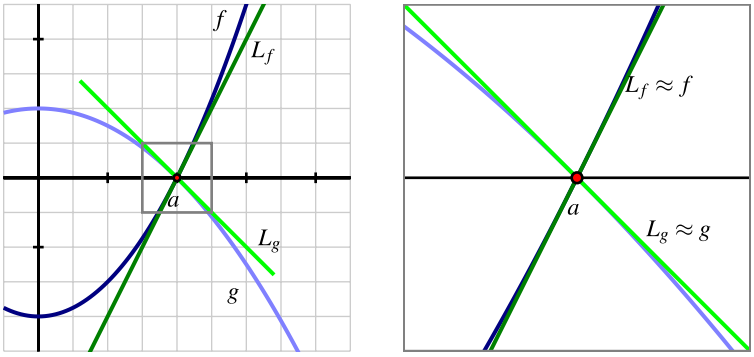


Figure 2.8.1. At left, the graphs of f and g near the value a, along with their tangent line approximations L_f and L_g at $x = a$. At right, zooming in on the point a and the four graphs.

The idea demonstrated in Preview Activity 2.8.1 — that we can evaluate an indeterminate limit of the form $\frac{0}{0}$ by replacing each of the numerator and denominator with their local linearizations at the point of interest — can be generalized in a way that enables us to evaluate a wide range of limits. We have a function $h(x)$ that can be written as a quotient $h(x) = \frac{f(x)}{g(x)}$, where f and g are both differentiable at $x = a$ and for which $f(a) = g(a) = 0$. We would like to evaluate the indeterminate limit given by $\lim_{x\to a} h(x)$. Figure 2.8.1 illustrates the situation. We see that both f and g have an x-intercept at $x = a$. Their respective tangent line approximations L_f and L_g at $x = a$ are also shown in the figure. We can take advantage of the fact that a function and its tangent line approximation become indistinguishable as $x \to a$.

First, let's recall that $L_f(x) = f'(a)(x - a) + f(a)$ and $L_g(x) = g'(a)(x - a) + g(a)$. Because x is getting arbitrarily close to a when we take the limit, we can replace f with L_f and replace g with L_g, and thus we observe that

$$\begin{aligned}\lim_{x\to a} \frac{f(x)}{g(x)} &= \lim_{x\to a} \frac{L_f(x)}{L_g(x)} \\ &= \lim_{x\to a} \frac{f'(a)(x-a) + f(a)}{g'(a)(x-a) + g(a)}.\end{aligned}$$

Next, we remember that both $f(a) = 0$ and $g(a) = 0$, which is precisely what makes the original limit indeterminate. Substituting these values for $f(a)$ and $g(a)$ in the limit above, we now have

$$\begin{aligned}\lim_{x\to a} \frac{f(x)}{g(x)} &= \lim_{x\to a} \frac{f'(a)(x-a)}{g'(a)(x-a)} \\ &= \lim_{x\to a} \frac{f'(a)}{g'(a)},\end{aligned}$$

where the latter equality holds because $\frac{x-a}{x-a} = 1$ when x is approaching (but not equal to) a. Finally, we note that $\frac{f'(a)}{g'(a)}$ is constant with respect to x, and thus

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)}.$$

This result holds as long as $g'(a)$ is not equal to zero. The formal name of the result is L'Hôpital's Rule.

L'Hôpital's Rule

Let f and g be differentiable at $x = a$, and suppose that $f(a) = g(a) = 0$ and that $g'(a) \neq 0$. Then $\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)}$.

In practice, we typically work with a slightly more general version of L'Hôpital's Rule, which states that (under the identical assumptions as the boxed rule above and the extra assumption that g' is continuous at $x = a$)

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)},$$

provided the righthand limit exists. This form reflects the basic idea of L'Hôpital's Rule: if $\frac{f(x)}{g(x)}$ produces an indeterminate limit of form $\frac{0}{0}$ as $x \to a$, that limit is equivalent to the limit of the quotient of the two functions' derivatives, $\frac{f'(x)}{g'(x)}$.

For example, if we consider the limit from Preview Activity 2.8.1 ,

$$\lim_{x\to 1} \frac{x^5 + x - 2}{x^2 - 1},$$

by L'Hôpital's Rule we have that

$$\lim_{x\to 1} \frac{x^5 + x - 2}{x^2 - 1} = \lim_{x\to 1} \frac{5x^4 + 1}{2x} = \frac{6}{2} = 3.$$

By replacing the numerator and denominator with their respective derivatives, we often replace an indeterminate limit with one whose value we can easily determine.

Activity 2.8.1

Evaluate each of the following limits. If you use L'Hôpital's Rule, indicate where it was used, and be certain its hypotheses are met before you apply it.

1. $\lim_{x\to 0} \frac{\ln(1+x)}{x}$
2. $\lim_{x\to \pi} \frac{\cos(x)}{x}$
3. $\lim_{x\to 1} \frac{2\ln(x)}{1 - e^{x-1}}$
4. $\lim_{x\to 0} \frac{\sin(x) - x}{\cos(2x) - 1}$

While L'Hôpital's Rule can be applied in an entirely algebraic way, it is important to remember that the justification of the rule is graphical: the main idea is that the slopes of the tangent lines to f and g at $x = a$ determine the value of the limit of $\frac{f(x)}{g(x)}$ as $x \to a$.

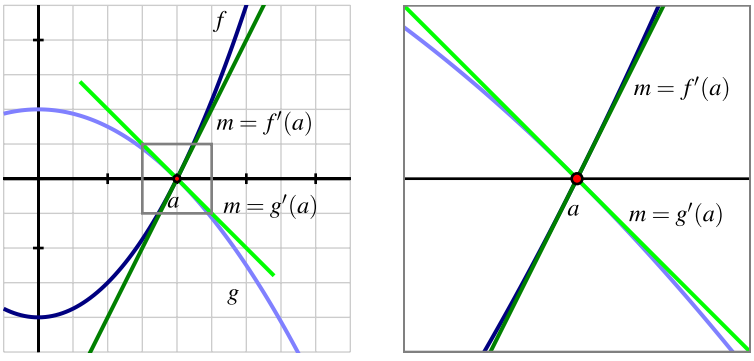


Figure 2.8.2. Two functions f and g that satisfy L'Hôpital's Rule.

We see this in Figure 2.8.2, where we can see from the grid that $f'(a) = 2$ and $g'(a) = -1$, hence by L'Hôpital's Rule,

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\frac{f'(a)}{g'(a)}=\frac{2}{-1}=-2.$$

It's not the fact that f and g both approach zero that matters most, but rather the *rate* at which each approaches zero that determines the value of the limit. This is a good way to remember what L'Hôpital's Rule says: if $f(a)=g(a)=0$, the the limit of $\frac{f(x)}{g(x)}$ as $x\to a$ is given by the ratio of the slopes of f and g at $x=a$.

Activity 2.8.2

In this activity, we reason graphically from the following figure to evaluate limits of ratios of functions about which some information is known.

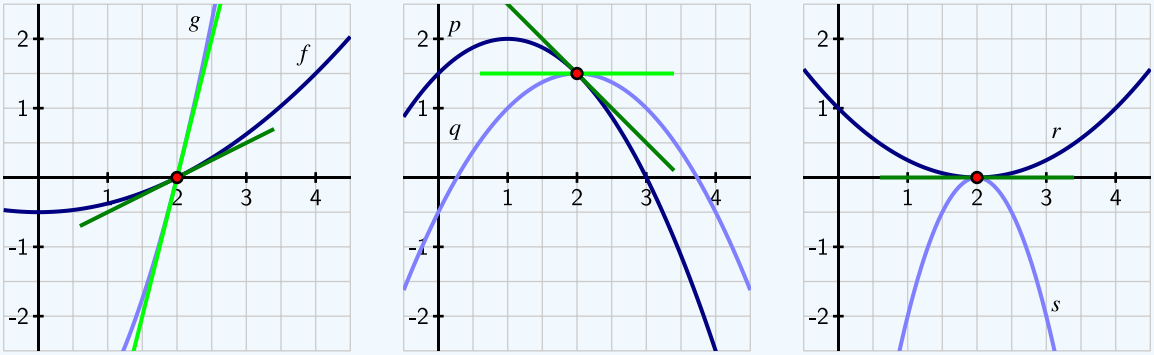


Figure 2.8.3. Three graphs referenced in the questions of Activity 2.8.3.

1. Use the left-hand graph to determine the values of $f(2)$, $f'(2)$, $g(2)$, and $g'(2)$. Then, evaluate $\lim_{x\to 2}\frac{f(x)}{g(x)}$.
2. Use the middle graph to find $p(2)$, $p'(2)$, $q(2)$, and $q'(2)$. Then, determine the value of $\lim_{x\to 2}\frac{p(x)}{q(x)}$.
3. Assume that r and s are functions whose for which $r''(2)\neq 0$ and $s''(2)\neq 0$ Use the right-hand graph to compute $r(2)$, $r'(2)$, $s(2)$, $s'(2)$. Explain why you cannot determine the exact value of $\lim_{x\to 2}\frac{r(x)}{s(x)}$ without further information being provided, but that you can determine the sign of $\lim_{x\to 2}\frac{r(x)}{s(x)}$. In addition, state what the sign of the limit will be, with justification.

2.8.2 Limits involving ∞

The concept of infinity, denoted ∞, arises naturally in calculus, as it does in much of mathematics. It is important to note from the outset that ∞ is a concept, but not a number itself. Indeed, the notion of ∞ naturally invokes the idea of limits. Consider, for example, the function $f(x)=\frac{1}{x}$, whose graph is pictured in Figure 2.8.4.

We note that $x=0$ is not in the domain of f, so we may naturally wonder what happens as $x\to 0$. As $x\to 0^+$, we observe that $f(x)$ *increases without bound*. That is, we can make the value of $f(x)$ as large as we like by taking x closer and closer (but not equal) to 0, while keeping $x>0$. This is a good way to think about what infinity represents: a quantity is tending to infinity if there is no single number that the quantity is always less than.

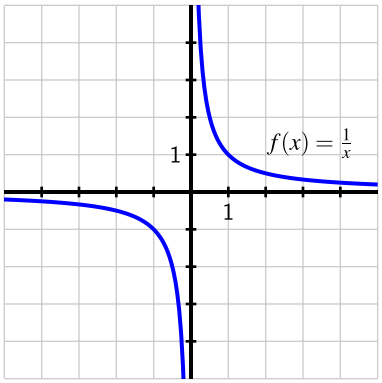


Figure 2.8.1. The graph of $f(x) = \frac{1}{x}$.

Recall that the statement $\lim_{x\to a} f(x) = L$, means that we can make $f(x)$ as close to L as we'd like by taking x sufficiently close (but not equal) to a. We now expand this notation and language to include the possibility that either L or a can be ∞. For instance, for $f(x) = \frac{1}{x}$, we now write

$$\lim_{x\to 0^+} \frac{1}{x} = \infty,$$

by which we mean that we can make $\frac{1}{x}$ as large as we like by taking x sufficiently close (but not equal) to 0. In a similar way, we write

$$\lim_{x\to \infty} \frac{1}{x} = 0,$$

since we can make $\frac{1}{x}$ as close to 0 as we'd like by taking x sufficiently large (i.e., by letting x increase without bound).

In general, the notation $\lim_{x\to a} f(x) = \infty$ means that we can make $f(x)$ as large as we like by taking x sufficiently close (but not equal) to a, and the notation $\lim_{x\to \infty} f(x) = L$ means that we can make $f(x)$ as close to L as we like by taking x sufficiently large. This notation also applies to left- and right-hand limits, and to limits involving $-\infty$. For example, returning to Figure 2.8.4 and $f(x) = \frac{1}{x}$, we can say that

$$\lim_{x\to 0^-} \frac{1}{x} = -\infty \quad \text{and} \quad \lim_{x\to -\infty} \frac{1}{x} = 0.$$

Finally, we write

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

if we can make the value of $f(x)$ as large as we'd like by taking x sufficiently large. For example,

$$\lim_{x\to \infty} x^2 = \infty.$$

Limits involving infinity identify *vertical* and *horizontal asymptotes* of a function. If $\lim_{x\to a} f(x) = \infty$, then $x = a$ is a vertical asymptote of f, while if $\lim_{x\to \infty} f(x) = L$, then $y = L$ is a horizontal asymptote of f. Similar statements can be made using $-\infty$, and with left- and right-hand limits as $x \to a^-$ or $x \to a^+$.

In precalculus classes, it is common to study the *end behavior* of certain families of functions, by which we mean the behavior of a function as $x \to \infty$ and as $x \to -\infty$. Here we briefly examine some familiar functions and note the values of several limits involving ∞.

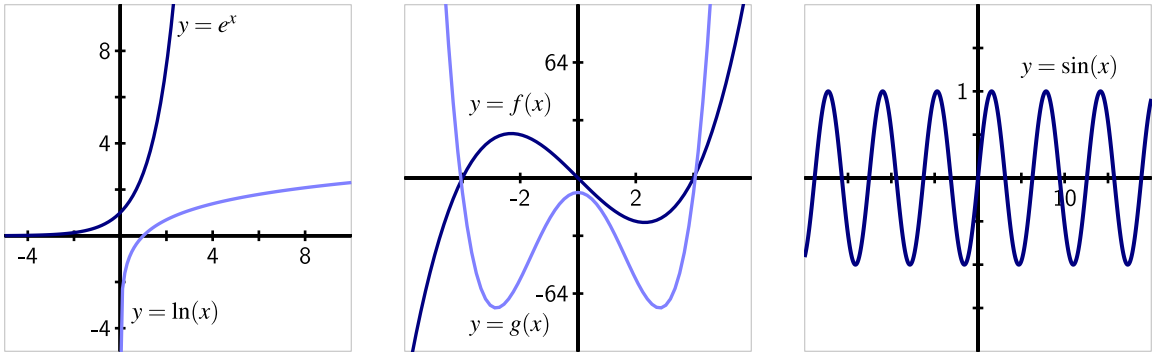


Figure 2.8.5. Graphs of some familiar functions whose end behavior as $x \to \pm\infty$ is known. In the middle graph, $f(x) = x^3 - 16x$ and $g(x) = x^4 - 16x^2 - 8$.

For the natural exponential function e^x, we note that $\lim_{x\to\infty} e^x = \infty$ and $\lim_{x\to-\infty} e^x = 0$. For the exponential decay function e^{-x}, these limits are reversed, with $\lim_{x\to\infty} e^{-x} = 0$ and $\lim_{x\to-\infty} e^{-x} = \infty$. Turning to the natural logarithm function, we have $\lim_{x\to0^+} \ln(x) = -\infty$ and $\lim_{x\to\infty} \ln(x) = \infty$. While both e^x and $\ln(x)$ grow without bound as $x \to \infty$, the exponential function does so much more quickly than the logarithm function does. We'll soon use limits to quantify what we mean by "quickly."

For polynomial functions of the form

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots a_1 x + a_0,$$

the end behavior depends on the sign of a_n and whether the highest power n is even or odd. If n is even and a_n is positive, then $\lim_{x\to\infty} p(x) = \infty$ and $\lim_{x\to-\infty} p(x) = \infty$, as in the plot of g in Figure 2.8.5. If instead a_n is negative, then $\lim_{x\to\infty} p(x) = -\infty$ and $\lim_{x\to-\infty} p(x) = -\infty$. In the situation where n is odd, then either $\lim_{x\to\infty} p(x) = \infty$ and $\lim_{x\to-\infty} p(x) = -\infty$ (which occurs when a_n is positive, as in the graph of f in Figure 2.8.5), or $\lim_{x\to\infty} p(x) = -\infty$ and $\lim_{x\to-\infty} p(x) = \infty$ (when a_n is negative).

A function can fail to have a limit as $x \to \infty$. For example, consider the plot of the sine function at right in Figure 2.8.5. Because the function continues oscillating between -1 and 1 as $x \to \infty$, we say that $\lim_{x\to\infty} \sin(x)$ does not exist.

Finally, it is straightforward to analyze the behavior of any rational function as $x \to \infty$.

Example 2.8.1

Determine the limit of the function

$$q(x) = \frac{3x^2 - 4x + 5}{7x^2 + 9x - 10}$$

as $x \to \infty$.

Note that both $(3x^2 - 4x + 5) \to \infty$ as $x \to \infty$ and $(7x^2 + 9x - 10) \to \infty$ as $x \to \infty$. Here we say that $\lim_{x\to\infty} q(x)$ has indeterminate form $\frac{\infty}{\infty}$. We can determine the value of this limit through a standard algebraic approach. Multiplying the numerator and denominator each by $\frac{1}{x^2}$, we find that

$$\lim_{x\to\infty} q(x) = \lim_{x\to\infty} \frac{3x^2 - 4x + 5}{7x^2 + 9x - 10} \cdot \frac{\frac{1}{x^2}}{\frac{1}{x^2}}$$

$$= \lim_{x\to\infty} \frac{3 - 4\frac{1}{x} + 5\frac{1}{x^2}}{7 + 9\frac{1}{x} - 10\frac{1}{x^2}} = \frac{3}{7}$$

since $\frac{1}{x^2} \to 0$ and $\frac{1}{x} \to 0$ as $x \to \infty$. This shows that the rational function q has a horizontal asymptote at $y = \frac{3}{7}$. A similar approach can be used to determine the limit of any rational function as $x \to \infty$.

But how should we handle a limit such as

$$\lim_{x\to\infty} \frac{x^2}{e^x}?$$

Here, both $x^2 \to \infty$ and $e^x \to \infty$, but there is not an obvious algebraic approach that enables us to find the limit's value. Fortunately, it turns out that L'Hôpital's Rule extends to cases involving infinity.

L'Hôpital's Rule (∞)

If f and g are differentiable and both approach zero or both approach $\pm\infty$ as $x \to a$ (where a is allowed to be ∞) , then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)},$$

provided the righthand limit exists.

(To be technically correct, we need to add the additional hypothesis that $g'(x) \neq 0$ on an open interval that contains a or in every neighborhood of infinity if a is ∞; this is almost always met in practice.)

To evaluate $\lim_{x\to\infty} \frac{x^2}{e^x}$, we can apply L'Hôpital's Rule, since both $x^2 \to \infty$ and $e^x \to \infty$. Doing so, it follows that

$$\lim_{x\to\infty} \frac{x^2}{e^x} = \lim_{x\to\infty} \frac{2x}{e^x}.$$

This updated limit is still indeterminate and of the form $\frac{\infty}{\infty}$, but it is simpler since $2x$ has replaced x^2. Hence, we can apply L'Hôpital's Rule again, and find that

$$\lim_{x\to\infty} \frac{x^2}{e^x} = \lim_{x\to\infty} \frac{2x}{e^x} = \lim_{x\to\infty} \frac{2}{e^x}.$$

Now, since 2 is constant and $e^x \to \infty$ as $x \to \infty$, it follows that $\frac{2}{e^x} \to 0$ as $x \to \infty$, which shows that

$$\lim_{x\to\infty} \frac{x^2}{e^x} = 0.$$

Activity 2.8.3

Evaluate each of the following limits. If you use L'Hôpital's Rule, indicate where it was used, and be certain its hypotheses are met before you apply it.

1. $\lim_{x\to\infty} \frac{x}{\ln(x)}$
2. $\lim_{x\to\infty} \frac{e^x + x}{2e^x + x^2}$
3. $\lim_{x\to 0^+} \frac{\ln(x)}{\frac{1}{x}}$
4. $\lim_{x\to \frac{\pi}{2}^-} \frac{\tan(x)}{x - \frac{\pi}{2}}$
5. $\lim_{x\to\infty} xe^{-x}$

To evaluate the limit of a quotient of two functions $\frac{f(x)}{g(x)}$ that results in an indeterminate form of $\frac{\infty}{\infty}$, in essence we are asking which function is growing faster without bound. We say that the function g *dominates* the function f as $x \to \infty$ provided that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0,$$

whereas f dominates g provided that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = \infty$. Finally, if the value of $\lim_{x\to\infty} \frac{f(x)}{g(x)}$ is finite and nonzero, we say that f and

g grow at the same rate. For example, we saw that $\lim_{x\to\infty} \frac{x^2}{e^x} = 0$, so e^x dominates x^2, while $\lim_{x\to\infty} \frac{3x^2-4x+5}{7x^2+9x-10} = \frac{3}{7}$, so $f(x) = 3x^2 - 4x + 5$ and $g(x) = 7x^2 + 9x - 10$ grow at the same rate.

Summary

- Derivatives can be used to help us evaluate indeterminate limits of the form $\frac{0}{0}$ through L'Hôpital's Rule, by replacing the functions in the numerator and denominator with their tangent line approximations. In particular, if $f(a) = g(a) = 0$ and f and g are differentiable at a, L'Hôpital's Rule tells us that
$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$
- When we write $x \to \infty$, this means that x is increasing without bound. Thus, $\lim_{x\to\infty} f(x) = L$ means that we can make $f(x)$ as close to L as we like by choosing x to be sufficiently large. Similarly, $\lim_{x\to a} f(x) = \infty$, means that we can make $f(x)$ as large as we like by choosing x sufficiently close to a.
- A version of L'Hôpital's Rule also helps us evaluate indeterminate limits of the form $\frac{\infty}{\infty}$. If f and g are differentiable and both approach zero or both approach $\pm\infty$ as $x \to a$ (where a is allowed to be ∞), then
$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

This page titled 2.8: Using Derivatives to Evaluate Limits is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.8: Using Derivatives to Evaluate Limits by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

2.9: Computing Derivatives (Exercises)

2.1: Elementary Derivative Rules

1. Derivative of a power function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

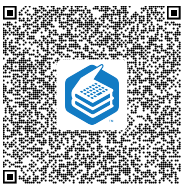
View this Interactive Content online



2. Derivative of a rational function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Derivative of a root function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Derivative of a quadratic

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5. Derivative of a sum of power functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



6. Simplifying a product before differentiating

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

7. Simplifying a quotient before differentiating

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

8. Finding a tangent line equation

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

9. Determining where $f'(x) = 0$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

10

Let f and g be differentiable functions for which the following information is known: $f(2) = 5$, $g(2) = -3$, $f'(2) = -1/2$, $g'(2) = 2$.

1. Let h be the new function defined by the rule $h(x) = 3f(x) - 4g(x)$. Determine $h(2)$ and $h'(2)$.
2. Find an equation for the tangent line to $y = h(x)$ at the point $(2, h(2))$.
3. Let p be the function defined by the rule $p(x) = -2f(x) + \frac{1}{2}g(x)$. Is p increasing, decreasing, or neither at $a = 2$? Why?
4. Estimate the value of $p(2.03)$ by using the local linearization of p at the point $(2, p(2))$.

11

Let functions p and q be the piecewise linear functions given by their respective graphs in Figure 2.1.6. Use the graphs to answer the following questions.

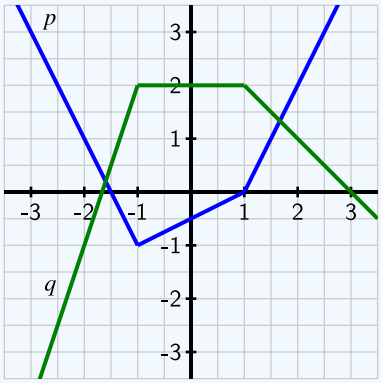


Figure 2.1.6. The graphs of p (in blue) and q (in green).

1. At what values of x is p not differentiable? At what values of x is q not differentiable? Why?
2. Let $r(x) = p(x) + 2q(x)$. At what values of x is r not differentiable? Why?
3. Determine $r'(-2)$ and $r'(0)$.
4. Find an equation for the tangent line to $y = r(x)$ at the point $(2, r(2))$.

12

Consider the functions $r(t) = t^t$ and $s(t) = \arccos(t)$, for which you are given the facts that $r'(t) = t^t(\ln(t) + 1)$ and $s'(t) = -\frac{1}{\sqrt{1-t^2}}$. Do not be concerned with where these derivative formulas come from. We restrict our interest in both functions to the domain $0 < t < 1$.

1. Let $w(t) = 3t^t - 2\arccos(t)$. Determine $w'(t)$.
2. Find an equation for the tangent line to $y = w(t)$ at the point $(\frac{1}{2}, w(\frac{1}{2}))$.
3. Let $v(t) = t^t + \arccos(t)$. Is v increasing or decreasing at the instant $t = \frac{1}{2}$? Why?

13

Let $f(x) = a^x$. The goal of this problem is to explore how the value of a affects the derivative of $f(x)$, without assuming we know the rule for $\frac{d}{dx}[a^x]$ that we have stated and used in earlier work in this section.

1. Use the limit definition of the derivative to show that
$$f'(x) = \lim_{h \to 0} \frac{a^x \cdot a^h - a^x}{h}.$$
2. Explain why it is also true that
$$f'(x) = a^x \cdot \lim_{h \to 0} \frac{a^h - 1}{h}.$$
3. Use computing technology and small values of h to estimate the value of
$$L = \lim_{h \to 0} \frac{a^h - 1}{h}$$
when $a = 2$. Do likewise when $a = 3$.
4. Note that it would be ideal if the value of the limit L was 1, for then f would be a particularly special function: its derivative would be simply a^x, which would mean that its derivative is itself. By experimenting with different values of a between 2 and 3, try to find a value for a for which

$$L = \lim_{h \to 0} \frac{a^h - 1}{h} = 1.$$

5. Compute $\ln(2)$ and $\ln(3)$. What does your work in (b) and (c) suggest is true about $\frac{d}{dx}[2^x]$ and $\frac{d}{dx}[3^x]$?
6. How do your investigations in (d) lead to a particularly important fact about the function $f(x) = e^x$?

2.2: The Sine and Cosine Function

1

Suppose that $V(t) = 24 \cdot 1.07^t + 6\sin(t)$ represents the value of a person's investment portfolio in thousands of dollars in year t, where $t = 0$ corresponds to January 1, 2010.

1. At what instantaneous rate is the portfolio's value changing on January 1, 2012? Include units on your answer.
2. Determine the value of $V''(2)$. What are the units on this quantity and what does it tell you about how the portfolio's value is changing?
3. On the interval $0 \le t \le 20$, graph the function $V(t) = 24 \cdot 1.07^t + 6\sin(t)$ and describe its behavior in the context of the problem. Then, compare the graphs of the functions $A(t) = 24 \cdot 1.07^t$ and $V(t) = 24 \cdot 1.07^t + 6\sin(t)$, as well as the graphs of their derivatives $A'(t)$ and $V'(t)$. What is the impact of the term $6\sin(t)$ on the behavior of the function $V(t)$?

2.

Let $f(x) = 3\cos(x) - 2\sin(x) + 6$.

1. Determine the exact slope of the tangent line to $y = f(x)$ at the point where $a = \frac{\pi}{4}$.
2. Determine the tangent line approximation to $y = f(x)$ at the point where $a = \pi$.
3. At the point where $a = \frac{\pi}{2}$, is f increasing, decreasing, or neither?
4. At the point where $a = \frac{3\pi}{2}$, does the tangent line to $y = f(x)$ lie above the curve, below the curve, or neither? How can you answer this question without even graphing the function or the tangent line?

3

In this exercise, we explore how the limit definition of the derivative more formally shows that $\frac{d}{dx}[\sin(x)] = \cos(x)$. Letting $f(x) = \sin(x)$, note that the limit definition of the derivative tells us that

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

1. Recall the trigonometric identity for the sine of a sum of angles α and β: $\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$. Use this identity and some algebra to show that
$$f'(x) = \lim_{h \to 0} \frac{\sin(x)(\cos(h) - 1) + \cos(x)\sin(h)}{h}.$$
2. Next, note that as h changes, x remains constant. Explain why it therefore makes sense to say that
$$f'(x) = \sin(x) \cdot \lim_{h \to 0} \frac{\cos(h) - 1}{h} + \cos(x) \cdot \lim_{h \to 0} \frac{\sin(h)}{h}.$$
3. Finally, use small values of h to estimate the values of the two limits in (c):
$$\lim_{h \to 0} \frac{\cos(h) - 1}{h} \quad \text{and} \quad \lim_{h \to 0} \frac{\sin(h)}{h}.$$
4. What do your results in (b) and (c) thus tell you about $f'(x)$?
5. By emulating the steps taken above, use the limit definition of the derivative to argue convincingly that $\frac{d}{dx}[\cos(x)] = -\sin(x)$.

2.3: The Product and Quotient Rules

1. Derivative of a basic product

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Derivative of a product

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Derivative of a quotient of linear functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Derivative of a rational function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Derivative of a product of trigonometric functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6. Derivative of a product of power and trigonmetric functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

7. Derivative of a sum that involves a product

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

8. Product and quotient rules with graphs

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

9. Product and quotient rules with given function values

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

10

Let f and g be differentiable functions for which the following information is known: $f(2) = 5$, $g(2) = -3$, $f'(2) = -1/2$, $g'(2) = 2$.

1. Let h be the new function defined by the rule $h(x) = g(x) \cdot f(x)$. Determine $h(2)$ and $h'(2)$.
2. Find an equation for the tangent line to $y = h(x)$ at the point $(2, h(2))$ (where h is the function defined in (a)).
3. Let r be the function defined by the rule $r(x) = \frac{g(x)}{f(x)}$. Is r increasing, decreasing, or neither at $a = 2$? Why?
4. Estimate the value of $r(2.06)$ (where r is the function defined in (c)) by using the local linearization of r at the point $(2, r(2))$.

11

Consider the functions $r(t) = t^t$ and $s(t) = \arccos(t)$, for which you are given the facts that $r'(t) = t^t(\ln(t) + 1)$ and $s'(t) = -\frac{1}{\sqrt{1-t^2}}$. Do not be concerned with where these derivative formulas come from. We restrict our interest in both functions to the domain $0 < t < 1$.

1. Let $w(t) = t^t \arccos(t)$. Determine $w'(t)$.
2. Find an equation for the tangent line to $y = w(t)$ at the point $(\frac{1}{2}, w(\frac{1}{2}))$.
3. Let $v(t) = \frac{t^t}{\arccos(t)}$. Is v increasing or decreasing at the instant $t = \frac{1}{2}$? Why?

12

Let functions p and q be the piecewise linear functions given by their respective graphs in Figure 2.3.6. Use the graphs to answer the following questions.

1. Let $r(x) = p(x) \cdot q(x)$. Determine $r'(-2)$ and $r'(0)$.
2. Are there values of x for which $r'(x)$ does not exist? If so, which values, and why?
3. Find an equation for the tangent line to $y = r(x)$ at the point $(2, r(2))$.
4. Let $z(x) = \frac{q(x)}{p(x)}$. Determine $z'(0)$ and $z'(2)$.
5. Are there values of x for which $z'(x)$ does not exist? If so, which values, and why?

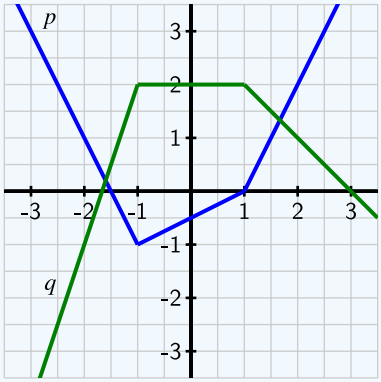


Figure 2.3.6. The graphs of p (in blue) and q (in green).

13

A farmer with large land holdings has historically grown a wide variety of crops. With the price of ethanol fuel rising, he decides that it would be prudent to devote more and more of his acreage to producing corn. As he grows more and more corn, he learns efficiencies that increase his yield per acre. In the present year, he used 7000 acres of his land to grow corn, and that land had an average yield of 170 bushels per acre. At the current time, he plans to increase his number of acres devoted to growing corn at a rate of 600 acres/year, and he expects that right now his average yield is increasing at a rate of 8 bushels per acre per year. Use this information to answer the following questions.

1. Say that the present year is $t = 0$, that $A(t)$ denotes the number of acres the farmer devotes to growing corn in year t, $Y(t)$ represents the average yield in year t (measured in bushels per acre), and $C(t)$ is the total number of bushels of corn the farmer produces. What is the formula for $C(t)$ in terms of $A(t)$ and $Y(t)$? Why?
2. What is the value of $C(0)$? What does it measure?
3. Write an expression for $C'(t)$ in terms of $A(t)$, $A'(t)$, $Y(t)$, and $Y'(t)$. Explain your thinking.
4. What is the value of $C'(0)$? What does it measure?
5. Based on the given information and your work above, estimate the value of $C(1)$.

14

Let $f(v)$ be the gas consumption (in liters/km) of a car going at velocity v (in km/hour). In other words, $f(v)$ tells you how many liters of gas the car uses to go one kilometer if it is traveling at v kilometers per hour. In addition, suppose that $f(80) = 0.05$ and $f'(80) = 0.0004$.

1. Let $g(v)$ be the distance the same car goes on one liter of gas at velocity v. What is the relationship between $f(v)$ and $g(v)$? Hence find $g(80)$ and $g'(80)$.
2. Let $h(v)$ be the gas consumption in liters per hour of a car going at velocity v. In other words, $h(v)$ tells you how many liters of gas the car uses in one hour if it is going at velocity v. What is the algebraic relationship between $h(v)$ and $f(v)$? Hence find $h(80)$ and $h'(80)$.
3. How would you explain the practical meaning of these function and derivative values to a driver who knows no calculus? Include units on each of the function and derivative values you discuss in your response.

2.4: Derivatives of Other Trigonometric Functions

1. A sum and product involving $\tan(x)$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. A quotient involving $\tan(t)$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. A quotient of trigonometric functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. A quotient that involves a product

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Finding a tangent line equation

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



6

An object moving vertically has its height at time t (measured in feet, with time in seconds) given by the function $h(t) = 3 + \frac{2\cos(t)}{1.2^t}$.

1. What is the object's instantaneous velocity when $t = 2$?
2. What is the object's acceleration at the instant $t = 2$?
3. Describe in everyday language the behavior of the object at the instant $t = 2$.

7

Let $f(x) = \sin(x)\cot(x)$.

1. Use the product rule to find $f'(x)$.
2. True or false: for all real numbers x, $f(x) = \cos(x)$.
3. Explain why the function that you found in (a) is almost the opposite of the sine function, but not quite. (Hint: convert all of the trigonometric functions in (a) to sines and cosines, and work to simplify. Think carefully about the domain of f and the domain of f'.)

8

Let $p(z)$ be given by the rule

$$p(z) = \frac{z\tan(z)}{z^2\sec(z) + 1} + 3e^z + 1.$$

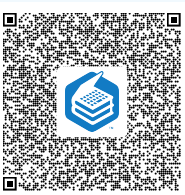
1. Determine $p'(z)$.
2. Find an equation for the tangent line to p at the point where $z = 0$.
3. At $z = 0$, is p increasing, decreasing, or neither? Why?

2.5: The Chain Rule

1. Mixing rules: chain, product, sum

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. Mixing rules: chain and product

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Using the chain rule repeatedly

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Derivative involving arbitrary constants a and b

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Chain rule with graphs

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6. Chain rule with function values

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

7. A product involving a composite function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

8

Consider the basic functions $f(x) = x^3$ and $g(x) = \sin(x)$.

1. Let $h(x) = f(g(x))$. Find the exact instantaneous rate of change of h at the point where $x = \frac{\pi}{4}$.
2. Which function is changing most rapidly at $x = 0.25$: $h(x) = f(g(x))$ or $r(x) = g(f(x))$? Why?
3. Let $h(x) = f(g(x))$ and $r(x) = g(f(x))$. Which of these functions has a derivative that is periodic? Why?

9

Let $u(x)$ be a differentiable function. For each of the following functions, determine the derivative. Each response will involve u and/or u'.

1. $p(x) = e^{u(x)}$
2. $q(x) = u(e^x)$
3. $r(x) = \cot(u(x))$
4. $s(x) = u(\cot(x))$
5. $a(x) = u(x^4)$
6. $b(x) = u^4(x)$

10

Let functions p and q be the piecewise linear functions given by their respective graphs in Figure 2.5.9. Use the graphs to answer the following questions.

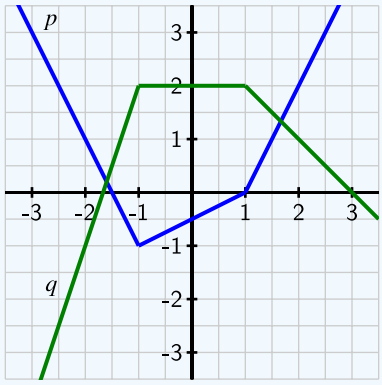


Figure 2.5.9. The graphs of p (in blue) and q (in green).

1. Let $C(x) = p(q(x))$. Determine $C'(0)$ and $C'(3)$.
2. Find a value of x for which $C'(x)$ does not exist. Explain your thinking.
3. Let $Y(x) = q(q(x))$ and $Z(x) = q(p(x))$. Determine $Y'(-2)$ and $Z'(0)$.

11

If a spherical tank of radius 4 feet has h feet of water present in the tank, then the volume of water in the tank is given by the formula

$$V = \frac{\pi}{3}h^2(12 - h).$$

1. At what instantaneous rate is the volume of water in the tank changing with respect to the *height* of the water at the instant $h = 1$? What are the units on this quantity?
2. Now suppose that the height of water in the tank is being regulated by an inflow and outflow (e.g., a faucet and a drain) so that the height of the water at time t is given by the rule $h(t) = \sin(\pi t) + 1$, where t is measured in hours (and h is still measured in feet). At what rate is the height of the water changing with respect to time at the instant $t = 2$?
3. Continuing under the assumptions in (b), at what instantaneous rate is the volume of water in the tank changing with respect to *time* at the instant $t = 2$?
4. What are the main differences between the rates found in (a) and (c)? Include a discussion of the relevant units.

2.6: Derivatives of Inverse Functions

1. Composite function involving logarithms and polynomials

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Composite function involving trigonometric functions and logarithms

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Product involving $\arcsin(w)$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Derivative involving $\arctan(x)$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5. Composite function from a graph

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



6. Composite function involving an inverse trigonometric function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



7. Mixing rules: product, chain, and inverse trig

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



8. Mixing rules: product and inverse trig

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



9

Determine the derivative of each of the following functions. Use proper notation and clearly identify the derivative rules you use.

1. $f(x) = \ln(2\arctan(x) + 3\arcsin(x) + 5)$
2. $r(z) = \arctan(\ln(\arcsin(z)))$
3. $q(t) = \arctan^2(3t)\arcsin^4(7t)$
4. $g(v) = \ln\left(\dfrac{\arctan(v)}{\arcsin(v) + v^2}\right)$

10

Consider the graph of $y = f(x)$ provided in Figure 2.6.7 and use it to answer the following questions.

1. Use the provided graph to estimate the value of $f'(1)$.
2. Sketch an approximate graph of $y = f^{-1}(x)$. Label at least three distinct points on the graph that correspond to three points on the graph of f.
3. Based on your work in (a), what is the value of $(f^{-1})'(-1)$? Why?

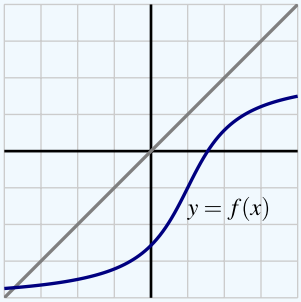


Figure 2.6.7. A function $y = f(x)$

11

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

1. Sketch a graph of $y = f(x)$ and explain why f is an invertible function.
2. Let g be the inverse of f and determine a formula for g.
3. Compute $f'(x)$, $g'(x)$, $f'(2)$, and $g'(6)$. What is the special relationship between $f'(2)$ and $g'(6)$? Why?

12

Let $h(x) = x + \sin(x)$.

1. Sketch a graph of $y = h(x)$ and explain why h must be invertible.
2. Explain why it does not appear to be algebraically possible to determine a formula for h^{-1}.
3. Observe that the point $(\frac{\pi}{2}, \frac{\pi}{2} + 1)$ lies on the graph of $y = h(x)$. Determine the value of $(h^{-1})'(\frac{\pi}{2} + 1)$.

2.7: Derivatives of Functions Given Implicitely

1. Implicit differentiaion in a polynomial equation

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. Implicit differentiation in an equation with logarithms

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

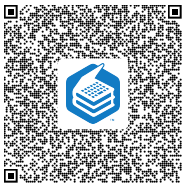
View this Interactive Content online



3. Implicit differentiation in an equation with inverse trigonometric functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

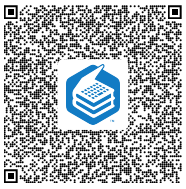
View this Interactive Content online



4. Slope of the tangent line to an implicit curve

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5. Equation of the tangent line to an implicit curve

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



6

Consider the curve given by the equation $2y^3 + y^2 - y^5 = x^4 - 2x^3 + x^2$. Find all points at which the tangent line to the curve is horizontal or vertical. Be sure to use a graphing utility to plot this implicit curve and to visually check the results of algebraic reasoning that you use to determine where the tangent lines are horizontal and vertical.

7

For the curve given by the equation $\sin(x + y) + \cos(x - y) = 1$, find the equation of the tangent line to the curve at the point $(\frac{\pi}{2}, \frac{\pi}{2})$.

8

Implicit differentiation enables us a different perspective from which to see why the rule $\frac{d}{dx}[a^x] = a^x \ln(a)$ holds, if we assume that $\frac{d}{dx}[\ln(x)] = \frac{1}{x}$. This exercise leads you through the key steps to do so.

1. Let $y = a^x$. Rewrite this equation using the natural logarithm function to write x in terms of y (and the constant a).
2. Differentiate both sides of the equation you found in (a) with respect to x, keeping in mind that y is implicitly a function of x.
3. Solve the equation you found in (b) for $\frac{dy}{dx}$, and then use the definition of y to write $\frac{dy}{dx}$ solely in terms of x. What have you found?

2.8: Using Derivatives to Evaluate Limits

1. L'Hôpital's Rule with graphs

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. L'Hôpital's Rule to evaluate a limit

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Determining if L'Hôpital's Rule applies

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Using L'Hôpital's Rule multiple times

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5

Let f and g be differentiable functions about which the following information is known: $f(3) = g(3) = 0$, $f'(3) = g'(3) = 0$, $f''(3) = -2$, and $g''(3) = 1$. Let a new function h be given by the rule $h(x) = \frac{f(x)}{g(x)}$. On the same set of axes, sketch possible graphs of f and g near $x = 3$, and use the provided information to determine the value of

$$\lim_{x \to 3} h(x).$$

Provide explanation to support your conclusion.

6

Find all vertical and horizontal asymptotes of the function

$$R(x) = \frac{3(x-a)(x-b)}{5(x-a)(x-c)},$$

where a, b, and c are distinct, arbitrary constants. In addition, state all values of x for which R is not continuous. Sketch a possible graph of R, clearly labeling the values of a, b, and c.

7

Consider the function $g(x) = x^{2x}$, which is defined for all $x > 0$. Observe that $\lim_{x \to 0^+} g(x)$ is indeterminate due to its form of 0^0. (Think about how we know that $0^k = 0$ for all $k > 0$, while $b^0 = 1$ for all $b \neq 0$, but that neither rule can apply to 0^0.)

1. Let $h(x) = \ln(g(x))$. Explain why $h(x) = 2x \ln(x)$.
2. Next, explain why it is equivalent to write $h(x) = \frac{2\ln(x)}{\frac{1}{x}}$.
3. Use L'Hôpital's Rule and your work in (b) to compute $\lim_{x \to 0^+} h(x)$.
4. Based on the value of $\lim_{x \to 0^+} h(x)$, determine $\lim_{x \to 0^+} g(x)$.

8

Recall we say that function g *dominates* function f provided that $\lim_{x\to\infty} f(x) = \infty$, $\lim_{x\to\infty} g(x) = \infty$, and $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$.

1. Which function dominates the other: $\ln(x)$ or $\sqrt{x}$?
2. Which function dominates the other: $\ln(x)$ or $\sqrt[n]{x}$? (n can be any positive integer)
3. Explain why e^x will dominate any polynomial function.
4. Explain why x^n will dominate $\ln(x)$ for any positive integer n.
5. Give any example of two nonlinear functions such that neither dominates the other

This page titled 2.9: Computing Derivatives (Exercises) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.E: Computing Derivatives (Exercises) has no license indicated.
- 2.1: Elementary Derivative Rules by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 2.2: The Sine and Cosine Function by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 2.3: The Product and Quotient Rules by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 2.4: Derivatives of Other Trigonometric Functions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 2.5: The Chain Rule by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 2.6: Derivatives of Inverse Functions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 2.7: Derivatives of Functions Given Implicitly by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 2.8: Using Derivatives to Evaluate Limits by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

3: USING DERIVATIVES

3.1: Using Derivatives to Identify Extreme Values
3.2: Using Derivatives to Describe Families of Functions
3.3: Global Optimization
3.4: Applied Optimization
3.5: Related Rates
3.6: Using Derivatives (Exercises)

This page titled 3: Using Derivatives is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 3: Using Derivatives by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

3.1: Using Derivatives to Identify Extreme Values

Motivating Questions

- What are the critical numbers of a function f and how are they connected to identifying the most extreme values the function achieves?
- How does the first derivative of a function reveal important information about the behavior of the function, including the function's extreme values?
- How can the second derivative of a function be used to help identify extreme values of the function?

In many different settings, we are interested in knowing where a function achieves its least and greatest values. These can be important in applications — say to identify a point at which maximum profit or minimum cost occurs — or in theory to characterize the behavior of a function or a family of related functions.

Consider the simple and familiar example of a parabolic function such as $s(t) = -16t^2 + 32t + 48$ (shown at left in Figure 3.1.1) that represents the height of an object tossed vertically: its maximum value occurs at the vertex of the parabola and represents the greatest height the object reaches. This maximum value is an especially important point on the graph, the point at which the curve changes from increasing to decreasing.

Definition 3.1.1

Given a function f, we say that $f(c)$ is a ***global*** or ***absolute maximum*** of f provided that $f(c) \geq f(x)$ for all x in the domain of f, and similarly we call $f(c)$ a ***global*** or ***absolute minimum*** of f whenever $f(c) \leq f(x)$ for all x in the domain of f.

For instance, in Figure 3.1.1, g has a global maximum of $g(c)$, but g does not appear to have a global minimum, as the graph of g seems to decrease without bound. Note that the point $(c, g(c))$ marks a fundamental change in the behavior of g, where g changes from increasing to decreasing; similar things happen at both $(a, g(a))$ and $(b, g(b))$, although these points are not global minima or maxima.

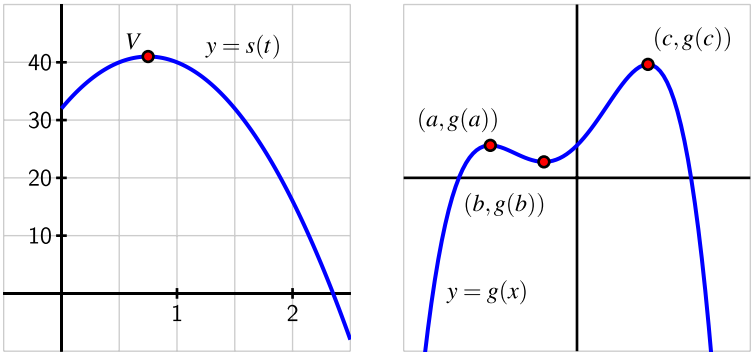


Figure 3.1.1. At left, $s(t) = -16t^2 + 24t + 32$ whose vertex is $(\frac{3}{4}, 41)$; at right, a function g that demonstrates several high and low points.

Definition 3.1.2

We say that $f(c)$ is a ***local maximum*** or ***relative maximum*** of f provided that $f(c) \geq f(x)$ for all x near c, and $f(c)$ is called a ***local*** or ***relative minimum*** of f whenever $f(c) \leq f(x)$ for all x near c.

For example, in Figure 3.1.1, g has a relative minimum of $g(b)$ at the point $(b, g(b))$ and a relative maximum of $g(a)$ at $(a, g(a))$. We

have already identified the global maximum of g as $g(c)$; it can also be considered a relative maximum. Any maximum or minimum may also be called an *extreme value* of f.

We would like to use calculus ideas to identify and classify key function behavior, including the location of relative extremes. Of course, if we are given a graph of a function, it is often straightforward to locate these important behaviors visually.

Preview Activity 3.1.1

Consider the function h given by the graph in Figure 3.1.2. Use the graph to answer each of the following questions.

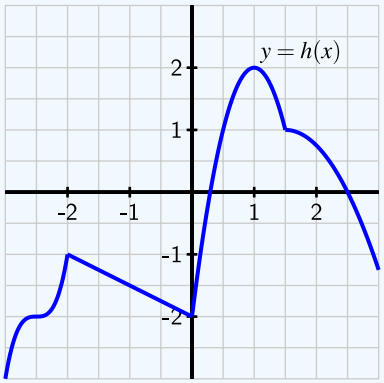


Figure 3.1.2. The graph of a function h on the interval $[-3, 3]$.

1. Identify all of the values of c such that $-3 < c < 3$ for which $h(c)$ is a local maximum of h.
2. Identify all of the values of c such that $-3 < c < 3$ for which $h(c)$ is a local minimum of h.
3. Does h have a global maximum on the interval $[-3, 3]$? If so, what is the value of this global maximum?
4. Does h have a global minimum on the interval $[-3, 3]$? If so, what is its value?
5. Identify all values of c for which $h'(c) = 0$.
6. Identify all values of c for which $h'(c)$ does not exist.
7. True or false: every relative maximum and minimum of h occurs at a point where $h'(c)$ is either zero or does not exist.
8. True or false: at every point where $h'(c)$ is zero or does not exist, h has a relative maximum or minimum.

Critical numbers and the first derivative test

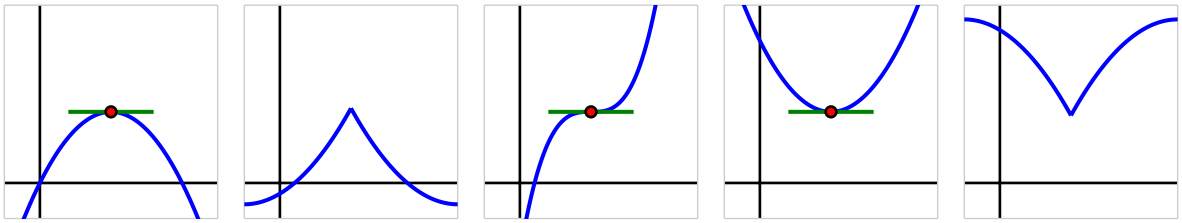


Figure 3.1.3. From left to right, a function with a relative maximum where its derivative is zero; a function with a relative maximum where its derivative is undefined; a function with neither a maximum nor a minimum at a point where its derivative is zero; a function with a relative minimum where its derivative is zero; and a function with a relative minimum where its derivative is undefined.

If a continuous function has a relative maximum at c, then it is both necessary and sufficient that the function change from being increasing just before c to decreasing just after c. A continuous function has a relative minimum at c if and only if the function changes from decreasing to increasing at c. (See Figure 3.1.3.) There are only two possible ways for these changes in behavior to occur: either $f'(c) = 0$ or $f'(c)$ is undefined. Because these values of c are so important, we call them *critical numbers*.

Definition 3.1.3

We say that a function f has a ***critical number*** at $x = c$ provided that c is in the domain of f, and $f'(c) = 0$ or $f'(c)$ is undefined.

Critical numbers are the only possible locations where the function f may have relative extremes. Note that not every critical number produces a maximum or minimum; in the middle graph of Figure 3.1.3, the function pictured there has a horizontal tangent line at the noted point, but the function is increasing before and increasing after, so the critical number does not yield a maximum or minimum.

When c is a critical number, we say that $(c, f(c))$ is a *critical point* of the function, or that $f(c)$ is a *critical value* . The *first derivative test* summarizes how sign changes in the first derivative (which can only occur at critical numbers) indicate the presence of a local maximum or minimum for a given function.

First Derivative Test

If p is a critical number of a continuous function f that is differentiable near p (except possibly at $x = p$), then f has a relative maximum at p if and only[1] if f' changes sign from positive to negative at p, and f has a relative minimum at p if and only if f' changes sign from negative to positive at p.

Technically, we also have to assume that f is not piecewise constant on any intervals. This is because every point on a horizontal line is a relative maximum (and relative minimum) despite the fact that the derivative doesn't change sign at any point along the horizontal line.

Example 3.1.1

Let f be a function whose derivative is given by the formula $f'(x) = e^{-2x}(3 - x)(x + 1)^2$. Determine all critical numbers of f and decide whether a relative maximum, relative minimum, or neither occurs at each.

Answer available at the end of this section.

Activity 3.1.2

Suppose that $g(x)$ is a function continuous for every value of $x \neq 2$ whose first derivative is $g'(x) = \frac{(x+4)(x-1)^2}{x-2}$. Further, assume that it is known that g has a vertical asymptote at $x = 2$.

1. Determine all critical numbers of g.
2. By developing a carefully labeled first derivative sign chart, decide whether g has as a local maximum, local minimum, or neither at each critical number.
3. Does g have a global maximum? global minimum? Justify your claims.
4. What is the value of $\lim_{x\to\infty} g'(x)$? What does the value of this limit tell you about the long-term behavior of g?
5. Sketch a possible graph of $y = g(x)$.

The second derivative test

Recall that the second derivative of a function tells us several important things about the behavior of the function itself. For instance, if f'' is positive on an interval, then we know that f' is increasing on that interval and, consequently, that f is concave up, so throughout that interval the tangent line to $y = f(x)$ lies below the curve at every point. At a point where $f'(p) = 0$, the sign of the second derivative determines whether f has a local minimum or local maximum at the critical number p.

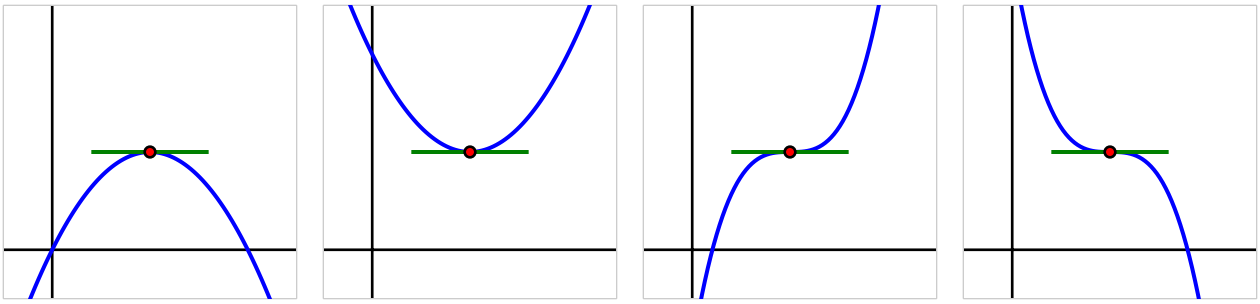


Figure 3.1.5. Four possible graphs of a function f with a horizontal tangent line at a critical point.

In Figure 3.1.5 , we see the four possibilities for a function f that has a critical number p at which $f'(p) = 0$, provided $f''(p)$ is not zero on an interval including p (except possibly at p). On either side of the critical number, f'' can be either positive or negative, and hence f can be either concave up or concave down. In the first two graphs, f does not change concavity at p, and in those situations, f has either a local minimum or local maximum. In particular, if $f'(p) = 0$ and $f''(p) < 0$, then f is concave down at p with a horizontal tangent line, so f has a local maximum there. This fact, along with the corresponding statement for when $f''(p)$ is positive, is the substance of the *second derivative test*.

Second Derivative Test

If p is a critical number of a continuous function f such that $f'(p) = 0$ and $f''(p) \neq 0$, then f has a relative maximum at p if and only if $f''(p) < 0$, and f has a relative minimum at p if and only if $f''(p) > 0$.

In the event that $f''(p) = 0$, the second derivative test is inconclusive. That is, the test doesn't provide us any information. This is because if $f''(p) = 0$, it is possible that f has a local minimum, local maximum, or neither. [2]

Consider the functions $f(x) = x^4$, $g(x) = -x^4$, and $h(x) = x^3$ at the critical point $p = 0$.

Just as a first derivative sign chart reveals all of the increasing and decreasing behavior of a function, we can construct a second derivative sign chart that demonstrates all of the important information involving concavity.

Example 3.1.2

Let $f(x)$ be a function whose first derivative is $f'(x) = 3x^4 - 9x^2$. Construct both first and second derivative sign charts for f, fully discuss where f is increasing and decreasing and concave up and concave down, identify all relative extreme values, and sketch a possible graph of f.

Answer available at the end of this section.

Points B, C, and D in Figure 3.1.8 are locations at which the concavity of f changes. We give a special name to any such point.

Definition 3.1.4

If p is a value in the domain of a continuous function f at which f changes concavity, then we say that $(p, f(p))$ is an ***inflection point*** (or ***point of inflection***) of f.

Just as we look for locations where f changes from increasing to decreasing at points where $f'(p) = 0$ or $f'(p)$ is undefined, so too we find where $f''(p) = 0$ or $f''(p)$ is undefined to see if there are points of inflection at these locations.

At this point in our study, it is important to remind ourselves of the big picture that derivatives help to paint: the sign of the first derivative f' tells us *whether* the function f is increasing or decreasing, while the sign of the second derivative f'' tells us *how*

the function f is increasing or decreasing.

Activity 3.1.3

Suppose that g is a function whose second derivative, g'', is given by the graph in Figure 3.1.9.

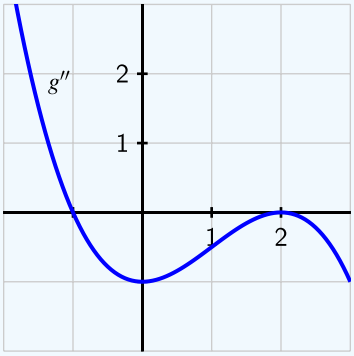


Figure 3.1.10. The graph of $y = g''(x)$.

1. Find the x-coordinates of all points of inflection of g.
2. Fully describe the concavity of g by making an appropriate sign chart.
3. Suppose you are given that $g'(-1.67857351) = 0$. Is there is a local maximum, local minimum, or neither (for the function g) at this critical number of g, or is it impossible to say? Why?
4. Assuming that $g''(x)$ is a polynomial (and that all important behavior of g'' is seen in the graph above), what degree polynomial do you think $g(x)$ is? Why?

As we will see in more detail in the following section, derivatives also help us to understand families of functions that differ only by changing one or more parameters. For instance, we might be interested in understanding the behavior of all functions of the form $f(x) = a(x - h)^2 + k$ where a, h, and k are parameters. Each parameter has considerable impact on how the graph appears.

Activity 3.1.4

Consider the family of functions given by $h(x) = x^2 + \cos(kx)$, where k is an arbitrary positive real number.

a. Use a graphing utility to sketch the graph of h for several different k-values, including $k = 1, 3, 5, 10$. Plot $h(x) = x^2 + \cos(3x)$ on the axes provided. What is the smallest value of k at which you think you can see (just by looking at the graph) at least one inflection point on the graph of h?

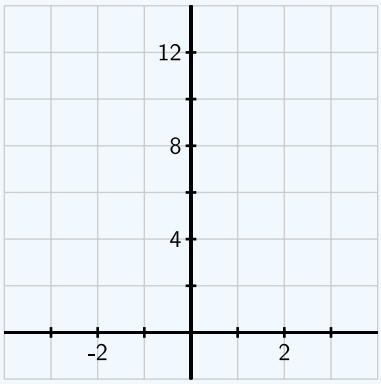


Figure 3.1.11. Axes for plotting $y = h(x)$.

b. Explain why the graph of h has no inflection points if $k \leq \sqrt{2}$, but infinitely many inflection points if $k > \sqrt{2}$.

c. Explain why, no matter the value of k, h can only have finitely many critical numbers.

Summary

- The critical numbers of a continuous function f are the values of p for which $f'(p) = 0$ or $f'(p)$ does not exist. These values are important because they identify horizontal tangent lines or corner points on the graph, which are the only possible locations at which a local maximum or local minimum can occur.
- Given a differentiable function f, whenever f' is positive, f is increasing; whenever f' is negative, f is decreasing. The first derivative test tells us that at any point where f changes from increasing to decreasing, f has a local maximum, while conversely at any point where f changes from decreasing to increasing f has a local minimum.
- Given a twice differentiable function f, if we have a horizontal tangent line at $x = p$ and $f''(p)$ is nonzero, the sign of f'' tells us the concavity of $f$and hence whether f has a maximum or minimum at $x = p$. In particular, if $f'(p) = 0$ and $f''(p) < 0$, then f is concave down at p and f has a local maximum there, while if $f'(p) = 0$ and $f''(p) > 0$, then f has a local minimum at p. If $f'(p) = 0$ and $f''(p) = 0$, then the second derivative does not tell us whether f has a local extreme at p or not.

Answers

Example 3.1.1

Since we already have $f'(x)$ written in factored form, it is straightforward to find the critical numbers of f. Because $f'(x)$ is defined for all values of x, we need only determine where $f'(x) = 0$. From the equation

$$e^{-2x}(3-x)(x+1)^2 = 0$$

and the zero product property, it follows that $x = 3$ and $x = -1$ are critical numbers of f. (There is no value of x that makes $e^{-2x} = 0$.)

Next, to apply the first derivative test, we'd like to know the sign of $f'(x)$ at inputs near the critical numbers. Because the critical numbers are the only locations at which f' can change sign, it follows that the sign of the derivative is the same on each of the intervals created by the critical numbers: for instance, the sign of f' must be the same for every $x < -1$. We create a first derivative sign chart to summarize the sign of f' on the relevant intervals, along with the corresponding behavior of f.

$$f'(x) = e^{-2x}(3-x)(x+1)^2$$

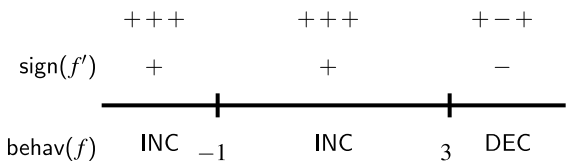


Figure 3.1.4. The first derivative sign chart for a function f whose derivative is given by the formula $f'(x) = e^{-2x}(3-x)(x+1)^2$.

To produce the first derivative sign chart in Figure 3.1.4 we identify the sign of each factor of $f'(x)$ at one selected point in each interval. For instance, for $x < -1$, we could determine the sign of e^{-2x}, $(3-x)$, and $(x+1)^2$ at the value $x = -2$. We note that both e^{-2x} and $(x+1)^2$ are positive regardless of the value of x, while $(3-x)$ is also positive at $x = -2$. Hence, each of the three terms in f' is positive, which we indicate by writing "$+++$." Taking the product of three positive terms results in a positive value for f', which we denote by the "$+$" in the interval to the left of $x = -1$. And, since f' is positive on that interval, we know that f is increasing, so we write "INC" to represent the behavior of f. In a similar way, we find that f' is positive and f is increasing on $-1 < x < 3$, and f' is negative and f is decreasing for $x > 3$.

Now we look for critical numbers at which f' changes sign. In this example, f' changes sign only at $x = 3$, from positive to negative, so f has a relative maximum at $x = 3$. Although f has a critical number at $x = -1$, since f is increasing both before and after $x = -1$, f has neither a minimum nor a maximum at $x = -1$.

Example 3.1.2

Since we know $f'(x) = 3x^4 - 9x^2$, we can find the critical numbers of f by solving $3x^4 - 9x^2 = 0$. Factoring, we observe that

$$0 = 3x^2(x^2-3) = 3x^2(x+\sqrt{3})(x-\sqrt{3}),$$

so that $x = 0, \pm\sqrt{3}$ are the three critical numbers of f. The first derivative sign chart for f is given in Figure 3.1.6 .

$$f'(x) = 3x^2(x+\sqrt{3})(x-\sqrt{3})$$

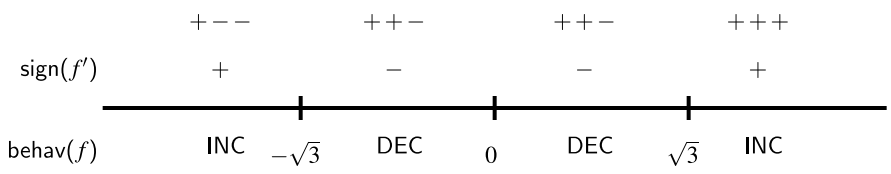


Figure 3.1.6 . The first derivative sign chart for f when $f'(x) = 3x^4 - 9x^2 = 3x^2(x^2-3)$.

We see that f is increasing on the intervals $(-\infty, -\sqrt{3})$ and $(\sqrt{3}, \infty)$, and f is decreasing on $(-\sqrt{3}, 0)$ and $(0, \sqrt{3})$. By the first derivative test, this information tells us that f has a local maximum at $x = -\sqrt{3}$ and a local minimum at $x = \sqrt{3}$. Although f also has a critical number at $x = 0$, neither a maximum nor minimum occurs there since f' does not change sign at $x = 0$.

Next, we move on to investigate concavity. Differentiating $f'(x) = 3x^4 - 9x^2$, we see that $f''(x) = 12x^3 - 18x$. Since we are interested in knowing the intervals on which f'' is positive and negative, we first find where $f''(x) = 0$. Observe that

$$0 = 12x^3 - 18x = 12x\left(x^2 - \frac{3}{2}\right) = 12x\left(x + \sqrt{\frac{3}{2}}\right)\left(x - \sqrt{\frac{3}{2}}\right).$$

This equation has solutions $x = 0, \pm\sqrt{\frac{3}{2}}$. Building a sign chart for f'' in the exact same way we do for f', we see the result shown in Figure 3.1.7.

$$f''(x) = 12x\left(x + \sqrt{\tfrac{3}{2}}\right)\left(x - \sqrt{\tfrac{3}{2}}\right)$$

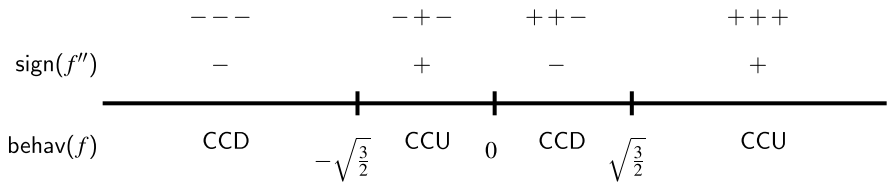


Figure 3.1.7. The second derivative sign chart for f when $f''(x) = 12x^3 - 18x = 12x^2\left(x^2 - \frac{3}{2}\right)$.

Therefore, f is concave down on the intervals $(-\infty, -\sqrt{\frac{3}{2}})$ and $(0, \sqrt{\frac{3}{2}})$, and concave up on $(-\sqrt{\frac{3}{2}}, 0)$ and $(\sqrt{\frac{3}{2}}, \infty)$.

Putting all of this information together, we now see a complete and accurate possible graph of f in Figure 3.1.8.

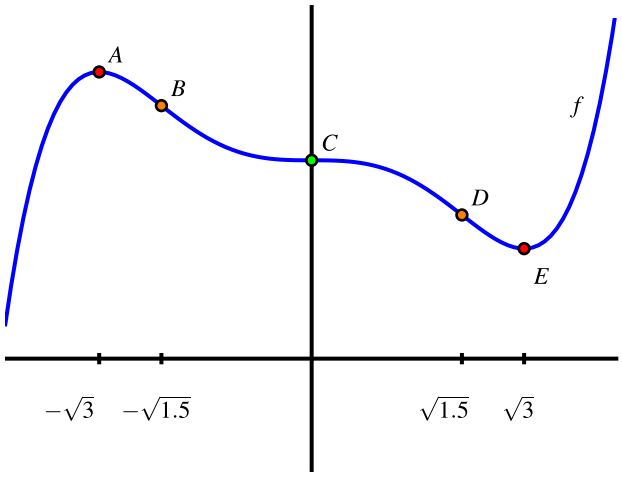


Figure 3.1.8 . A possible graph of the function f in Example 3.1.10.

The point $A = (-\sqrt{3}, f(-\sqrt{3}))$ is a local maximum, because f is increasing prior to A and decreasing after; similarly, the point $E = (\sqrt{3}, f(\sqrt{3})$ is a local minimum. Note, too, that f is concave down at A and concave up at B, which is consistent both with our second derivative sign chart and the second derivative test. At points B and D, concavity changes, as we saw in the results of the second derivative sign chart in Figure 3.1.8 . Finally, at point C, f has a critical point with a horizontal tangent line, but neither a maximum nor a minimum occurs there, since f is decreasing both before and after C. It is also the case that concavity changes at C.

While we completely understand where f is increasing and decreasing, where f is concave up and concave down, and where f has relative extremes, we do not know any specific information about the y-coordinates of points on the curve. For instance, while we know that f has a local maximum at $x = -\sqrt{3}$, we don't know the value of that maximum because we do not know $f(-\sqrt{3})$. Any vertical translation of our sketch of f in Figure 3.1.8 would satisfy the given criteria for f.

This page titled 3.1: Using Derivatives to Identify Extreme Values is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 3.1: Using Derivatives to Identify Extreme Values by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

3.2: Using Derivatives to Describe Families of Functions

Motivating Questions

- Given a family of functions that depends on one or more parameters, how does the shape of the graph of a typical function in the family depend on the value of the parameters?
- How can we construct first and second derivative sign charts of functions that depend on one or more parameters while allowing those parameters to remain arbitrary constants?

Mathematicians are often interested in making general observations, say by describing patterns that hold in a large number of cases. Think about the Pythagorean Theorem: it doesn't tell us something about a single right triangle, but rather a fact about *every* right triangle. In the next part of our studies, we use calculus to make general observations about families of functions that depend on one or more parameters. People who use applied mathematics, such as engineers and economists, often encounter the same types of functions where only small changes to certain constants occur. These constants are called *parameters*.

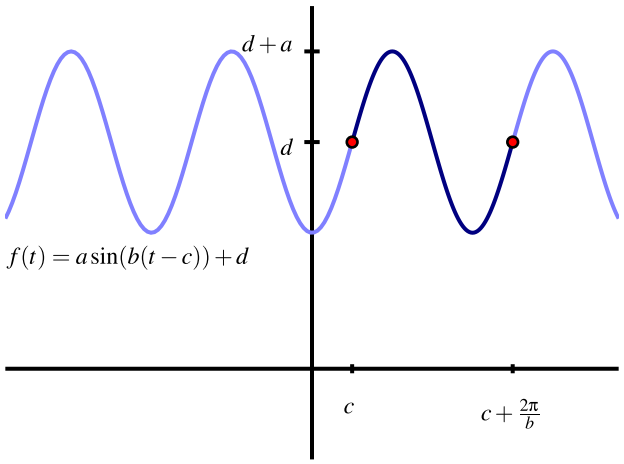


Figure 3.2.1 The graph of $f(t)=a\sin(b(t-c))+d$ based on parameters a, b, c, and d.

You are already familiar with certain families of functions. For example, $f(t)=a\sin(b(t-c))+d$ is a stretched and shifted version of the sine function with amplitude a, period $\frac{2\pi}{b}$, phase shift c, and vertical shift d. We know that a affects the size of the oscillation, b the rapidity of oscillation, and c where the oscillation starts, as shown in Figure 3.2.1, while d affects the vertical positioning of the graph.

As another example, every function of the form $y=mx+b$ is a line with slope m and y-intercept $(0,b)$. The value of m affects the line's steepness, and the value of b situates the line vertically on the coordinate axes. These two parameters describe all possible non-vertical lines.

For other less familiar families of functions, we can use calculus to discover where key behavior occurs: where members of the family are increasing or decreasing, concave up or concave down, where relative extremes occur, and more, all in terms of the parameters involved. To get started, we revisit a common collection of functions to see how calculus confirms things we already know.

Preview Activity 3.2.1

Let a, h, and k be arbitrary real numbers with $a\neq 0$, and let f be the function given by the rule $f(x)=a(x-h)^2+k$.

1. What familiar type of function is f? What information do you know about f just by looking at its form? (Think about the roles of a, h, and k.)
2. Next we use some calculus to develop familiar ideas from a different perspective. To start, treat a, h, and k as constants and compute $f'(x)$.
3. Find all critical numbers of f. (These will depend on at least one of a, h, and k.)
4. Assume that $a < 0$. Construct a first derivative sign chart for f.
5. Based on the information you've found above, classify the critical values of f as maxima or minima.

Describing families of functions in terms of parameters

Our goal is to describe the key characteristics of the overall behavior of each member of a family of functions in terms of its parameters. By finding the first and second derivatives and constructing sign charts (each of which may depend on one or more of the parameters), we can often make broad conclusions about how each member of the family will appear.

Example 3.2.1

Consider the two-parameter family of functions given by $g(x) = axe^{-bx}$, where a and b are positive real numbers. Fully describe the behavior of a typical member of the family in terms of a and b, including the location of all critical numbers, where g is increasing, decreasing, concave up, and concave down, and the long term behavior of g.

Answer available at the end of this section.

The work we've completed in Example 3.2.1 can often be replicated for other families of functions that depend on parameters. Normally we are most interested in determining all critical numbers, a first derivative sign chart, a second derivative sign chart, and the limit of the function as $x \to \infty$. Throughout, we prefer to work with the parameters as arbitrary constants. In addition, we can experiment with some particular values of the parameters present to reduce the algebraic complexity of our work. The following activities offer several key examples where we see that the values of the parameters substantially affect the behavior of individual functions within a given family.

Activity 3.2.2

Consider the family of functions defined by $p(x) = x^3 - ax$, where $a \neq 0$ is an arbitrary constant.

1. Find $p'(x)$ and determine the critical numbers of p. How many critical numbers does p have?
2. Construct a first derivative sign chart for p. What can you say about the overall behavior of p if the constant a is positive? Why? What if the constant a is negative? In each case, describe the relative extremes of p.
3. Find $p''(x)$ and construct a second derivative sign chart for p. What does this tell you about the concavity of p? What role does a play in determining the concavity of p?
4. Without using a graphing utility, sketch and label typical graphs of $p(x)$ for the cases where $a > 0$ and $a < 0$. Label all inflection points and local extrema.
5. Finally, use a graphing utility to test your observations above by entering and plotting the function $p(x) = x^3 - ax$ for at least four different values of a. Write several sentences to describe your overall conclusions about how the behavior of p depends on a.

Activity 3.2.3

Consider the two-parameter family of functions of the form $h(x) = a(1 - e^{-bx})$, where a and b are positive real numbers.

1. Find the first derivative and the critical numbers of h. Use these to construct a first derivative sign chart and determine for which values of x the function h is increasing and decreasing.
2. Find the second derivative and build a second derivative sign chart. For which values of x is a function in this family concave up? concave down?

3. What is the value of $\lim_{x\to\infty} a(1-e^{-bx})$? $\lim_{x\to-\infty} a(1-e^{-bx})$?
4. How does changing the value of b affect the shape of the curve?
5. Without using a graphing utility, sketch the graph of a typical member of this family. Write several sentences to describe the overall behavior of a typical function h and how this behavior depends on a and b.

Activity 3.2.4

Let $L(t) = \frac{A}{1+ce^{-kt}}$, where A, c, and k are all positive real numbers.

1. Observe that we can equivalently write $L(t) = A(1+ce^{-kt})^{-1}$. Find $L'(t)$ and explain why L has no critical numbers. Is L always increasing or always decreasing? Why?
2. Given the fact that
$$L''(t) = Ack^2e^{-kt}\frac{ce^{-kt}-1}{(1+ce^{-kt})^3},$$
find all values of t such that $L''(t) = 0$ and hence construct a second derivative sign chart. For which values of t is a function in this family concave up? concave down?
3. What is the value of $\lim_{t\to\infty} \frac{A}{1+ce^{-kt}}$? $\lim_{t\to-\infty} \frac{A}{1+ce^{-kt}}$?
4. Find the value of $L(x)$ at the inflection point found in (b).
5. Without using a graphing utility, sketch the graph of a typical member of this family. Write several sentences to describe the overall behavior of a typical function L and how this behavior depends on A, c, and k number.
6. Explain why it is reasonable to think that the function $L(t)$ models the growth of a population over time in a setting where the largest possible population the surrounding environment can support is A.

Summary

- Given a family of functions that depends on one or more parameters, by investigating how critical numbers and locations where the second derivative is zero depend on the values of these parameters, we can often accurately describe the shape of the function in terms of the parameters.
- In particular, just as we can created first and second derivative sign charts for a single function, we often can do so for entire families of functions where critical numbers and possible inflection points depend on arbitrary constants. These sign charts then reveal where members of the family are increasing or decreasing, concave up or concave down, and help us to identify relative extremes and inflection points.

Answers

Example 3.2.1

We begin by computing $g'(x)$. By the product rule,
$$g'(x) = ax\frac{d}{dx}\left[e^{-bx}\right] + e^{-bx}\frac{d}{dx}[ax].$$
By applying the chain rule and constant multiple rule, we find that
$$g'(x) = axe^{-bx}(-b) + e^{-bx}(a).$$
To find the critical numbers of g, we solve the equation $g'(x) = 0$. By factoring $g'(x)$, we find
$$0 = ae^{-bx}(-bx+1).$$
Since we are given that $a \neq 0$ and we know that $e^{-bx} \neq 0$ for all values of x, the only way this equation can hold is when $-bx+1 = 0$. Solving for x, we find $x = \frac{1}{b}$, and this is therefore the only critical number of g.

We construct the first derivative sign chart for g that is shown in Figure 3.2.2.

$$g'(x) = ae^{-bx}(1-bx)$$

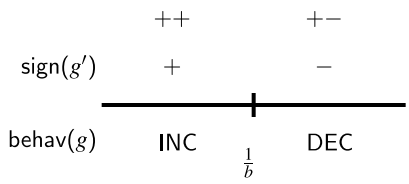


Figure 3.2.2. The first derivative sign chart for $g(x) = axe^{-bx}$.

Because the factor ae^{-bx} is always positive, the sign of g' depends on the linear factor $(1-bx)$, which is positive for $x < \frac{1}{b}$ and negative for $x > \frac{1}{b}$. Hence we can not only conclude that g is always increasing for $x < \frac{1}{b}$ and decreasing for $x > \frac{1}{b}$, but also that g has a global maximum at $(\frac{1}{b}, g(\frac{1}{b}))$ and no local minimum.

We turn next to analyzing the concavity of g. With $g'(x) = -abxe^{-bx} + ae^{-bx}$, we differentiate to find that

$$g''(x) = -abxe^{-bx}(-b) + e^{-bx}(-ab) + ae^{-bx}(-b).$$

Combining like terms and factoring, we now have

$$g''(x) = ab^2xe^{-bx} - 2abe^{-bx} = abe^{-bx}(bx-2).$$

$$g''(x) = abe^{-bx}(bx-2)$$

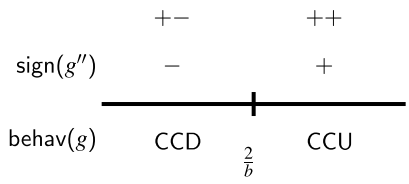


Figure 3.2.3. The second derivative sign chart for $g(x) = axe^{-bx}$.

We observe that abe^{-bx} is always positive, and thus the sign of g'' depends on the sign of $(bx-2)$, which is zero when $x = \frac{2}{b}$. Since b is positive, the value of $(bx-2)$ is negative for $x < \frac{2}{b}$ and positive for $x > \frac{2}{b}$. The sign chart for g'' is shown in Figure 3.2.3. Thus, g is concave down for all $x < \frac{2}{b}$ and concave up for all $x > \frac{2}{b}$.

Finally, we analyze the long term behavior of g by considering two limits. First, we note that

$$\lim_{x\to\infty} g(x) = \lim_{x\to\infty} axe^{-bx} = \lim_{x\to\infty} \frac{ax}{e^{bx}}.$$

This limit has indeterminate form $\frac{\infty}{\infty}$, so we apply L'Hôpital's Rule and find that $\lim_{x\to\infty} g(x) = 0$. In the other direction,

$$\lim_{x\to-\infty} g(x) = \lim_{x\to-\infty} axe^{-bx} = -\infty,$$

because $ax \to -\infty$ and $e^{-bx} \to \infty$ as $x \to -\infty$. Hence, as we move left on its graph, g decreases without bound, while as we move to the right, $g(x) \to 0$.

All of this information now helps us produce the graph of a typical member of this family of functions without using a graphing utility (and without choosing particular values for a and b), as shown in Figure 3.2.4.

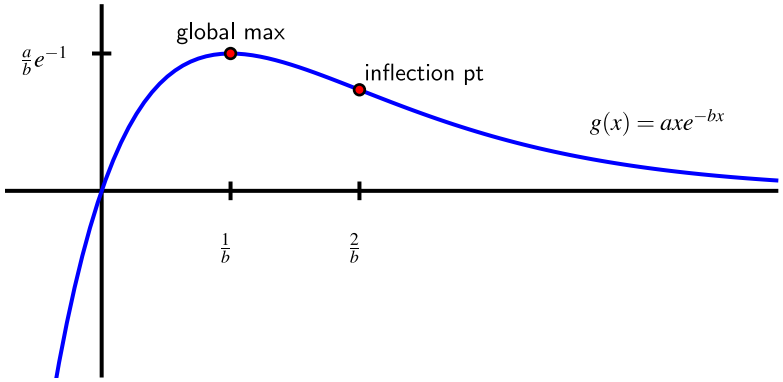


Figure 3.2.4. The graph of $g(x) = axe^{-bx}$.

Note that the value of b controls the horizontal location of the global maximum and the inflection point, as neither depends on a. The value of a affects the vertical stretch of the graph. For example, the global maximum occurs at the point $(\frac{1}{b}, g(\frac{1}{b})) = (\frac{1}{b}, \frac{a}{b}e^{-1})$, so the larger the value of a, the greater the value of the global maximum.

This page titled 3.2: Using Derivatives to Describe Families of Functions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 3.2: Using Derivatives to Describe Families of Functions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

3.3: Global Optimization

Motivating Questions

- What are the differences between finding relative extreme values and global extreme values of a function?
- How is the process of finding the global maximum or minimum of a function over the function's entire domain different from determining the global maximum or minimum on a restricted domain?
- For a function that is guaranteed to have both a global maximum and global minimum on a closed, bounded interval, what are the possible points at which these extreme values occur?

We have seen that we can use the first derivative of a function to determine where the function is increasing or decreasing, and the second derivative to know where the function is concave up or concave down. This information helps us determine the overall shape and behavior of the graph, as well as whether the function has relative extrema.

Remember the difference between a relative maximum and a global maximum: there is a *relative* maximum of f at $x = p$ if $f(p) \geq f(x)$ for all x *near* p, while there is a *global* maximum at p if $f(p) \geq f(x)$ for *all* x in the domain of f.

For instance, in Figure 3.3.1, we see a function f that has a global maximum at $x = c$ and a relative maximum at $x = a$, since $f(c)$ is greater than $f(x)$ for every value of x, while $f(a)$ is only greater than the value of $f(x)$ for x near a. Since the function appears to decrease without bound, f has no global minimum, though clearly f has a relative minimum at $x = b$.

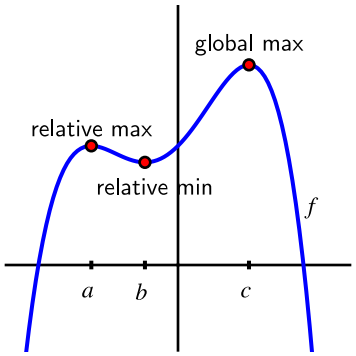


Figure 3.3.1 A function f with a global maximum, but no global minimum.

Our emphasis in this section is on finding the global extreme values of a function (if they exist), either over its entire domain or on some restricted portion.

Preview Activity 3.3.1

Let $f(x) = 2 + \frac{3}{1+(x+1)^2}$.

1. Determine all of the critical numbers of f.
2. Construct a first derivative sign chart for f and thus determine all intervals on which f is increasing or decreasing.
3. Does f have a global maximum? If so, why, and what is its value and where is the maximum attained? If not, explain why.
4. Determine $\lim_{x\to\infty} f(x)$ and $\lim_{x\to-\infty} f(x)$.
5. Explain why $f(x) > 2$ for every value of x.
6. Does f have a global minimum? If so, why, and what is its value and where is the minimum attained? If not, explain why.

Global Optimization

In Figure 3.3.1 and Preview Activity 3.3.1, we were interested in finding the global minimum and global maximum for f on its entire domain. At other times, we might focus on some restriction of the domain.

For example, rather than considering $f(x) = 2 + \frac{3}{1+(x+1)^2}$ for every value of x, perhaps instead we are only interested in those x for which $0 \leq x \leq 4$, and we would like to know which values of x in the interval $[0, 4]$ produce the largest possible and smallest possible values of f. We are accustomed to critical numbers playing a key role in determining the location of extreme values of a function; now, by restricting the domain to an interval, it makes sense that the endpoints of the interval will also be important to consider, as we see in the following activity. When limiting ourselves to a particular interval, we will often refer to the *absolute* maximum or minimum value, rather than the *global* maximum or minimum.

Activity 3.3.2

Let $g(x) = \frac{1}{3}x^3 - 2x + 2$.

1. Find all critical numbers of g that lie in the interval $-2 \leq x \leq 3$.
2. Use a graphing utility to construct the graph of g on the interval $-2 \leq x \leq 3$.
3. From the graph, determine the x-values at which the absolute minimum and absolute maximum of g occur on the interval $[-2, 3]$.
4. How do your answers change if we instead consider the interval $-2 \leq x \leq 2$?
5. What if we instead consider the interval $-2 \leq x \leq 1$?

In Activity 3.3.2, we saw how the absolute maximum and absolute minimum of a function on a closed, bounded interval $[a, b]$, depend not only on the critical numbers of the function, but also on the values of a and b. These observations demonstrate several important facts that hold more generally. First, we state an important result called the Extreme Value Theorem.

The Extreme Value Theorem

If f is a continuous function on a closed interval $[a, b]$, then f attains both an absolute minimum and absolute maximum on $[a, b]$. That is, for some value x_m such that $a \leq x_m \leq b$, it follows that $f(x_m) \leq f(x)$ for all x in $[a, b]$. Similarly, there is a value x_M in $[a, b]$ such that $f(x_M) \geq f(x)$ for all x in $[a, b]$. Letting $m = f(x_m)$ and $M = f(x_M)$, it follows that $m \leq f(x) \leq M$ for all x in $[a, b]$.

The Extreme Value Theorem tells us that on any closed interval $[a, b]$, a continuous function has to achieve both an absolute minimum and an absolute maximum. The theorem does not tell us where these extreme values occur, but rather only that they must exist. As we saw in Activity 3.3.1.2, the only possible locations for relative extremes are at the endpoints of the interval or at a critical number.

Note 3.3.1

Thus, we have the following approach to finding the absolute maximum and minimum of a continuous function f on the interval $[a, b]$:

- find all critical numbers of f that lie in the interval;
- evaluate the function f at each critical number in the interval and at each endpoint of the interval;
- from among those function values, the smallest is the absolute minimum of f on the interval, while the largest is the absolute maximum.

Activity 3.3.3

Find the *exact* absolute maximum and minimum of each function on the stated interval.

1. $h(x) = xe^{-x}$, $[0, 3]$
2. $p(t) = \sin(t) + \cos(t)$, $[-\frac{\pi}{2}, \frac{\pi}{2}]$
3. $q(x) = \frac{x^2}{x-2}$, $[3, 7]$
4. $f(x) = 4 - e^{-(x-2)^2}$, $(-\infty, \infty)$
5. $h(x) = xe^{-ax}$, $[0, \frac{2}{a}]$ $(a > 0)$
6. $f(x) = b - e^{-(x-a)^2}$, $(-\infty, \infty)$, $a, b > 0$

The interval we choose has nearly the same influence on extreme values as the function under consideration. Consider, for instance, the function pictured in Figure 3.3.2.

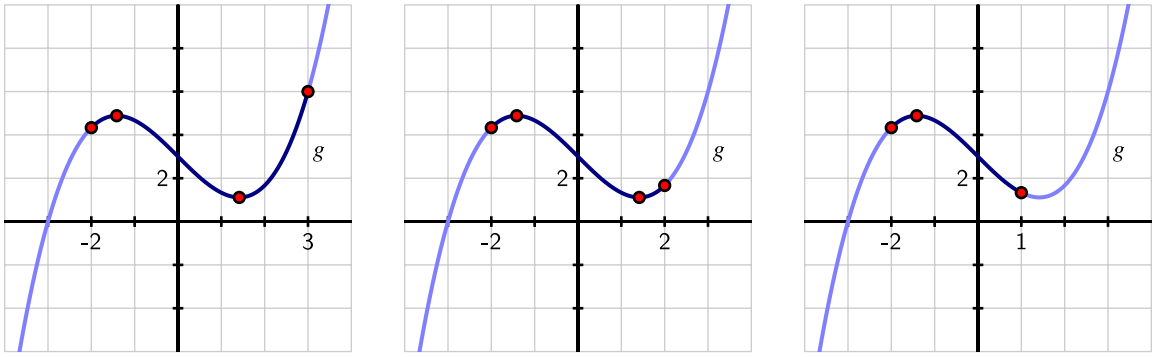


Figure 3.3.2. A function g considered on three different intervals.

In sequence, from left to right, the interval under consideration is changed from $[-2, 3]$ to $[-2, 2]$ to $[-2, 1]$.

- On the interval $[-2, 3]$, there are two critical numbers, with the absolute minimum at one critical number and the absolute maximum at the right endpoint.
- On the interval $[-2, 2]$, both critical numbers are in the interval, with the absolute minimum and maximum at the two critical numbers.
- On the interval $[-2, 1]$, just one critical number lies in the interval, with the absolute maximum at one critical number and the absolute minimum at one endpoint.

Remember to consider only the critical numbers that lie within the interval.

Moving toward applications

We conclude this section with an example of an applied optimization problem. It highlights the role that a closed, bounded domain can play in finding absolute extrema.

Example 3.3.1

A 20 cm piece of wire is cut into two pieces. One piece is used to form a square and the other to form an equilateral triangle. How should the wire be cut to maximize the total area enclosed by the square and triangle? to minimize the area?

Answer available at the end of this section.

Activity 3.3.4

A piece of cardboard that is 10×15 (each measured in inches) is being made into a box without a top. To do so, squares are cut from each corner of the box and the remaining sides are folded up. If the box needs to be at least 1 inch deep and no more than 3 inches deep, what is the maximum possible volume of the box? what is the minimum volume? Justify your answers using calculus.

1. Draw a labeled diagram that shows the given information. What variable should we introduce to represent the choice we make in creating the box? Label the diagram appropriately with the variable, and write a sentence to state what the variable represents.
2. Determine a formula for the function V (that depends on the variable in (a)) that tells us the volume of the box.
3. What is the domain of the function V? That is, what values of x make sense for input? Are there additional restrictions provided in the problem?
4. Determine all critical numbers of the function V.
5. Evaluate V at each of the endpoints of the domain and at any critical numbers that lie in the domain.
6. What is the maximum possible volume of the box? the minimum?

Example 3.3.1 and Activity 3.3.4 illustrate standard steps that we undertake in almost every applied optimization problem: we draw a picture to demonstrate the situation, introduce one or more variables to represent quantities that are changing, find a function that models the quantity to be optimized, and then decide on an appropriate domain for that function. Once that is done, we are in the familiar situation of finding the absolute minimum and maximum of a function over a particular domain, so we apply the calculus ideas that we have been studying to this point in Chapter 3.

Summary

- To find relative extreme values of a function, we use a first derivative sign chart and classify all of the function's critical numbers. If instead we are interested in absolute extreme values, we first decide whether we are considering the entire domain of the function or a particular interval.
- In the case of finding global extremes over the function's entire domain, we again use a first or second derivative sign chart. If we are working to find absolute extremes on a restricted interval, then we first identify all critical numbers of the function that lie in the interval.
- For a continuous function on a closed, bounded interval, the only possible points at which absolute extreme values occur are the critical numbers and the endpoints. Thus, we simply evaluate the function at each endpoint and each critical number in the interval, and compare the results to decide which is largest (the absolute maximum) and which is smallest (the absolute minimum).

Answers

Example 3.3.1

We begin by sketching a picture that illustrates the situation. The variable in the problem is where we decide to cut the wire. We thus label the cut point at a distance x from one end of the wire, and note that the remaining portion of the wire then has length $20 - x$

As shown in Figure 3.3.3, we see that the x cm of wire that is used to form the equilateral triangle with three sides of length $\frac{x}{3}$. For the remaining $20 - x$ cm of wire, the square that results will have each side of length $\frac{20-x}{4}$.

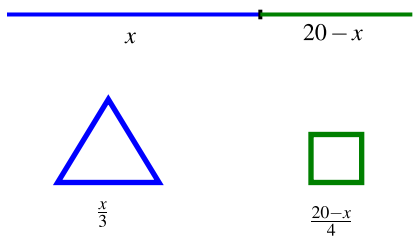


Figure 3.3.3. A 20 cm piece of wire cut into two pieces, one of which forms an equilateral triangle, the other which yields a square.

At this point, we note that there are obvious restrictions on x: in particular, $0 \leq x \leq 20$. In the extreme cases, all of the wire is being used to make just one figure. For instance, if $x = 0$, then all 20 cm of wire are used to make a square that is 5×5.

Now, our overall goal is to find the minimum and maximum areas that can be enclosed. Because the height of an equilateral triangle is $\sqrt{3}$ times half the length of the base, the area of the triangle is

$$A_\Delta = \frac{1}{2}bh = \frac{1}{2}\cdot\frac{x}{3}\cdot\frac{x\sqrt{3}}{6}.$$

The area of the square is $A_\square = s^2 = \left(\frac{20-x}{4}\right)^2$. Therefore, the total area function is

$$A(x) = \frac{\sqrt{3}x^2}{36} + \left(\frac{20-x}{4}\right)^2.$$

Remember that we are considering this function only on the restricted domain $[0, 20]$.

Differentiating $A(x)$, we have

$$A'(x) = \frac{\sqrt{3}x}{18} + 2\left(\frac{20-x}{4}\right)\left(-\frac{1}{4}\right) = \frac{\sqrt{3}}{18}x + \frac{1}{8}x - \frac{5}{2}.$$

When we set $A'(x) = 0$, we find that $x = \frac{180}{4\sqrt{3}+9} \approx 11.3007$ is the only critical number of A in the interval $[0, 20]$.

Evaluating A at the critical number and endpoints, we see that

$$A\left(\frac{180}{4\sqrt{3}+9}\right) = \frac{\sqrt{3}(\frac{180}{4\sqrt{3}+9})^2}{4} + \left(\frac{20-\frac{180}{4\sqrt{3}+9}}{4}\right)^2 \approx 10.8741\ A(0) = 25\ A(20) = \frac{\sqrt{3}}{36}(400) = \frac{100}{9}\sqrt{3} \approx 19.2450$$

Thus, the absolute minimum occurs when $x \approx 11.3007$ and results in the minimum area of approximately 10.8741 square centimeters. The absolute maximum occurs when we invest all of the wire in the square (and none in the triangle), resulting in 25 square centimeters of area. These results are confirmed by a plot of $y = A(x)$ on the interval $[0, 20]$, as shown in Figure 3.3.4.

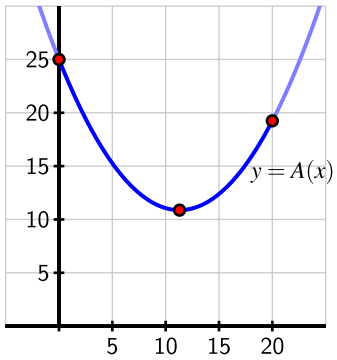


Figure 3.3.4. A plot of the area function from Example 3.3.1.4.

This page titled 3.3: Global Optimization is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 3.3: Global Optimization by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

3.4: Applied Optimization

Motivating Questions

- In a setting where a situation is described for which optimal parameters are sought, how do we develop a function that models the situation and use calculus to find the desired maximum or minimum?

Near the conclusion of Section 3.3, we considered two optimization problems where determining the function to be optimized was part of the problem. In Example 3.3.4, we sought to use a single piece of wire to build an equilateral triangle and square in order to maximize the total combined area enclosed. In the subsequent Activity 3.3.4 we investigated how the volume of a box constructed from a piece of cardboard by removing squares from each corner and folding up the sides depends on the size of the squares removed.

In neither of these problems was a function to optimize explicitly provided. Rather, we first tried to understand the problem by drawing a figure and introducing variables, and then sought to develop a formula for a function that modeled the quantity to be optimized. Once the function was established, we then considered what domain was appropriate. At that point, we were finally ready to apply the ideas of calculus to determine the absolute minimum or maximum.

Throughout what follows in the current section, the primary emphasis is on the reader solving problems. Initially, some substantial guidance is provided, with the problems progressing to require greater independence as we move along.

Preview Activity 3.4.1

According to U.S. postal regulations, the girth plus the length of a parcel sent by mail may not exceed 108 inches, where by "girth" we mean the perimeter of the smallest end. What is the largest possible volume of a rectangular parcel with a square end that can be sent by mail? What are the dimensions of the package of largest volume?

a. Let x represent the length of one side of the square end and y the length of the longer side. Label these quantities appropriately on the image shown in Figure 3.4.1.

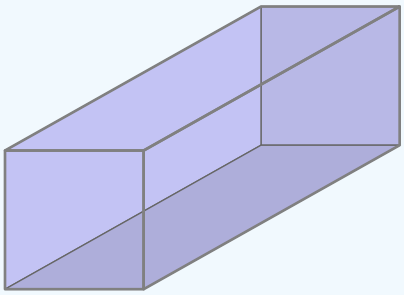


Figure 3.4.1. A rectangular parcel with a square end.

b. What is the quantity to be optimized in this problem? Find a formula for this quantity in terms of x and y.

c. The problem statement tells us that the parcel's girth plus length may not exceed 108 inches. In order to maximize volume, we assume that we will actually need the girth plus length to equal 108 inches. What equation does this produce involving x and y?

d. Solve the equation you found in (c) for one of x or y (whichever is easier).

e. Now use your work in (b) and (d) to determine a formula for the volume of the parcel so that this formula is a function of a single variable.

f. Over what domain should we consider this function? Note that both x and y must be positive; how does the constraint that girth plus length is 108 inches produce intervals of possible values for x and y?

g. Find the absolute maximum of the volume of the parcel on the domain you established in (f) and hence also determine the dimensions of the box of greatest volume. Justify that you've found the maximum using calculus.

More applied optimization problems

Many of the steps in Preview Activity 3.4.1 are ones that we will execute in any applied optimization problem. We briefly summarize those here to provide an overview of our approach in subsequent questions.

Note 3.4.1

- Draw a picture and introduce variables. It is essential to first understand what quantities are allowed to vary in the problem and then to represent those values with variables. Constructing a figure with the variables labeled is almost always an essential first step. Sometimes drawing several diagrams can be especially helpful to get a sense of the situation. A nice example of this can be seen at http://gvsu.edu/s/99, where the choice of where to bend a piece of wire into the shape of a rectangle determines both the rectangle's shape and area.
- Identify the quantity to be optimized as well as any key relationships among the variable quantities. Essentially this step involves writing equations that involve the variables that have been introduced: one to represent the quantity whose minimum or maximum is sought, and possibly others that show how multiple variables in the problem may be interrelated.
- Determine a function of a single variable that models the quantity to be optimized; this may involve using other relationships among variables to eliminate one or more variables in the function formula. For example, in Preview Activity 3.4.11, we initially found that $V = x^2y$, but then the additional relationship that $4x + y = 108$ (girth plus length equals 108 inches) allows us to relate x and y and thus observe equivalently that $y = 108 - 4x$. Substituting for y in the volume equation yields $V(x) = x^2(108 - 4x)$, and thus we have written the volume as a function of the single variable x.
- Decide the domain on which to consider the function being optimized. Often the physical constraints of the problem will limit the possible values that the independent variable can take on. Thinking back to the diagram describing the overall situation and any relationships among variables in the problem often helps identify the smallest and largest values of the input variable.
- Use calculus to identify the absolute maximum and/or minimum of the quantity being optimized. This always involves finding the critical numbers of the function first. Then, depending on the domain, we either construct a first derivative sign chart (for an open or unbounded interval) or evaluate the function at the endpoints and critical numbers (for a closed, bounded interval), using ideas we've studied so far in Chapter 3.
- Finally, we make certain we have answered the question: does the question seek the absolute maximum of a quantity, or the values of the variables that produce the maximum? That is, finding the absolute maximum volume of a parcel is different from finding the dimensions of the parcel that produce the maximum.

Activity 3.4.2

A soup can in the shape of a right circular cylinder is to be made from two materials. The material for the side of the can costs \$0.015 per square inch and the material for the lids costs \$0.027 per square inch. Suppose that we desire to construct a can that has a volume of 16 cubic inches. What dimensions minimize the cost of the can?

1. Draw a picture of the can and label its dimensions with appropriate variables.
2. Use your variables to determine expressions for the volume, surface area, and cost of the can.
3. Determine the total cost function as a function of a single variable. What is the domain on which you should consider this function?
4. Find the absolute minimum cost and the dimensions that produce this value.

Familiarity with common geometric formulas is particularly helpful in problems such as the one in Activity 3.4.12. Sometimes those involve perimeter, area, volume, or surface area. At other times, the constraints of a problem introduce right triangles (where the Pythagorean Theorem applies) or other functions whose formulas provide relationships among the variables.

Activity 3.4.3

A hiker starting at a point P on a straight road walks east towards point Q, which is on the road and 3 kilometers from point P.

Two kilometers due north of point Q is a cabin. The hiker will walk down the road for a while, at a pace of 8 kilometers per hour. At some point Z between P and Q, the hiker leaves the road and makes a straight line towards the cabin through the woods, hiking at a pace of 3 kph, as pictured in Figure 3.4.2. In order to minimize the time to go from P to Z to the cabin, where should the hiker turn into the forest?

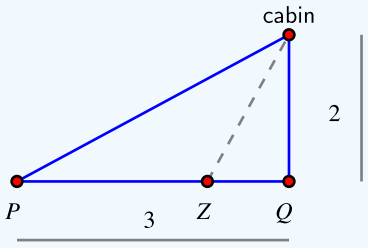


Figure 3.4.2. A hiker walks from P to Z to the cabin, as pictured.

In more geometric problems, we often use curves or functions to provide natural constraints. For instance, we could investigate which isosceles triangle that circumscribes a unit circle has the smallest area, which you can explore for yourself at http://gvsu.edu/s/9b. Or similarly, for a region bounded by a parabola, we might seek the rectangle of largest area that fits beneath the curve, as shown at http://gvsu.edu/s/9c. The next activity is similar to the latter problem.

Activity 3.4.4

Consider the region in the x-y plane that is bounded by the x-axis and the function $f(x) = 25 - x^2$. Construct a rectangle whose base lies on the x-axis and is centered at the origin, and whose sides extend vertically until they intersect the curve $y = 25 - x^2$. Which such rectangle has the maximum possible area? Which such rectangle has the greatest perimeter? Which has the greatest combined perimeter and area? (Challenge: answer the same questions in terms of positive parameters a and b for the function $f(x) = b - ax^2$.)

Activity 3.4.5

A trough is being constructed by bending a 4×24 (measured in feet) rectangular piece of sheet metal.

Two symmetric folds 2 feet apart will be made parallel to the longest side of the rectangle so that the trough has cross-sections in the shape of a trapezoid, as pictured in Figure 3.4.3. At what angle should the folds be made to produce the trough of maximum volume?

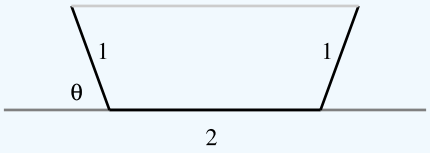


Figure 3.4.3. A cross-section of the trough formed by folding to an angle of θ.

Summary

- While there is no single algorithm that works in every situation where optimization is used, in most of the problems we consider, the following steps are helpful: draw a picture and introduce variables; identify the quantity to be optimized and find relationships among the variables; determine a function of a single variable that models the quantity to be optimized; decide the domain on which to consider the function being optimized; use calculus to identify the absolute maximum and/or minimum of the quantity being optimized.

This page titled 3.4: Applied Optimization is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 3.4: Applied Optimization by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

3.5: Related Rates

Motivating Questions

- If two quantities that are related, such as the radius and volume of a spherical balloon, are both changing as implicit functions of time, how are their rates of change related? That is, how does the relationship between the values of the quantities affect the relationship between their respective derivatives with respect to time?

In most of our applications of the derivative so far, we have been interested in the instantaneous rate at which one variable, say y, changes with respect to another, say x, leading us to compute and interpret $\frac{dy}{dx}$. We next consider situations where several variable quantities are related, but where each quantity is implicitly a function of time, which will be represented by the variable t. Through knowing how the quantities are related, we will be interested in determining how their respective rates of change with respect to time are related.

For example, suppose that air is being pumped into a spherical balloon so that its volume increases at a constant rate of 20 cubic inches per second. Since the balloon's volume and radius are related, by knowing how fast the volume is changing, we ought to be able to discover how fast the radius is changing. We are interested in questions such as: can we determine how fast the radius of the balloon is increasing at the moment the balloon's diameter is 12 inches?

Preview Activity 3.5.1

A spherical balloon is being inflated at a constant rate of 20 cubic inches per second. How fast is the radius of the balloon changing at the instant the balloon's diameter is 12 inches? Is the radius changing more rapidly when $d = 12$ or when $d = 16$? Why?

1. Draw several spheres with different radii, and observe that as volume changes, the radius, diameter, and surface area of the balloon also change.
2. Recall that the volume of a sphere of radius r is $V = \frac{4}{3}\pi r^3$. Note well that in the setting of this problem, *both* V and r are changing as time t changes, and thus both V and r may be viewed as implicit functions of t, with respective derivatives $\frac{dV}{dt}$ and $\frac{dr}{dt}$. Differentiate both sides of the equation $V = \frac{4}{3}\pi r^3$ with respect to t (using the chain rule on the right) to find a formula for $\frac{dV}{dt}$ that depends on both r and $\frac{dr}{dt}$.
3. At this point in the problem, by differentiating we have "related the rates" of change of V and r. Recall that we are given in the problem that the balloon is being inflated at a constant *rate* of 20 cubic inches per second. Is this rate the value of $\frac{dr}{dt}$ or $\frac{dV}{dt}$? Why?
4. From part (c), we know the value of $\frac{dV}{dt}$ at every value of t. Next, observe that when the diameter of the balloon is 12, we know the value of the radius. In the equation $\frac{dV}{dt} = 4\pi r^2 \frac{dr}{dt}$, substitute these values for the relevant quantities and solve for the remaining unknown quantity, which is $\frac{dr}{dt}$. How fast is the radius changing at the instant $d = 12$?
5. How is the situation different when $d = 16$? When is the radius changing more rapidly, when $d = 12$ or when $d = 16$?

Related Rates Problems

In problems where two or more quantities can be related to one another, and all of the variables involved are implicitly functions of time, t, we are often interested in how their rates are related; we call these *related rates* problems. Once we have an equation establishing the relationship among the variables, we differentiate implicitly with respect to time to find connections among the rates of change.

Example 3.5.1

Sand is being dumped by a conveyor belt onto a pile so that the sand forms a right circular cone, as pictured in Figure 3.5.1. How are the instantaneous rates of change of the sand's volume, height, and radius related to one another?

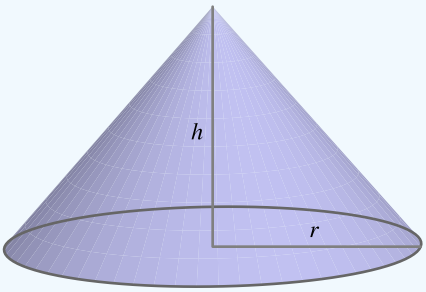


Figure 3.5.1. A conical pile of sand.

Answer available at the end of this section.

If we are given sufficient additional information, we may then find the value of one or more of these rates of change at a specific point in time.

Example 3.5.2

In the setting of Example 3.5.1, suppose we also know the following: (a) sand falls from the conveyor in such a way that the height of the pile is always half the radius, and (b) sand falls from the conveyor belt at a constant rate of 10 cubic feet per minute. How fast is the height of the sandpile changing at the moment the radius is 4 feet?

Answer available at the end of this section.

Note the difference between the notations $\frac{dr}{dt}$ and $\frac{dr}{dt}\Big|_{r=4}$. The former represents the rate of change of r with respect to t at an arbitrary value of t, while the latter is the rate of change of r with respect to t at a particular moment, the moment when $r = 4$.

Had we known that $h = \frac{1}{2}r$ at the beginning of Example 3.5.1, we could have immediately simplified our work by writing V solely in terms of r to have

$$V = \frac{1}{3}\pi r^2\left(\frac{1}{2}h\right) = \frac{1}{6}\pi r^3.$$

From this last equation, differentiating with respect to t implies

$$\frac{dV}{dt} = \frac{1}{2}\pi r^2\frac{dr}{dt},$$

from which the same conclusions can be made.

Our work with the sandpile problem above is similar in many ways to our approach in Preview Activity 3.5.11, and these steps are typical of most related rates problems. In certain ways, they also resemble work we do in applied optimization problems, and here we summarize the main approach for consideration in subsequent problems.

Note 3.5.1

- Identify the quantities in the problem that are changing and choose clearly defined variable names for them. Draw one or more figures that clearly represent the situation.
- Determine all rates of change that are known or given and identify the rate(s) of change to be found.
- Find an equation that relates the variables whose rates of change are known to those variables whose rates of change are to be found.
- Differentiate implicitly with respect to t to relate the rates of change of the involved quantities.

- Evaluate the derivatives and variables at the information relevant to the instant at which a certain rate of change is sought. Use proper notation to identify when a derivative is being evaluated at a particular instant, such as $\frac{dr}{dt}\big|_{r=4}$.

When identifying variables and drawing a picture, it is important to think about the dynamic ways in which the quantities change. Sometimes a sequence of pictures can be helpful; for some pictures that can be easily modified as applets built in Geogebra, see the following links,[1] which represent

- how a circular oil slick's area grows as its radius increases http://gvsu.edu/s/9n;
- how the location of the base of a ladder and its height along a wall change as the ladder slides http://gvsu.edu/s/9o;
- how the water level changes in a conical tank as it fills with water at a constant rate http://gvsu.edu/s/9p (compare the problem in Activity 3.5.12);
- how a skateboarder's shadow changes as he moves past a lamppost http://gvsu.edu/s/9q.

We again refer to the work of Prof. Marc Renault of Shippensburg University, found at http://gvsu.edu/s/5p.

Drawing well-labeled diagrams and envisioning how different parts of the figure change is a key part of understanding related rates problems and being successful at solving them.

Activity 3.5.2

A water tank has the shape of an inverted circular cone (point down) with a base of radius 6 feet and a depth of 8 feet. Suppose that water is being pumped into the tank at a constant instantaneous rate of 4 cubic feet per minute.

1. Draw a picture of the conical tank, including a sketch of the water level at a point in time when the tank is not yet full. Introduce variables that measure the radius of the water's surface and the water's depth in the tank, and label them on your figure.
2. Say that r is the radius and h the depth of the water at a given time, t. What equation relates the radius and height of the water, and why?
3. Determine an equation that relates the volume of water in the tank at time t to the depth h of the water at that time.
4. Through differentiation, find an equation that relates the instantaneous rate of change of water volume with respect to time to the instantaneous rate of change of water depth at time t.
5. Find the instantaneous rate at which the water level is rising when the water in the tank is 3 feet deep.
6. When is the water rising most rapidly: at $h = 3$, $h = 4$, or $h = 5$?

Recognizing which geometric relationships are relevant in a given problem is often the key to finding the function to optimize. For instance, although the problem in Activity 3.5.2 is about a conical tank, the most important fact is that there are two similar right triangles involved. In another setting, we might use the Pythagorean Theorem to relate the legs of the triangle. But in the conical tank, the fact that the water fills the tank so that that the ratio of radius to depth is constant turns out to be the important relationship. In other situations where a changing angle is involved, trigonometric functions may provide the means to find relationships among various parts of the triangle.

Activity 3.5.3

A television camera is positioned 4000 feet from the base of a rocket launching pad. The angle of elevation of the camera has to change at the correct rate in order to keep the rocket in sight. In addition, the auto-focus of the camera has to take into account the increasing distance between the camera and the rocket. We assume that the rocket rises vertically. (A similar problem is discussed and pictured dynamically at http://gvsu.edu/s/9t. Exploring the applet at the link will be helpful to you in answering the questions that follow.)

1. Draw a figure that summarizes the given situation. What parts of the picture are changing? What parts are constant? Introduce appropriate variables to represent the quantities that are changing.
2. Find an equation that relates the camera's angle of elevation to the height of the rocket, and then find an equation that relates the instantaneous rate of change of the camera's elevation angle to the instantaneous rate of change of the rocket's height (where all rates of change are with respect to time).

3. Find an equation that relates the distance from the camera to the rocket to the rocket's height, as well as an equation that relates the instantaneous rate of change of distance from the camera to the rocket to the instantaneous rate of change of the rocket's height (where all rates of change are with respect to time).
4. Suppose that the rocket's speed is 600 ft/sec at the instant it has risen 3000 feet. How fast is the distance from the television camera to the rocket changing at that moment? If the camera is following the rocket, how fast is the camera's angle of elevation changing at that same moment?
5. If from an elevation of 3000 feet onward the rocket continues to rise at 600 feet/sec, will the rate of change of distance with respect to time be greater when the elevation is 4000 feet than it was at 3000 feet, or less? Why?

In addition to finding instantaneous rates of change at particular points in time, we can often make more general observations about how particular rates themselves will change over time. For instance, when a conical tank is filling with water at a constant rate, it seems obvious that the depth of the water should increase more slowly over time. Note how carefully we must phrase the relationship: we mean to say that while the depth, h, of the water is increasing, its rate of change, $\frac{dh}{dt}$, is decreasing (both as a function of t and as a function of h). We make this observation by solving the equation that relates the various rates for one particular rate, without substituting any particular values for known variables or rates. For instance, in the conical tank problem in Activity 3.5.2, we established that

$$\frac{dV}{dt} = \frac{1}{16}\pi h^2 \frac{dh}{dt},$$

and hence

$$\frac{dh}{dt} = \frac{16}{\pi h^2}\frac{dV}{dt}.$$

Provided that $\frac{dV}{dt}$ is constant, it is immediately apparent that as h gets larger, $\frac{dh}{dt}$ will get smaller but remain positive. Hence, the depth of the water is increasing at a decreasing rate.

Activity 3.5.4

As pictured in the applet at http://gvsu.edu/s/9q, a skateboarder who is 6 feet tall rides under a 15 foot tall lamppost at a constant rate of 3 feet per second. We are interested in understanding how fast his shadow is changing at various points in time.

1. Draw an appropriate right triangle that represents a snapshot in time of the skateboarder, lamppost, and his shadow. Let x denote the horizontal distance from the base of the lamppost to the skateboarder and s represent the length of his shadow. Label these quantities, as well as the skateboarder's height and the lamppost's height on the diagram.
2. Observe that the skateboarder and the lamppost represent parallel line segments in the diagram, and thus similar triangles are present. Use similar triangles to establish an equation that relates x and s.
3. Use your work in (b) to find an equation that relates $\frac{dx}{dt}$ and $\frac{ds}{dt}$.
4. At what rate is the length of the skateboarder's shadow increasing at the instant the skateboarder is 8 feet from the lamppost?
5. As the skateboarder's distance from the lamppost increases, is his shadow's length increasing at an increasing rate, increasing at a decreasing rate, or increasing at a constant rate?
6. Which is moving more rapidly: the skateboarder or the tip of his shadow? Explain, and justify your answer.

In the first three activities of this section, we provided guided instruction to build a solution in a step by step way. For the closing activity and the following exercises, most of the detailed work is left to the reader.

Activity 3.5.5

A baseball diamond is $90'$ square. A batter hits a ball along the third base line and runs to first base. At what rate is the distance between the ball and first base changing when the ball is halfway to third base, if at that instant the ball is traveling 100 feet/sec? At what rate is the distance between the ball and the runner changing at the same instant, if at the same instant the runner is $1/8$ of the way to first base running at 30 feet/sec?

Summary

- When two or more related quantities are changing as implicit functions of time, their rates of change can be related by implicitly differentiating the equation that relates the quantities themselves. For instance, if the sides of a right triangle are all changing as functions of time, say having lengths x, y, and z, then these quantities are related by the Pythagorean Theorem: $x^2 + y^2 = z^2$. It follows by implicitly differentiating with respect to t that their rates are related by the equation

$$2x\frac{dx}{dt} + 2y\frac{dy}{dt} = 2z\frac{dz}{dt},$$

  so that if we know the values of x, y, and z at a particular time, as well as two of the three rates, we can deduce the value of the third.

Answers

Example 3.5.1

As sand falls from the conveyor belt, several features of the sand pile will change: the volume of the pile will grow, the height will increase, and the radius will get bigger, too. All of these quantities are related to one another, and the rate at which each is changing is related to the rate at which sand falls from the conveyor.

We begin by identifying which variables are changing and how they are related. In this problem, we observe that the radius and height of the pile are related to its volume by the standard equation for the volume of a cone,

$$V = \frac{1}{3}\pi r^2 h.$$

Viewing each of V, r, and h as functions of t, we differentiate implicitly to arrive at an equation that relates their respective rates of change. Taking the derivative of each side of the equation with respect to t, we find

$$\frac{d}{dt}[V] = \frac{d}{dt}\left[\frac{1}{3}\pi r^2 h\right].$$

On the left, $\frac{d}{dt}[V]$ is simply $\frac{dV}{dt}$. On the right, the situation is more complicated, as both r and h are implicit functions of t. Hence we need the product and chain rules. We find that

$$\begin{aligned}\frac{dV}{dt} &= \frac{d}{dt}\left[\frac{1}{3}\pi r^2 h\right] \\ &= \frac{1}{3}\pi r^2 \frac{d}{dt}[h] + \frac{1}{3}\pi h \frac{d}{dt}[r^2] \\ &= \frac{1}{3}\pi r^2 \frac{dh}{dt} + \frac{1}{3}\pi h 2r\frac{dr}{dt}\end{aligned}$$

(Note particularly how we are using ideas from Section 2.7 on implicit differentiation. There we found that when y is an implicit function of x, $\frac{d}{dx}[y^2] = 2y\frac{dy}{dx}$. The same principles are applied here when we compute $\frac{d}{dt}[r^2] = 2r\frac{dr}{dt}$.)

The equation

$$\frac{dV}{dt} = \frac{1}{3}\pi r^2 \frac{dh}{dt} + \frac{2}{3}\pi r h \frac{dr}{dt},$$

relates the rates of change of V, h, and r.

Example 3.5.2

The information that the height is always half the radius tells us that for all values of t, $h = \frac{1}{2}r$. Differentiating with respect to t, it follows that $\frac{dh}{dt} = \frac{1}{2}\frac{dr}{dt}$. These relationships enable us to relate $\frac{dV}{dt}$ to just one of r or h. Substituting the expressions involving r and $\frac{dr}{dt}$ for h and $\frac{dh}{dt}$, we now have that

$$\frac{dV}{dt} = \frac{1}{3}\pi r^2 \cdot \frac{1}{2}\frac{dr}{dt} + \frac{2}{3}\pi r \cdot \frac{1}{2}r \cdot \frac{dr}{dt}. \tag{3.5.11}$$

Since sand falls from the conveyor at the constant rate of 10 cubic feet per minute, the value of $\frac{dV}{dt}$, the rate at which the volume of the sand pile changes, is $\frac{dV}{dt} = 10$ ft^3/min. We are interested in how fast the height of the pile is changing at the instant when $r = 4$, so we substitute $r = 4$ and $\frac{dV}{dt} = 10$ into Equation (3.5.11), to find

$$10 = \frac{1}{3}\pi 4^2 \cdot \frac{1}{2}\frac{dr}{dt}\Big|_{r=4} + \frac{2}{3}\pi 4 \cdot \frac{1}{2}4 \cdot \frac{dr}{dt}\Big|_{r=4} = \frac{8}{3}\pi \frac{dr}{dt}\Big|_{r=4} + \frac{16}{3}\pi \frac{dr}{dt}\Big|_{r=4}.$$

Only the value of $\frac{dr}{dt}\big|_{r=4}$ remains unknown. We combine like terms on the right side of the equation above to get $10 = 8\pi \frac{dr}{dt}\big|_{r=4}$, and solve for $\frac{dr}{dt}\big|_{r=4}$ to find

$$\frac{dr}{dt}\Big|_{r=4} = \frac{10}{8\pi} \approx 0.39789$$

feet per second. Because we were interested in how fast the height of the pile was changing at this instant, we want to know $\frac{dh}{dt}$ when $r = 4$. Since $\frac{dh}{dt} = \frac{1}{2}\frac{dr}{dt}$ for all values of t, it follows

$$\frac{dh}{dt}\Big|_{r=4} = \frac{5}{8\pi} \approx 0.19894 \text{ ft/min}.$$

This page titled 3.5: Related Rates is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 3.5: Related Rates by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

3.6: Using Derivatives (Exercises)

3.1: Using Derivatives to Identify Extreme Values

1. Finding critical points and inflection points

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Finding inflection points

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Matching graphs of f, f', f''

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4.

This problem concerns a function about which the following information is known:

- f is a differentiable function defined at every real number x
- $f(0) = -1/2$
- $y = f'(x)$ has its graph given at center in Figure 3.1.17

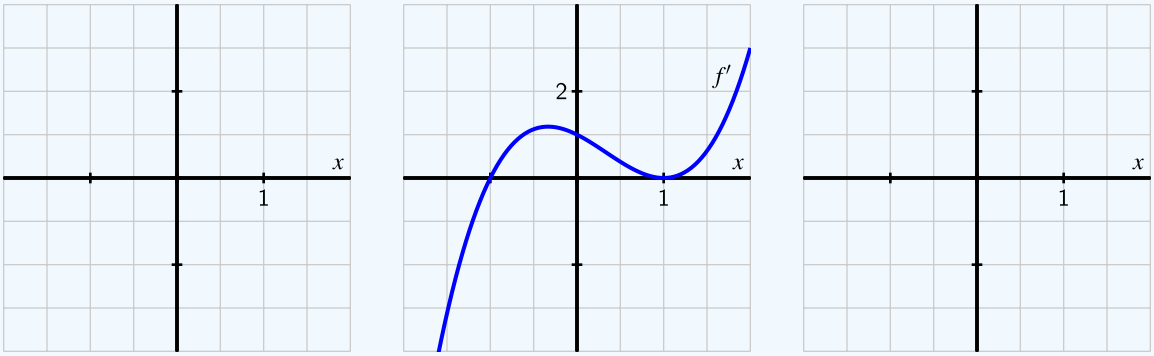


Figure 3.1.17. At center, a graph of $y = f'(x)$; at left, axes for plotting $y = f(x)$; at right, axes for plotting $y = f''(x)$.

1. Construct a first derivative sign chart for f. Clearly identify all critical numbers of f, where f is increasing and decreasing, and where f has local extrema.
2. On the right-hand axes, sketch an approximate graph of $y = f''(x)$.
3. Construct a second derivative sign chart for f. Clearly identify where f is concave up and concave down, as well as all inflection points.
4. On the left-hand axes, sketch a possible graph of $y = f(x)$.

5.

Suppose that g is a differentiable function and $g'(2) = 0$. In addition, suppose that on $1 < x < 2$ and $2 < x < 3$ it is known that $g'(x)$ is positive.

1. Does g have a local maximum, local minimum, or neither at $x = 2$? Why?
2. Suppose that $g''(x)$ exists for every x such that $1 < x < 3$. Reasoning graphically, describe the behavior of $g''(x)$ for x -values near 2.
3. Besides being a critical number of g, what is special about the value $x = 2$ in terms of the behavior of the graph of g?

6.

Suppose that h is a differentiable function whose first derivative is given by the graph in Figure 3.1.18.

1. How many real number solutions can the equation $h(x) = 0$ have? Why?
2. If $h(x) = 0$ has two distinct real solutions, what can you say about the signs of the two solutions? Why?
3. Assume that $\lim_{x\to\infty} h'(x) = 3$, as appears to be indicated in Figure 3.1.18. How will the graph of $y = h(x)$ appear as $x \to \infty$? Why?
4. Describe the concavity of $y = h(x)$ as fully as you can from the provided information.

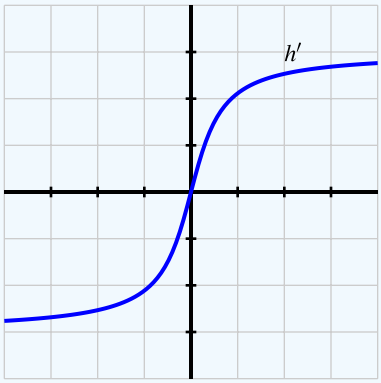


Figure 3.1.18. The graph of $y = h'(x)$.

7.

Let p be a function whose second derivative is $p''(x) = (x+1)(x-2)e^{-x}$.

1. Construct a second derivative sign chart for p and determine all inflection points of p.
2. Suppose you also know that $x = \frac{\sqrt{5}-1}{2}$ is a critical number of p. Does p have a local minimum, local maximum, or neither at $x = \frac{\sqrt{5}-1}{2}$? Why?
3. If the point $(2, \frac{12}{e^2})$ lies on the graph of $y = p(x)$ and $p'(2) = -\frac{5}{e^2}$, find the equation of the tangent line to $y = p(x)$ at the point where $x = 2$. Does the tangent line lie above the curve, below the curve, or neither at this value? Why?

3.2: Using Derivatives to Describe Families of Functions

1. Drug dosage with a parameter

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Using the graph of g'

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3

Consider the one-parameter family of functions given by $p(x) = x^3 - ax^2$, where $a > 0$.

1. Sketch a plot of a typical member of the family, using the fact that each is a cubic polynomial with a repeated zero at $x = 0$ and another zero at $x = a$.
2. Find all critical numbers of p.
3. Compute p'' and find all values for which $p''(x) = 0$. Hence construct a second derivative sign chart for p.
4. Describe how the location of the critical numbers and the inflection point of p change as a changes. That is, if the value of a is increased, what happens to the critical numbers and inflection point?

4

Let $q(x) = \frac{e^{-x}}{x-c}$ be a one-parameter family of functions where $c > 0$.

1. Explain why q has a vertical asymptote at $x = c$.
2. Determine $\lim_{x\to\infty} q(x)$ and $\lim_{x\to-\infty} q(x)$.
3. Compute $q'(x)$ and find all critical numbers of q.
4. Construct a first derivative sign chart for q and determine whether each critical number leads to a local minimum, local maximum, or neither for the function q.
5. Sketch a typical member of this family of functions with important behaviors clearly labeled.

5

Let $E(x) = e^{-\frac{(x-m)^2}{2s^2}}$, where m is any real number and s is a positive real number.

1. Compute $E'(x)$ and hence find all critical numbers of E.
2. Construct a first derivative sign chart for E and classify each critical number of the function as a local minimum, local maximum, or neither.
3. It can be shown that $E''(x)$ is given by the formula

$$E''(x) = e^{-\frac{(x-m)^2}{2s^2}} \left(\frac{(x-m)^2 - s^2}{s^4} \right).$$

   Find all values of x for which $E''(x) = 0$.

4. Determine $\lim_{x\to\infty} E(x)$ and $\lim_{x\to-\infty} E(x)$.
5. Construct a labeled graph of a typical function E that clearly shows how important points on the graph of $y = E(x)$ depend on m and s.

3.3: Global Optimization

1

Based on the given information about each function, decide whether the function has global maximum, a global minimum, neither, both, or that it is not possible to say without more information. Assume that each function is twice differentiable and defined for all real numbers, unless noted otherwise. In each case, write one sentence to explain your conclusion.

1. f is a function such that $f''(x) < 0$ for every x.
2. g is a function with two critical numbers a and b (where $a < b$), and $g'(x) < 0$ for $x < a$, $g'(x) < 0$ for $a < x < b$, and $g'(x) > 0$ for $x > b$.
3. h is a function with two critical numbers a and b (where $a < b$), and $h'(x) < 0$ for $x < a$, $h'(x) > 0$ for $a < x < b$, and $h'(x) < 0$ for $x > b$. In addition, $\lim_{x\to\infty} h(x) = 0$ and $\lim_{x\to-\infty} h(x) = 0$.
4. p is a function differentiable everywhere except at $x = a$ and $p''(x) > 0$ for $x < a$ and $p''(x) < 0$ for $x > a$.

2

For each family of functions that depends on one or more parameters, determine the function's absolute maximum and absolute minimum on the given interval.

1. $p(x) = x^3 - a^2x$, $[0, a]$ $(a > 0)$
2. $r(x) = axe^{-bx}$, $[\frac{1}{2b}, \frac{2}{b}]$ $(a > 0, b > 1)$
3. $w(x) = a(1 - e^{-bx})$, $[b, 3b]$ $(a, b > 0)$
4. $s(x) = \sin(kx)$, $\left[\frac{\pi}{3k}, \frac{5\pi}{6k}\right]$ $(k > 0)$

3

For each of the functions described below (each continuous on $[a, b]$), state the location of the function's absolute maximum and absolute minimum on the interval $[a, b]$, or say there is not enough information provided to make a conclusion. Assume that any critical numbers mentioned in the problem statement represent all of the critical numbers the function has in $[a, b]$. In each case, write one sentence to explain your answer.

1. $f'(x) \leq 0$ for all x in $[a, b]$
2. g has a critical number at c such that $a < c < b$ and $g'(x) > 0$ for $x < c$ and $g'(x) < 0$ for $x > c$
3. $h(a) = h(b)$ and $h''(x) < 0$ for all x in $[a, b]$
4. $p(a) > 0$, $p(b) < 0$, and for the critical number c such that $a < c < b$, $p'(x) < 0$ for $x < c$ and $p'(x) > 0$ for $x > c$

4

Let $s(t) = 3\sin(2(t - \frac{\pi}{6})) + 5$. Find the exact absolute maximum and minimum of s on the provided intervals by testing the endpoints and finding and evaluating all relevant critical numbers of s.

1. $[\frac{\pi}{6}, \frac{7\pi}{6}]$
2. $[0, \frac{\pi}{2}]$
3. $[0, 2\pi]$
4. $[\frac{\pi}{3}, \frac{5\pi}{6}]$

3.4: Applied Optimization

1. Maximizing the volume of a box

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Minimizing the cost of a container

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Maximizing area contained by a fence

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Minimizing the area of a poster

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Maximizing the area of a rectangle

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6

A rectangular box with a square bottom and closed top is to be made from two materials. The material for the side costs $1.50 per square foot and the material for the top and bottom costs $3.00 per square foot. If you are willing to spend $15 on the box, what is the largest volume it can contain? Justify your answer completely using calculus.

7

A farmer wants to start raising cows, horses, goats, and sheep, and desires to have a rectangular pasture for the animals to graze in. However, no two different kinds of animals can graze together. In order to minimize the amount of fencing she will need, she has decided to enclose a large rectangular area and then divide it into four equally sized pens by adding three segments of fence inside the large rectangle that are parallel to two existing sides. She has decided to purchase 7500 ft of fencing. What is the maximum possible area that each of the four pens will enclose?

8

Two vertical poles of heights 60 ft and 80 ft stand on level ground, with their bases 100 ft apart. A cable that is stretched from the top of one pole to some point on the ground between the poles, and then to the top of the other pole. What is the minimum possible length of cable required? Justify your answer completely using calculus.

9

A company is designing propane tanks that are cylindrical with hemispherical ends. Assume that the company wants tanks that will hold 1000 cubic feet of gas, and that the ends are more expensive to make, costing $5 per square foot, while the cylindrical barrel between the ends costs $2 per square foot. Use calculus to determine the minimum cost to construct such a tank.

3.5: Related Rates

1. Height of a conical pile of gravel

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. Movement of a shadow

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. A leaking conical tank

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4

A sailboat is sitting at rest near its dock. A rope attached to the bow of the boat is drawn in over a pulley that stands on a post on the end of the dock that is 5 feet higher than the bow. If the rope is being pulled in at a rate of 2 feet per second, how fast is the boat approaching the dock when the length of rope from bow to pulley is 13 feet?

5

A swimming pool is 60 feet long and 25 feet wide. Its depth varies uniformly from 3 feet at the shallow end to 15 feet at the deep end, as shown in the Figure 3.5.5.

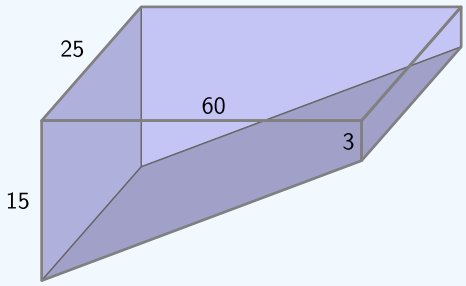


Figure 3.5.5. The swimming pool.

Suppose the pool has been emptied and is now being filled with water at a rate of 800 cubic feet per minute. At what rate is the depth of water (measured at the deepest point of the pool) increasing when it is 5 feet deep at that end? Over time, describe how the depth of the water will increase: at an increasing rate, at a decreasing rate, or at a constant rate. Explain.

6

A baseball diamond is a square with sides 90 feet long. Suppose a baseball player is advancing from second to third base at the rate of 24 feet per second, and an umpire is standing on home plate. Let θ be the angle between the third baseline and the line of sight from the umpire to the runner. How fast is θ changing when the runner is 30 feet from third base?

7

Sand is being dumped off a conveyor belt onto a pile in such a way that the pile forms in the shape of a cone whose radius is always equal to its height. Assuming that the sand is being dumped at a rate of 10 cubic feet per minute, how fast is the height of the pile changing when there are 1000 cubic feet on the pile?

This page titled 3.6: Using Derivatives (Exercises) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 3.E: Using Derivatives (Exercises) has no license indicated.
- 3.1: Using Derivatives to Identify Extreme Values by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 3.2: Using Derivatives to Describe Families of Functions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 3.3: Global Optimization by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 3.4: Applied Optimization by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 3.5: Related Rates by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

4: THE DEFINITE INTEGRAL

4.1: Determining Distance Traveled from Velocity

4.2: Riemann Sums

4.3: The Definite Integral

4.4: The Fundamental Theorem of Calculus

4.5: The Definite Integral (Exercises)

This page titled 4: The Definite Integral is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 4: The Definite Integral by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

4.1: Determining Distance Traveled from Velocity

Motivating Questions

- If we know the velocity of a moving body at every point in a given interval, can we determine the distance the object has traveled on the time interval?
- How is the problem of finding distance traveled related to finding the area under a certain curve?
- What does it mean to antidifferentiate a function and why is this process relevant to finding distance traveled?
- If velocity is negative, how does this impact the problem of finding distance traveled?

In the first section of the text, we considered a moving object with known position at time t, namely, a tennis ball tossed into the air with height s (in feet) at time t (in seconds) given by $s(t) = 64 - 16(t-1)^2$. We investigated the average velocity of the ball on an interval $[a, b]$, computed by the difference quotient $\frac{s(b)-s(a)}{b-a}$. We found that we could determine the instantaneous velocity of the ball at time t by taking the derivative of the position function,

$$s'(t) = \lim_{h \to 0} \frac{s(t+h) - s(t)}{h}.$$

Thus, if its position function is differentiable, we can find the velocity of a moving object at any point in time.

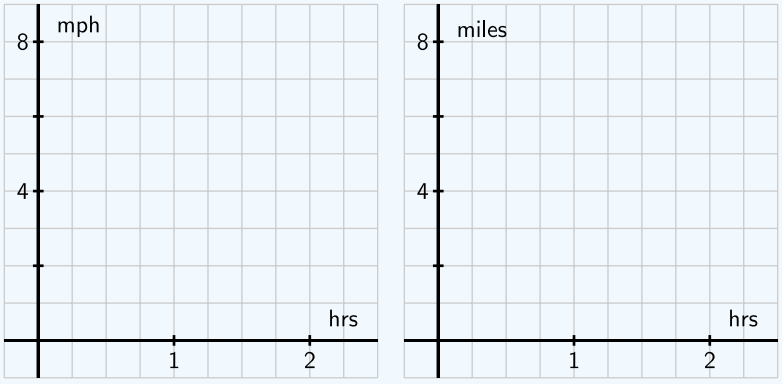
From this study of position and velocity we have learned a great deal. We can use the derivative to find a function's instantaneous rate of change at any point in the domain, to find where the function is increasing or decreasing, where it is concave up or concave down, and to locate relative extremes. The vast majority of the problems and applications we have considered have involved the situation where a particular function is known and we seek information that relies on knowing the function's instantaneous rate of change. For all these tasks, we proceed from a function f to its derivative, f', and use the meaning of the derivative to help us answer important questions.

We have also encountered the reverse situation, where we know the derivative of a function, f', and try to deduce information about f. We will focus our attention in Chapter 4 on this problem: if we know the instantaneous rate of change of a function, can we find the function itself? We start with a more specific question: if we know the instantaneous velocity of an object moving along a straight line path, can we find its corresponding position function?

Preview Activity 4.1.1

Suppose that a person is taking a walk along a long straight path and walks at a constant rate of 3 miles per hour.

1. On the left-hand axes provided in Figure 4.1.1, sketch a labeled graph of the velocity function $v(t) = 3$.



1.

Figure 4.1.1. At left, axes for plotting $y = v(t)$; at right, for plotting $y = s(t)$.

Note that while the scale on the two sets of axes is the same, the units on the right-hand axes differ from those on the left. The right-hand axes will be used in question (d).

2. How far did the person travel during the two hours? How is this distance related to the area of a certain region under the graph of $y = v(t)$?
3. Find an algebraic formula, $s(t)$, for the position of the person at time t, assuming that $s(0) = 0$. Explain your thinking.
4. On the right-hand axes provided in Figure 4.1.11, sketch a labeled graph of the position function $y = s(t)$.
5. For what values of t is the position function s increasing? Explain why this is the case using relevant information about the velocity function v.

Area under the graph of the velocity function

In Preview Activity 4.1.11, we learned that when the velocity of a moving object's velocity is constant (and positive), the area under the velocity curve over an interval of time tells us the distance the object traveled.

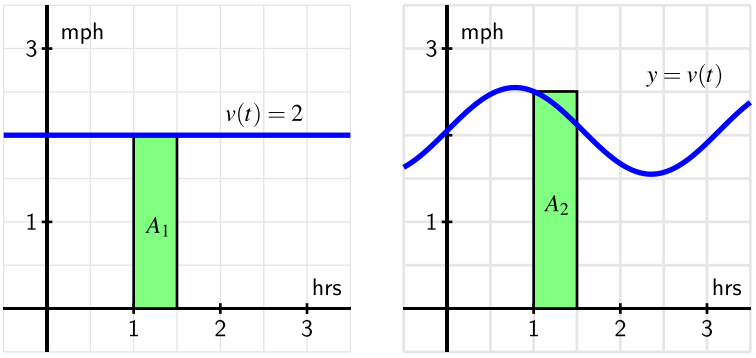


Figure 4.1.2. At left, a constant velocity function; at right, a non-constant velocity function.

The left-hand graph of Figure 4.1.2 shows the velocity of an object moving at 2 miles per hour over the time interval $[1, 1.5]$. The area A_1 of the shaded region under $y = v(t)$ on $[1, 1.5]$ is

$$A_1 = 2\,\frac{\text{miles}}{\text{hour}} \cdot \frac{1}{2}\,\text{hours} = 1\,\text{mile}.$$

This result is simply the fact that distance equals rate times time, provided the rate is constant. Thus, if $v(t)$ is constant on the interval $[a, b]$, the distance traveled on $[a, b]$ is equal to the area A given by

$$A = v(a)(b - a) = v(a)\Delta t,$$

where Δt is the change in t over the interval. (Since the velocity is constant, we can use any value of $v(t)$ on the interval $[a, b]$, we simply chose $v(a)$, the value at the interval's left endpoint.) For several examples where the velocity function is piecewise constant, see http://gvsu.edu/s/9T.[1]

Marc Renault, calculus applets.

The situation is more complicated when the velocity function is not constant. But on relatively small intervals where $v(t)$ does not vary much, we can use the area principle to estimate the distance traveled. The graph at right in Figure 4.1.2 shows a non-constant velocity function. On the interval $[1, 1.5]$, the velocity varies from $v(1) = 2.5$ down to $v(1.5) \approx 2.1$. One estimate for the distance traveled is the area of the pictured rectangle,

$$A_2 = v(1)\Delta t = 2.5\,\frac{\text{miles}}{\text{hour}} \cdot \frac{1}{2}\,\text{hours} = 1.25\,\text{miles}.$$

Note that because v is decreasing on $[1, 1.5]$, $A_2 = 1.25$ is an over-estimate of the actual distance traveled.

To estimate the area under this non-constant velocity function on a wider interval, say $[0, 3]$, one rectangle will not give a good approximation. Instead, we could use the six rectangles pictured in Figure 4.1.3, find the area of each rectangle, and add up the total. Obviously there are choices to make and issues to understand: How many rectangles should we use? Where should we evaluate the function to decide the rectangle's height? What happens if the velocity is sometimes negative? Can we find the exact area under any non-constant curve?

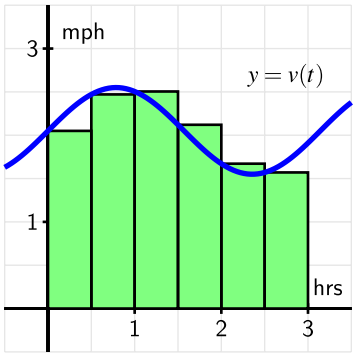


Figure 4.1.3. Using six rectangles to estimate the area under $y = v(t)$ on $[0, 3]$.

We will study these questions and more in what follows; for now it suffices to observe that the simple idea of the area of a rectangle gives us a powerful tool for estimating distance traveled from a velocity function, as well as for estimating the area under an arbitrary curve. To explore the use of multiple rectangles to approximate area under a non-constant velocity function, see the applet found at http://gvsu.edu/s/9U.[2]

Activity 4.1.2

Suppose that a person is walking in such a way that her velocity varies slightly according to the information given in Table 4.1.1 and graph given in Figure 4.1.4.

Table 4.1.1. Velocity data for the person walking.

t	$v(t)$
0.00	1.500
0.25	1.789
0.50	1.938
0.75	1.992
1.00	2.000
1.25	2.008
1.50	2.063
1.75	2.211
2.00	2.500

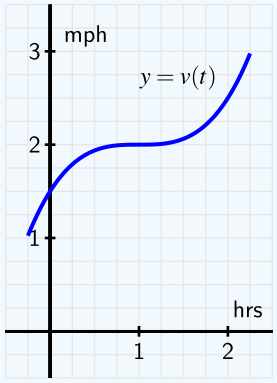


Figure 4.1.4. The graph of $y = v(t)$.

1. Using the grid, graph, and given data appropriately, estimate the distance traveled by the walker during the two hour interval from $t = 0$ to $t = 2$. You should use time intervals of width $\Delta t = 0.5$, choosing a way to use the function consistently to determine the height of each rectangle in order to approximate distance traveled.
2. How could you get a better approximation of the distance traveled on $[0, 2]$? Explain, and then find this new estimate.
3. Now suppose that you know that v is given by $v(t) = 0.5t^3 - 1.5t^2 + 1.5t + 1.5$. Remember that v is the derivative of the walker's position function, s. Find a formula for s so that $s' = v$.
4. Based on your work in (c), what is the value of $s(2) - s(0)$? What is the meaning of this quantity?

Two approaches: area and antidifferentiation

When the velocity of a moving object is positive, the object's position is always increasing. (We will soon consider situations where velocity is negative; for now, we focus on the situation where velocity is always positive.) We have established that whenever v is constant on an interval, the exact distance traveled is the area under the velocity curve. When v is not constant, we can estimate the total distance traveled by finding the areas of rectangles that approximate the area under the velocity curve.

Thus, we see that finding the area between a curve and the horizontal axis is an important exercise: besides being an interesting geometric question, if the curve gives the velocity of a moving object, the area under the curve tells us the exact distance traveled on an interval. We can estimate this area if we have a graph or a table of values for the velocity function.

In Activity 4.1.12, we encountered an alternate approach to finding the distance traveled. If $y = v(t)$ is a formula for the instantaneous velocity of a moving object, then v must be the derivative of the object's position function, s. If we can find a formula for $s(t)$ from the formula for $v(t)$, we will know the position of the object at time t, and the change in position over a particular time interval tells us the distance traveled on that interval.

For a simple example, consider the situation from Preview Activity 4.1.1, where a person is walking along a straight line with velocity function $v(t) = 3$ mph.

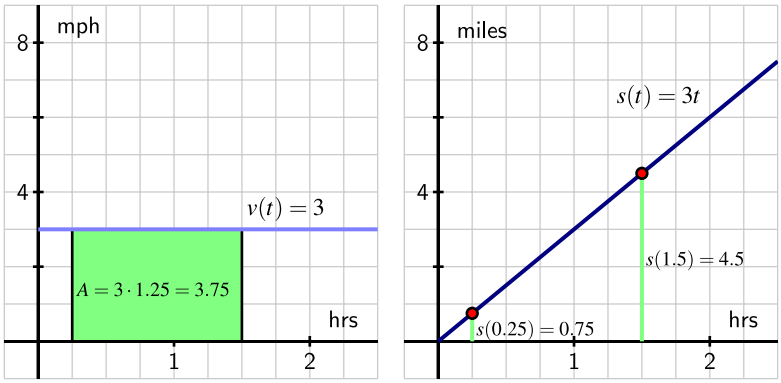


Figure 4.1.5. The velocity function $v(t) = 3$ and corresponding position function $s(t) = 3t$.

On the left-hand graph of the velocity function in Figure 4.1.5, we see the relationship between area and distance traveled,

$$A = 3\,\frac{\text{miles}}{\text{hour}} \cdot 1.25\,\text{hours} = 3.75\,\text{miles}.$$

In addition, we observe[3] that if $s(t) = 3t$, then $s'(t) = 3$, so $s(t) = 3t$ is the position function whose derivative is the given velocity function, $v(t) = 3$. The respective locations of the person at times $t = 0.25$ and $t = 1.5$ are $s(1.5) = 4.5$ and $s(0.25) = 0.75$, and therefore

$$s(1.5) - s(0.25) = 4.5 - 0.75 = 3.75 \text{ miles}.$$

Here we are making the implicit assumption that $s(0) = 0$; we will discuss different possibilities for values of $s(0)$ in subsequent study.

This is the person's change in position on $[0.25, 1.5]$, which is precisely the distance traveled. In this example there are profound ideas and connections that we will study throughout Chapter 4.

For now, observe that if we know a formula for a velocity function v, it can be very helpful to find a function s that satisfies $s' = v$. We say that s is an *antiderivative* of v. More generally, we have the following formal definition.

Definition 4.1.1

If g and G are functions such that $G' = g$, we say that G is an *antiderivative* of g.

For example, if $g(x) = 3x^2 + 2x$, $G(x) = x^3 + x^2$ is an antiderivative of g, because $G'(x) = g(x)$. Note that we say "an" antiderivative of g rather than "the" antiderivative of g, because $H(x) = x^3 + x^2 + 5$ is also a function whose derivative is g, and thus H is another antiderivative of g.

Activity 4.1.3

A ball is tossed vertically in such a way that its velocity function is given by $v(t) = 32 - 32t$, where t is measured in seconds and v in feet per second. Assume that this function is valid for $0 \le t \le 2$.

a. For what values of t is the velocity of the ball positive? What does this tell you about the motion of the ball on this interval of time values?

b. Find an antiderivative, s, of v that satisfies $s(0) = 0$.

c. Compute the value of $s(1) - s(\frac{1}{2})$. What is the meaning of the value you find?

d. Using the graph of $y = v(t)$ provided in Figure 4.1.6, find the exact area of the region under the velocity curve between $t = \frac{1}{2}$ and $t = 1$. What is the meaning of the value you find?

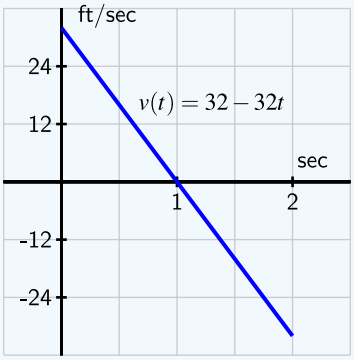


Figure 4.1.6. The graph of $y = v(t)$.

e. Answer the same questions as in (c) and (d) but instead using the interval $[0, 1]$.

f. What is the value of $s(2) - s(0)$? What does this result tell you about the flight of the ball? How is this value connected to the provided graph of $y = v(t)$? Explain.

When velocity is negative

The assumption that its velocity is positive on a given interval guarantees that the movement of an object is always in a single direction, and hence ensures that its change in position is the same as the distance it travels. As we saw in Activity 4.1.3, there are natural settings in which an object's velocity is negative, and we would like to understand this scenario as well.

Consider a simple example where a woman goes for a walk on the beach along a stretch of very straight shoreline that runs east-west. We assume that her initial position is $s(0) = 0$, and that her position function increases as she moves east from her starting location. For instance, $s = 1$ mile represents one mile east of the start location, while $s = -1$ tells us she is one mile west of where she began walking on the beach.

Now suppose she walks in the following manner. From the outset at $t = 0$, she walks due east at a constant rate of 3 mph for 1.5 hours. After 1.5 hours, she stops abruptly and begins walking due west at a constant rate of 4 mph and does so for 0.5 hours. Then, after another abrupt stop and start, she resumes walking at a constant rate of 3 mph to the east for one more hour. What is the total distance she traveled on the time interval from $t = 0$ to $t = 3$? What the total change in her position over that time?

These questions are possible to answer without calculus because the velocity is constant on each interval. From $t = 0$ to $t = 1.5$, she traveled

$$D_{[0,1.5]} = 3 \text{ miles per hour} \cdot 1.5 \text{ hours} = 4.5 \text{ miles}.$$

On $t = 1.5$ to $t = 2$, the distance traveled is

$$D_{[1.5,2]} = 4 \text{ miles per hour} \cdot 0.5 \text{ hours} = 2 \text{ miles}.$$

Finally, in the last hour she walked

$$D_{[2,3]} = 3 \text{ miles per hour} \cdot 1 \text{ hours} = 3 \text{ miles},$$

so the total distance she traveled is

$$D = D_{[0,1.5]} + D_{[1.5,2]} + D_{[2,3]} = 4.5 + 2 + 3 = 9.5 \text{ miles}.$$

Since the velocity for $1.5 < t < 2$ is $v = -4$, indicating motion in the westward direction, the woman first walked 4.5 miles east,

then 2 miles west, followed by 3 more miles east. Thus, the total change in her position is

$$\text{change in position} = 4.5 - 2 + 3 = 5.5 \text{ miles}.$$

We have been able to answer these questions fairly easily, and if we think about the problem graphically, we can generalize our solution to the more complicated setting when velocity is not constant, and possibly negative.

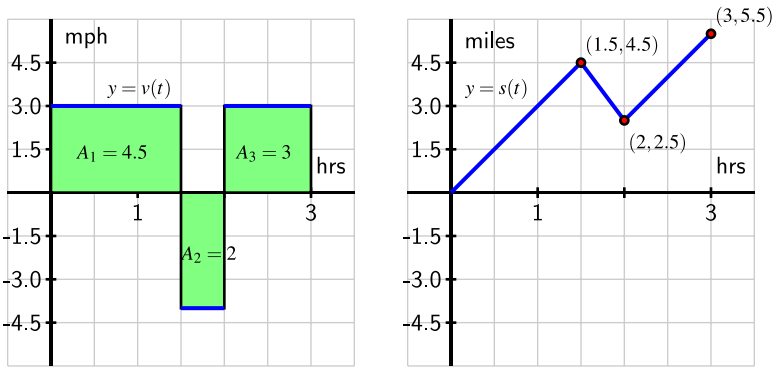


Figure 4.1.7. At left, the velocity function of the person walking; at right, the corresponding position function.

In Figure 4.1.7, we see how the distances we computed can be viewed as areas: $A_1 = 4.5$ comes from multiplyimg rate times time $(3 \cdot 1.5)$, as do A_2 and A_3. But while A_2 is an area (and is therefore positive), because the velocity function is negative for $1.5 < t < 2$, this area has a negative sign associated with it. The negative area distinguishes between distance traveled and change in position.

The distance traveled is the sum of the areas,

$$D = A_1 + A_2 + A_3 = 4.5 + 2 + 3 = 9.5 \text{ miles}.$$

But the change in position has to account for travel in the negative direction. An area above the t-axis is considered positive because it represents distance traveled in the positive direction, while one below the t-axis is viewed as negative because it represents travel in thenegative direction. Thus, the change in the woman's position is

$$s(3) - s(0) = (+4.5) + (-2) + (+3) = 5.5 \text{ miles}.$$

In other words, the woman walks 4.5 miles in the positive direction, followed by two miles in the negative direction, and then 3 more miles in the positive direction.

Negative velocity is also seen in the graph of the position function $y = s(t)$. Its slope is negative (specifically, -4) on the interval $1.5 < t < 2$ because the velocity is -4 on that interval. The negative slope shows the position function is decreasing because the woman is walking east, rather than west.

To summarize, we see that if velocity is sometimes negative, a moving object's change in position different from its distance traveled. If we compute separately the distance traveled on each interval where velocity is positive or negative, we can calculate either the total distance traveled or the total change in position. We assign a negative value to distances traveled in the negative direction when we calculate change in position, but a positive value when we calculate the total distance traveled.

Activity 4.1.4

Suppose that an object moving along a straight line path has its velocity v (in meters per second) at time t (in seconds) given by the piecewise linear function whose graph is pictured at left in Figure 4.1.8. We view movement to the right as being in the positive direction (with positive velocity), while movement to the left is in the negative direction.

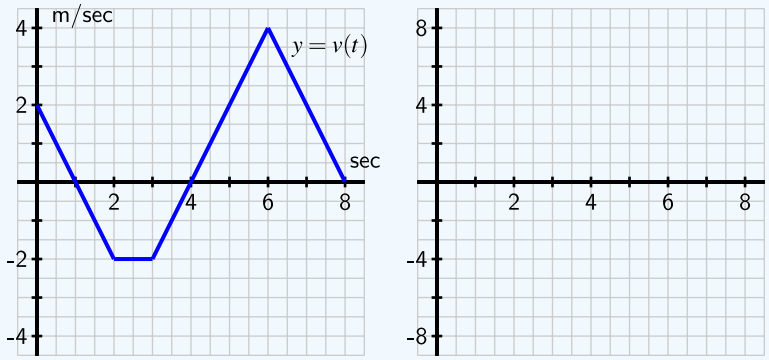


Figure 4.1.8. The velocity function of a moving object.

Suppose further that the object's initial position at time $t = 0$ is $s(0) = 1$.

1. Determine the total distance traveled and the total change in position on the time interval $0 \le t \le 2$. What is the object's position at $t = 2$?
2. On what time intervals is the moving object's position function increasing? Why? On what intervals is the object's position decreasing? Why?
3. What is the object's position at $t = 8$? How many total meters has it traveled to get to this point (including distance in both directions)? Is this different from the object's total change in position on $t = 0$ to $t = 8$?
4. Find the exact position of the object at $t = 1, 2, 3, \ldots, 8$ and use this data to sketch an accurate graph of $y = s(t)$ on the axes provided at right in Figure 4.1.11. How can you use the provided information about $y = v(t)$ to determine the concavity of s on each relevant interval?

Summary

- If we know the velocity of a moving body at every point in a given interval and the velocity is positive throughout, we can estimate the object's distance traveled and in some circumstances determine this value exactly.
- In particular, when velocity is positive on an interval, we can find the total distance traveled by finding the area under the velocity curve and above the t-axis on the given time interval. We may only be able to estimate this area, depending on the shape of the velocity curve.
- An antiderivative of a function f is a new function F whose derivative is f. That is, F is an antiderivative of f provided that $F' = f$. In the context of velocity and position, if we know a velocity function v, an antiderivative of v is a position function s that satisfies $s' = v$. If v is positive on a given interval, say $[a, b]$, then the change in position, $s(b) - s(a)$, measures the distance the moving object traveled on $[a, b]$.
- If its velocity is sometimes negative, a moving object is sometimes traveling in the opposite direction or backtracking. To determine distance traveled, we have to compute the distance separately on intervals where velocity is positive or negative, and account for the change in position on each such interval.

This page titled 4.1: Determining Distance Traveled from Velocity is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 4.1: Determining Distance Traveled from Velocity by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

4.2: Riemann Sums

Motivating Questions

- How can we use a Riemann sum to estimate the area between a given curve and the horizontal axis over a particular interval?
- What are the differences among left, right, middle, and random Riemann sums?
- How can we write Riemann sums in an abbreviated form?

In Section 4.1, we learned that if an object moves with positive velocity v, the area between $y = v(t)$ and the t-axis over a given time interval tells us the distance traveled by the object over that time period. If $v(t)$ is sometimes negative and we view the area of any region below the t-axis as having an associated negative sign, then the sum of these signed areas tells us the moving object's change in position over a given time interval.

For instance, for the velocity function given in Figure 4.2.1, if the areas of shaded regions are A_1, A_2, and A_3 as labeled, then the total distance D traveled by the moving object on $[a, b]$ is

$$D = A_1 + A_2 + A_3,$$

while the total change in the object's position on $[a, b]$ is

$$s(b) - s(a) = A_1 - A_2 + A_3.$$

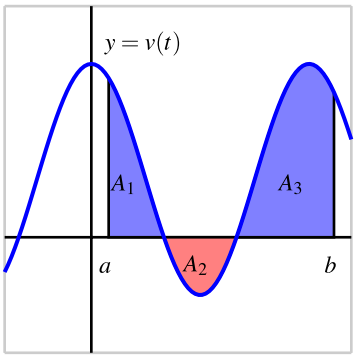


Figure 4.2.1. A velocity function that is sometimes negative.

Because the motion is in the negative direction on the interval where $v(t) < 0$, we subtract A_2 to determine the object's total change in position.

Of course, finding D and $s(b) - s(a)$ for the graph in Figure 4.2.1 presumes that we can actually find the areas A_1, A_2, and A_3. So far, we have worked with velocity functions that were either constant or linear, so that the area bounded by the velocity function and the horizontal axis is a combination of rectangles and triangles, and we can find the area exactly. But when the curve bounds a region that is not a familiar geometric shape, we cannot find its area exactly. Indeed, this is one of our biggest goals in Chapter 4: to learn how to find the exact area bounded between a curve and the horizontal axis for as many different types of functions as possible.

In Activity 4.1.2, we approximated the area under a nonlinear velocity function using rectangles. In the following preview activity, we consider three different options for the heights of the rectangles we will use.

Preview Activity 4.2.1

A person walking along a straight path has her velocity in miles per hour at time t given by the function $v(t) = 0.25t^3 - 1.5t^2 + 3t + 0.25$, for times in the interval $0 \le t \le 2$. The graph of this function is also given in each of the three diagrams in Figure 4.2.2.

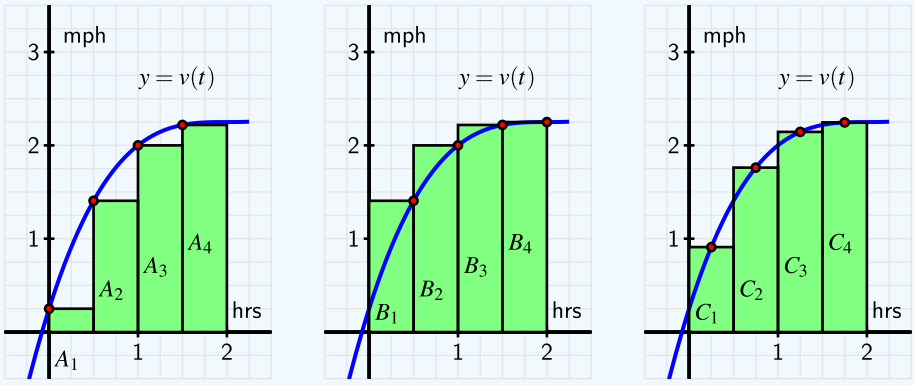


Figure 4.2.2. Three approaches to estimating the area under $y = v(t)$ on the interval $[0, 2]$.

Note that in each diagram, we use four rectangles to estimate the area under $y = v(t)$ on the interval $[0, 2]$, but the method by which the four rectangles' respective heights are decided varies among the three individual graphs.

1. How are the heights of rectangles in the left-most diagram being chosen? Explain, and hence determine the value of
$$S = A_1 + A_2 + A_3 + A_4$$
by evaluating the function $y = v(t)$ at appropriately chosen values and observing the width of each rectangle. Note, for example, that
$$A_3 = v(1) \cdot \frac{1}{2} = 2 \cdot \frac{1}{2} = 1.$$
2. Explain how the heights of rectangles are being chosen in the middle diagram and find the value of
$$T = B_1 + B_2 + B_3 + B_4.$$
3. Likewise, determine the pattern of how heights of rectangles are chosen in the right-most diagram and determine
$$U = C_1 + C_2 + C_3 + C_4.$$
4. Of the estimates S, T, and U, which do you think is the best approximation of D, the total distance the person traveled on $[0, 2]$? Why?

Sigma Notation

We have used sums of areas of rectangles to approximate the area under a curve. Intuitively, we expect that using a larger number of thinner rectangles will provide a better estimate for the area. Consequently, we anticipate dealing with sums of a large number of terms. To do so, we introduce *sigma notation*, named for the Greek letter Σ, which is the capital letter S in the Greek alphabet.

For example, say we are interested in the sum
$$1 + 2 + 3 + \cdots + 100,$$
the sum of the first 100 natural numbers. In sigma notation we write

$$\sum_{k=1}^{100} k = 1 + 2 + 3 + \cdots + 100.$$

We read the symbol $\sum_{k=1}^{100} k$ as "the sum from k equals 1 to 100 of k." The variable k is called the index of summation, and any letter can be used for this variable. The pattern in the terms of the sum is denoted by a function of the index; for example,

$$\sum_{k=1}^{10}(k^2 + 2k) = (1^2 + 2 \cdot 1) + (2^2 + 2 \cdot 2) + (3^2 + 2 \cdot 3) + \cdots + (10^2 + 2 \cdot 10),$$

and more generally,

$$\sum_{k=1}^{n} f(k) = f(1) + f(2) + \cdots + f(n).$$

Sigma notation allows us to vary easily the function being used to describe the terms in the sum, and to adjust the number of terms in the sum simply by changing the value of n. We test our understanding of this new notation in the following activity.

Activity 4.2.2

For each sum written in sigma notation, write the sum long-hand and evaluate the sum to find its value. For each sum written in expanded form, write the sum in sigma notation.

1. $\sum_{k=1}^{5}(k^2 + 2)$
2. $\sum_{i=3}^{6}(2i - 1)$
3. $3 + 7 + 11 + 15 + \cdots + 27$
4. $4 + 8 + 16 + 32 + \cdots + 256$
5. $\sum_{i=1}^{6} \frac{1}{2^i}$

Riemann Sums

When a moving body has a positive velocity function $y = v(t)$ on a given interval $[a, b]$, the area under the curve over the interval gives the total distance the body travels on $[a, b]$. We are also interested in finding the exact area bounded by $y = f(x)$ on an interval $[a, b]$, regardless of the meaning or context of the function f. For now, we continue to focus on finding an accurate estimate of this area by using a sum of the areas of rectangles. Unless otherwise indicated, we assume that f is continuous and non-negative on $[a, b]$.

The first choice we make in such an approximation is the number of rectangles.

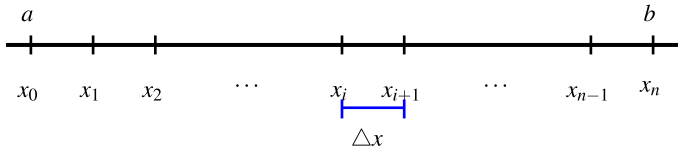


Figure 4.2.3. Subdividing the interval $[a, b]$ into n subintervals of equal length Δx.

If we desire n rectangles of equal width to subdivide the interval $[a, b]$, then each rectangle must have width $\Delta x = \frac{b-a}{n}$. We let $x_0 = a$, $x_n = b$, and define $x_i = a + i\Delta x$, so that $x_1 = x_0 + \Delta x$, $x_2 = x_0 + 2\Delta x$, and so on, as pictured in Figure 4.2.13.

We use each subinterval $[x_i, x_{i+1}]$ as the base of a rectangle, and next choose the height of the rectangle on that subinterval. There are three standard choices: we can use the left endpoint of each subinterval, the right endpoint of each subinterval, or the midpoint of each. These are precisely the options encountered in Preview Activity 4.2.11 and seen in Figure 4.2.12. We next

explore how these choices can be described in sigma notation.

Consider an arbitrary positive function f on $[a,b]$ with the interval subdivided as shown in Figure 4.2.13, and choose to use left endpoints. Then on each interval $[x_i, x_{i+1}]$, the area of the rectangle formed is given by

$$A_{i+1} = f(x_i) \cdot \Delta x,$$

as seen in Figure 4.2.4.

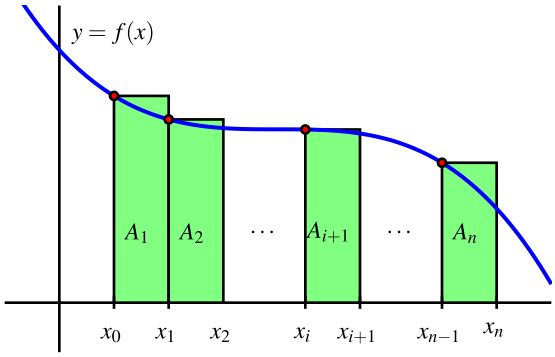


Figure 4.2.4. Subdividing the interval $[a,b]$ into n subintervals of equal length Δx and approximating the area under $y = f(x)$ over $[a,b]$ using left rectangles.

If we let L_n denote the sum of the areas of these rectangles, we see that

$$\begin{aligned} L_n &= A_1 + A_2 + \cdots + A_{i+1} + \cdots + A_n \\ &= f(x_0) \cdot \Delta x + f(x_1) \cdot \Delta x + \cdots + f(x_i) \cdot \Delta x + \cdots + f(x_{n-1}) \cdot \Delta x. \end{aligned}$$

In the more compact sigma notation, we have

$$L_n = \sum_{i=0}^{n-1} f(x_i) \Delta x.$$

Note that since the index of summation begins at 0 and ends at $n-1$, there are indeed n terms in this sum. We call L_n the *left Riemann sum* for the function f on the interval $[a,b]$.

To see how the Riemann sums for right endpoints and midpoints are constructed, we consider Figure 4.2.5.

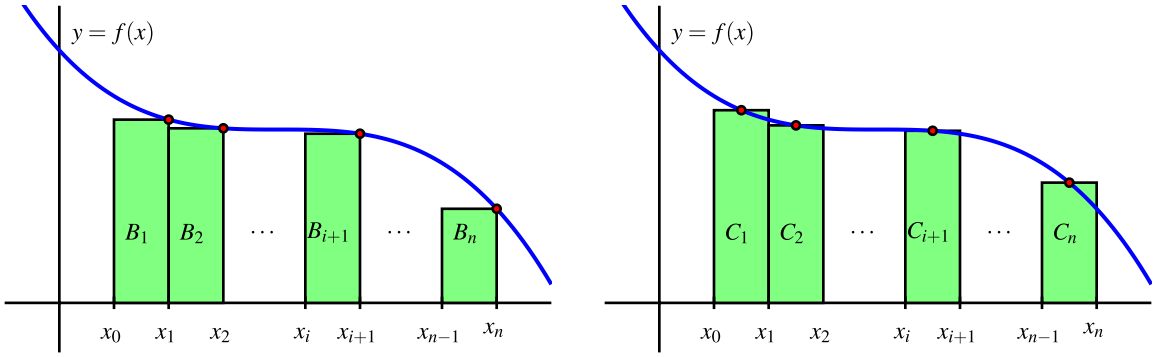


Figure 4.2.5. Riemann sums using right endpoints and midpoints.

For the sum with right endpoints, we see that the area of the rectangle on an arbitrary interval $[x_i, x_{i+1}]$ is given by $B_{i+1} = f(x_{i+1}) \cdot \Delta x$, and that the sum of all such areas of rectangles is given by

$$\begin{aligned} R_n &= B_1 + B_2 + \cdots + B_{i+1} + \cdots + B_n \\ &= f(x_1) \cdot \Delta x + f(x_2) \cdot \Delta x + \cdots + f(x_{i+1}) \cdot \Delta x + \cdots + f(x_n) \cdot \Delta x \\ &= \sum_{i=1}^{n} f(x_i) \Delta x. \end{aligned}$$

We call R_n the *right Riemann sum* for the function f on the interval $[a, b]$.

For the sum that uses midpoints, we introduce the notation

$$\overline{x}_{i+1} = \frac{x_i + x_{i+1}}{2}$$

so that $\overline{x}_{i+1}$ is the midpoint of the interval $[x_i, x_{i+1}]$. For instance, for the rectangle with area C_1 in Figure 4.2.15, we now have

$$C_1 = f(\overline{x}_1) \cdot \Delta x.$$

Hence, the sum of all the areas of rectangles that use midpoints is

$$\begin{aligned} M_n &= C_1 + C_2 + \cdots + C_{i+1} + \cdots + C_n \\ &= f(\overline{x_1}) \cdot \Delta x + f(\overline{x_2}) \cdot \Delta x + \cdots + f(\overline{x}_{i+1}) \cdot \Delta x + \cdots + f(\overline{x}_n) \cdot \Delta x \\ &= \sum_{i=1}^{n} f(\overline{x}_i) \Delta x, \end{aligned}$$

and we say that M_n is the *middle Riemann sum* for f on $[a, b]$.

Thus, we have two variables to explore: the number of rectangles and the height of each rectangle. We can explore these choices dynamically, and the applet [1] found at http://gvsu.edu/s/a9 is a particularly useful one. There we see the image shown in Figure 4.2.6, but with the opportunity to adjust the slider bars for the heights and the number of rectangles.

Marc Renault, Geogebra Calculus Applets.

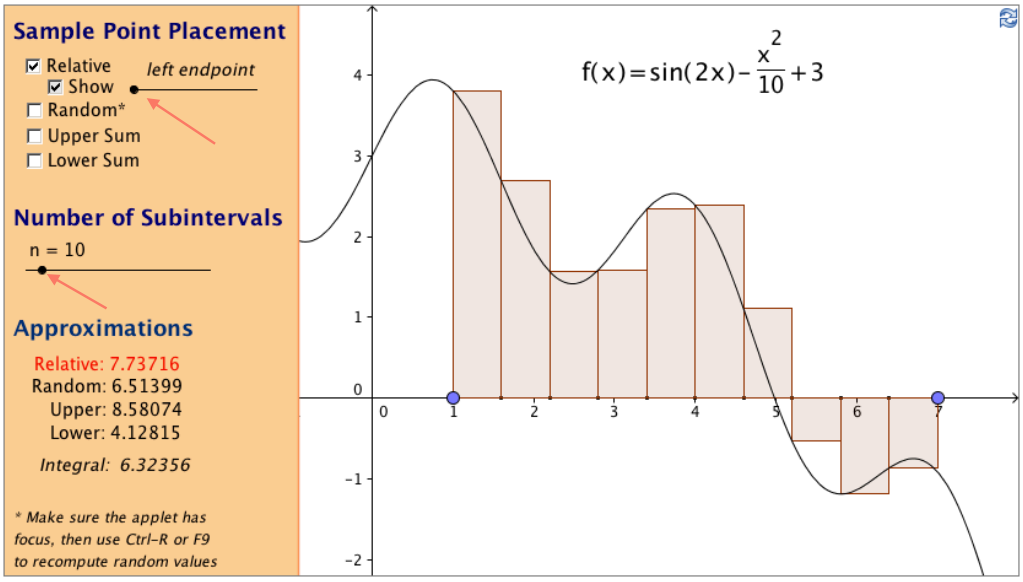


Figure 4.2.6. A snapshot of the applet found at http://gvsu.edu/s/a9.

By moving the sliders, we can see how the heights of the rectangles change as we consider left endpoints, midpoints, and right endpoints, as well as the impact that a larger number of narrower rectangles has on the approximation of the exact area bounded by the function and the horizontal axis.

When $f(x) \geq 0$ on $[a, b]$, each of the Riemann sums L_n, R_n, and M_n provides an estimate of the area under the curve $y = f(x)$ over the interval $[a, b]$. We also recall that in the context of a nonnegative velocity function $y = v(t)$, the corresponding Riemann sums approximate the distance traveled on $[a, b]$ by a moving object with velocity function v.

There is a more general way to think of Riemann sums, and that is to allow any choice of where the function is evaluated to determine the rectangle heights. Rather than saying we'll always choose left endpoints, or always choose midpoints, we simply say that a point x_{i+1}^* will be selected at random in the interval $[x_i, x_{i+1}]$ (so that $x_i \leq x_{i+1}^* \leq x_{i+1}$). The Riemann sum is then given by

$$f(x_1^*) \cdot \Delta x + f(x_2^*) \cdot \Delta x + \cdots + f(x_{i+1}^*) \cdot \Delta x + \cdots + f(x_n^*) \cdot \Delta x = \sum_{i=1}^{n} f(x_i^*) \Delta x.$$

At http://gvsu.edu/s/a9, the applet noted earlier and referenced in Figure 4.2.6, by unchecking the "relative" box at the top left, and instead checking "random," we can easily explore the effect of using random point locations in subintervals on a Riemann sum. In computational practice, we most often use L_n, R_n, or M_n, while the random Riemann sum is useful in theoretical discussions. In the following activity, we investigate several different Riemann sums for a particular velocity function.

Activity 4.2.3

Suppose that an object moving along a straight line path has its velocity in feet per second at time t in seconds given by $v(t) = \frac{2}{9}(t-3)^2 + 2$.

1. Carefully sketch the region whose exact area will tell you the value of the distance the object traveled on the time interval $2 \leq t \leq 5$.
2. Estimate the distance traveled on $[2, 5]$ by computing L_4, R_4, and M_4.
3. Does averaging L_4 and R_4 result in the same value as M_4? If not, what do you think the average of L_4 and R_4 measures?
4. For this question, think about an arbitrary function f, rather than the particular function v given above. If f is positive and increasing on $[a, b]$, will L_n over-estimate or under-estimate the exact area under f on $[a, b]$? Will R_n over- or under-estimate the exact area under f on $[a, b]$? Explain.

When the function is sometimes negative

For a Riemann sum such as

$$L_n = \sum_{i=0}^{n-1} f(x_i) \Delta x,$$

we can of course compute the sum even when f takes on negative values. We know that when f is positive on $[a, b]$, a Riemann sum estimates the area bounded between f and the horizontal axis over the interval.

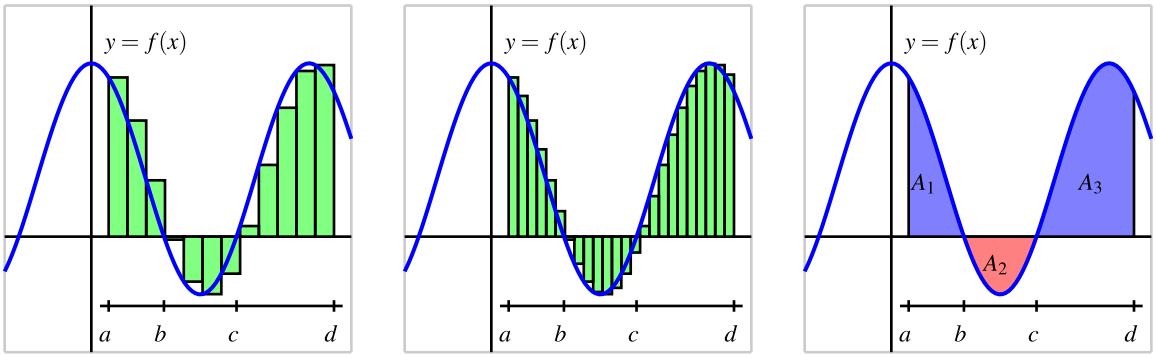


Figure 4.2.7. At left and center, two left Riemann sums for a function f that is sometimes negative; at right, the areas bounded by f on the interval $[a, d]$.

For the function pictured in the first graph of Figure 4.2.7, a left Riemann sum with 12 subintervals over $[a, d]$ is shown. The function is negative on the interval $b \leq x \leq c$, so at the four left endpoints that fall in $[b, c]$, the terms $f(x_i)\Delta x$ are negative. This means that those four terms in the Riemann sum produce an estimate of the *opposite* of the area bounded by $y = f(x)$ and the x-axis on $[b, c]$.

In the middle graph of Figure 4.2.7, we see that by increasing the number of rectangles the approximation of the area (or the opposite of the area) bounded by the curve appears to improve.

In general, any Riemann sum of a continuous function f on an interval $[a, b]$ approximates the difference between the area that lies above the horizontal axis on $[a, b]$ and under f and the area that lies below the horizontal axis on $[a, b]$ and above f. In the notation of Figure 4.2.7, we may say that

$$L_{24} \approx A_1 - A_2 + A_3,$$

where L_{24} is the left Riemann sum using 24 subintervals shown in the middle graph. A_1 and A_3 are the areas of the regions where f is positive, and A_2 is the area where f is negative. We will call the quantity $A_1 - A_2 + A_3$ the *net signed area* bounded by f over the interval $[a, d]$, where by the phrase "signed area" we indicate that we are attaching a minus sign to the areas of regions that fall below the horizontal axis.

Finally, we recall that if the function f represents the velocity of a moving object, the sum of the areas bounded by the curve tells us the total distance traveled over the relevant time interval, while the net signed area bounded by the curve computes the object's change in position on the interval.

Activity 4.2.4

Suppose that an object moving along a straight line path has its velocity v (in feet per second) at time t (in seconds) given by

$$v(t) = \frac{1}{2}t^2 - 3t + \frac{7}{2}.$$

1. Compute M_5, the middle Riemann sum, for v on the time interval $[1, 5]$. Be sure to clearly identify the value of Δt as well as the locations of $t_0, t_1, \cdots, t_5$. In addition, provide a careful sketch of the function and the corresponding rectangles that are being used in the sum.
2. Building on your work in (a), estimate the total change in position of the object on the interval $[1, 5]$.
3. Building on your work in (a) and (b), estimate the total distance traveled by the object on $[1, 5]$.
4. Use appropriate computing technology[2] to compute M_{10} and M_{20}. What exact value do you think the middle sum eventually approaches as n increases without bound? What does that number represent in the physical context of the overall problem?

For instance, consider the applet at http://gvsu.edu/s/a9 and change the function and adjust the locations of the blue points that represent the interval endpoints a and b.

Summary

- A Riemann sum is simply a sum of products of the form $f(x_i^*)\Delta x$ that estimates the area between a positive function and the horizontal axis over a given interval. If the function is sometimes negative on the interval, the Riemann sum estimates the difference between the areas that lie above the horizontal axis and those that lie below the axis.
- The three most common types of Riemann sums are left, right, and middle sums, but we can also work with a more general Riemann sum. The only difference among these sums is the location of the point at which the function is evaluated to determine the height of the rectangle whose area is being computed. For a left Riemann sum, we evaluate the function at the left endpoint of each subinterval, while for right and middle sums, we use right endpoints and midpoints, respectively.
- The left, right, and middle Riemann sums are denoted L_n, R_n, and M_n, with formulas

$$L_n = f(x_0)\Delta x + f(x_1)\Delta x + \cdots + f(x_{n-1})\Delta x = \sum_{i=0}^{n-1} f(x_i)\Delta x,$$

$$R_n = f(x_1)\Delta x + f(x_2)\Delta x + \cdots + f(x_n)\Delta x = \sum_{i=1}^{n} f(x_i)\Delta x,$$

$$M_n = f(\overline{x}_1)\Delta x + f(\overline{x}_2)\Delta x + \cdots + f(\overline{x}_n)\Delta x = \sum_{i=1}^{n} f(\overline{x}_i)\Delta x,$$

where $x_0 = a$, $x_i = a + i\Delta x$, and $x_n = b$, using $\Delta x = \frac{b-a}{n}$. For the midpoint sum, $\overline{x}_i = (x_{i-1} + x_i)/2$.

This page titled 4.2: Riemann Sums is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 4.2: Riemann Sums by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

4.3: The Definite Integral

Motivating Questions

- How does increasing the number of subintervals affect the accuracy of the approximation generated by a Riemann sum?
- What is the definition of the definite integral of a function f over the interval $[a, b]$?
- What does the definite integral measure exactly, and what are some of the key properties of the definite integral?

In Figure 4.3.1, we see evidence that increasing the number of rectangles in a Riemann sum improves the accuracy of the approximation of the net signed area bounded by the given function.

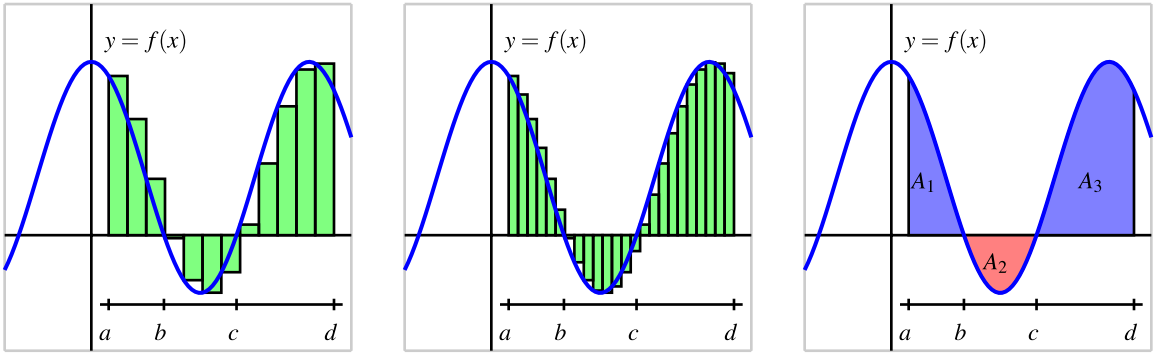


Figure 4.3.1. At left and center, two left Riemann sums for a function f that is sometimes negative; at right, the exact areas bounded by f on the interval $[a, d]$.

We therefore explore the natural idea of allowing the number of rectangles to increase without bound. In an effort to compute the exact net signed area we also consider the differences among left, right, and middle Riemann sums and the different results they generate as the value of n increases. We begin with functions that are exclusively positive on the interval under consideration.

Preview Activity 4.3.1

Consider the applet found at http://gvsu.edu/s/a9 [1]. There, you will initially see the situation shown in Figure 4.3.2.

Marc Renault, Shippensburg University, Geogebra Applets for Calclulus, http://gvsu.edu/s/5p.

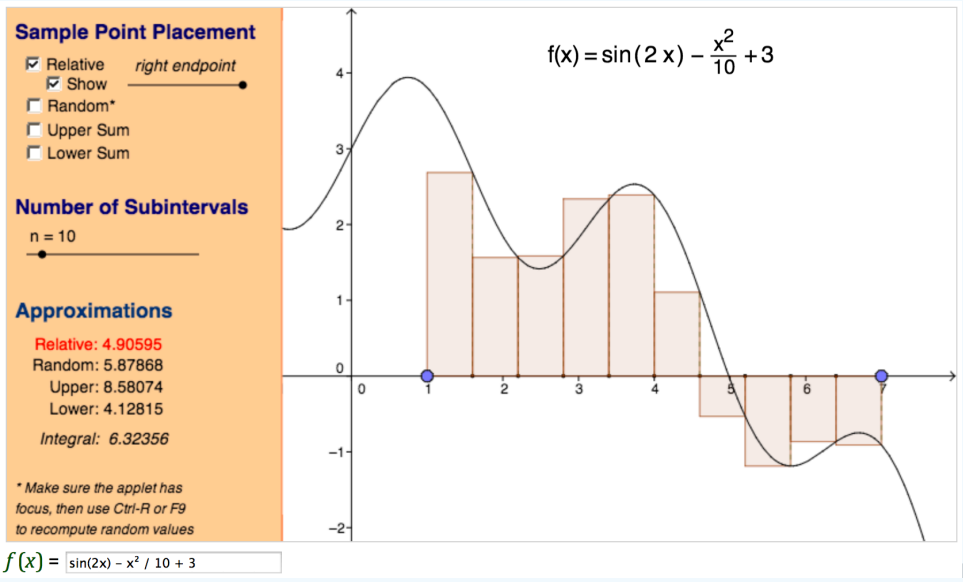


Figure 4.3.2. A right Riemann sum with 10 subintervals for the function $f(x) = \sin(2x) - \frac{x^2}{10} + 3$ on the interval $[1, 7]$. The value of the sum is $R_{10} = 4.90595$.

Note that the value of the chosen Riemann sum is displayed next to the word “relative,” and that you can change the type of Riemann sum being computed by dragging the point on the slider bar below the phrase “sample point placement.”

Explore to see how you can change the window in which the function is viewed, as well as the function itself. You can set the minimum and maximum values of x by clicking and dragging on the blue points that set the endpoints; you can change the function by typing a new formula in the “f(x)” window at the bottom; and you can adjust the overall window by “panning and zooming” by using the Shift key and the scrolling feature of your mouse. More information on how to pan and zoom can be found at http://gvsu.edu/s/Fl.

Work accordingly to adjust the applet so that it uses a left Riemann sum with $n = 5$ subintervals for the function is $f(x) = 2x + 1$. You should see the updated figure shown in Figure 4.3.13. Then, answer the following questions.

1. Update the applet (and view window, as needed) so that the function being considered is $f(x) = 2x + 1$ on $[1, 4]$, as directed above. For this function on this interval, compute L_n, M_n, R_n for $n = 5$, $n = 25$, and $n = 100$. What appears to be the exact area bounded by $f(x) = 2x + 1$ and the x-axis on $[1, 4]$?
2. Use basic geometry to determine the exact area bounded by $f(x) = 2x + 1$ and the x-axis on $[1, 4]$.
3. Based on your work in (a) and (b), what do you observe occurs when we increase the number of subintervals used in the Riemann sum?
4. Update the applet to consider the function $f(x) = x^2 + 1$ on the interval $[1, 4]$ (note that you need to enter “`x ^ 2 + 1`” for the function formula). Use the applet to compute L_n, M_n, R_n for $n = 5$, $n = 25$, and $n = 100$. What do you conjecture is the exact area bounded by $f(x) = x^2 + 1$ and the x-axis on $[1, 4]$?
5. Why can we not compute the exact value of the area bounded by $f(x) = x^2 + 1$ and the x-axis on $[1, 4]$ using a formula like we did in (b)?

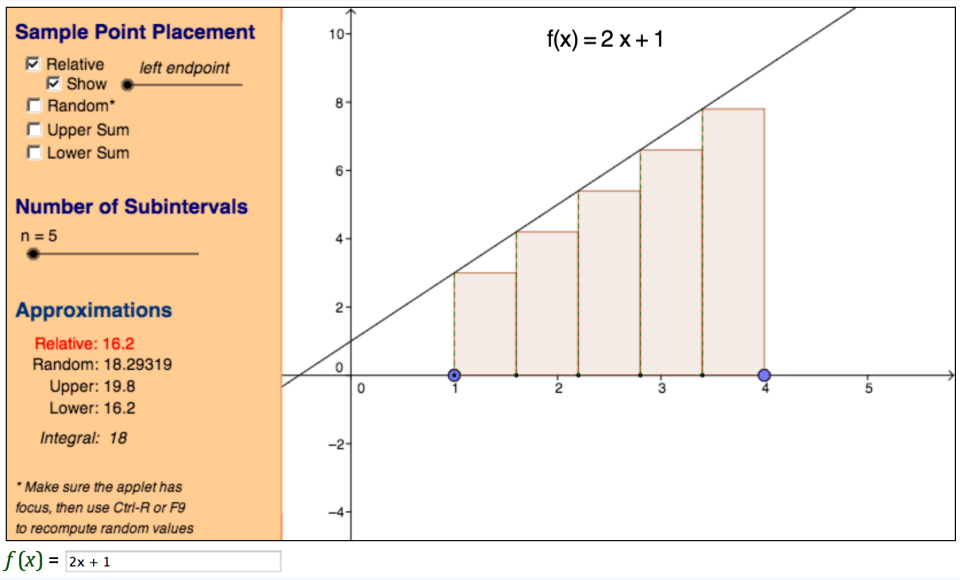


Figure 4.3.3. A left Riemann sum with 5 subintervals for the function $f(x) = 2x + 1$ on the interval $[1, 4]$. The value of the sum is $L_5 = 16.2$.

The definition of the definite integral

In Preview Activity 4.3.1, we saw that as the number of rectangles got larger and larger, the values of L_n, M_n, and R_n all grew closer and closer to the same value. It turns out that this occurs for any continuous function on an interval $[a, b]$, and also for a Riemann sum using any point x^*_{i+1} in the interval $[x_i, x_{i+1}]$. Thus, as we let $n \to \infty$, it doesn't really matter where we choose to evaluate the function within a given subinterval, because

$$\lim_{n\to\infty} L_n = \lim_{n\to\infty} R_n = \lim_{n\to\infty} M_n = \lim_{n\to\infty} \sum_{i=1}^{n} f(x^*_i)\Delta x.$$

That these limits always exist (and share the same value) when f is continuous[2] allows us to make the following definition.

It turns out that a function need not be continuous in order to have a definite integral. For our purposes, we assume that the functions we consider are continuous on the interval(s) of interest. It is straightforward to see that any function that is piecewise continuous on an interval of interest will also have a well-defined definite integral.

Definition 4.3.1

The ***definite integral*** of a continuous function f on the interval $[a, b]$, denoted $\int_a^b f(x)\ dx$, is the real number given by

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x^*_i)\Delta x,$$

where $\Delta x = \frac{b-a}{n}$, $x_i = a + i\Delta x$ (for $i = 0, \ldots, n$), and x^*_i satisfies $x_{i-1} \le x^*_i \le x_i$ (for $i = 1, \ldots, n$).

We call the symbol $\int$ the *integral sign*, the values a and b the *limits of integration*, and the function f the *integrand*. The process of determining the real number $\int_a^b f(x)\ dx$ is called *evaluating the definite integral*. While there are several different interpretations of the definite integral, for now the most important is that $\int_a^b f(x)\ dx$ measures the net signed area bounded by $y = f(x)$ and the x-axis on the interval $[a, b]$.

For example, if f is the function pictured in Figure 4.3.4, and A_1, A_2, and A_3 are the exact areas bounded by f and the x-axis on the respective intervals $[a, b]$, $[b, c]$, and $[c, d]$, then

$$\int_a^b f(x)\,dx = A_1, \quad \int_b^c f(x)\,dx = -A_2,$$

$$\int_c^d f(x)\,dx = A_3,$$

$$\text{and } \int_a^d f(x)\,dx = A_1 - A_2 + A_3.$$

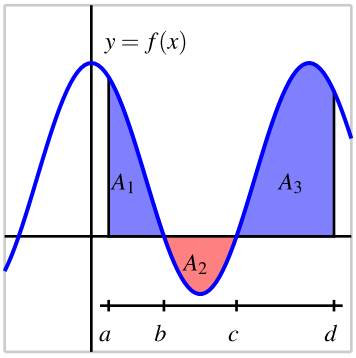


Figure 4.3.4. A continuous function f on the interval $[a, d]$.

We can also use definite integrals to express the change in position and the distance traveled by a moving object. If v is a velocity function on an interval $[a, b]$, then the change in position of the object, $s(b) - s(a)$, is given by

$$s(b) - s(a) = \int_a^b v(t)\,dt.$$

If the velocity function is nonnegative on $[a, b]$, then $\int_a^b v(t)\ dt$ tells us the distance the object traveled. If the velocity is sometimes negative on $[a, b]$, we can use definite integrals to find the areas bounded by the function on each interval where v does not change sign, and the sum of these areas will tell us the distance the object traveled.

To compute the value of a definite integral from the definition, we have to take the limit of a sum. While this is possible to do in select circumstances, it is also tedious and time-consuming, and does not offer much additional insight into the meaning or interpretation of the definite integral. Instead, in Section 4.4, we will learn the Fundamental Theorem of Calculus, which provides a shortcut for evaluating a large class of definite integrals. This will enable us to determine the exact net signed area bounded by a continuous function and the x-axis in many circumstances.

For now, our goal is to understand the meaning and properties of the definite integral, rather than to compute its value. To do this, we will rely on the net signed area interpretation of the definite integral. So we will use as examples curves that produce regions whose areas we can compute exactly through area formulas. We can thus compute the exact value of the corresponding integral.

For instance, if we wish to evaluate the definite integral $\int_1^4 (2x + 1)\ dx$, we observe that the region bounded by this function and the x-axis is the trapezoid shown in Figure 4.3.5. By the formula for the area of a trapezoid, $A = \frac{1}{2}(3 + 9) \cdot 3 = 18$, so

$$\int_1^4 (2x + 1)\,dx = 18.$$

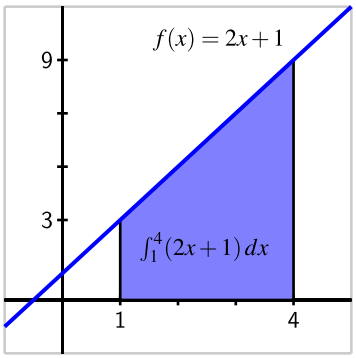
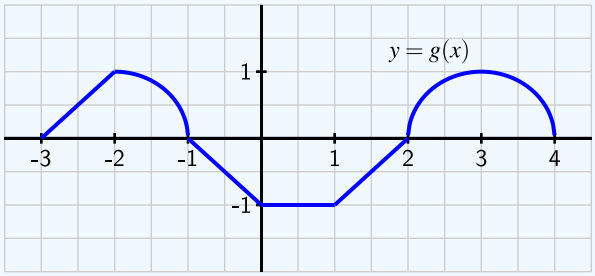


Figure 4.3.5. The area bounded by $f(x) = 2x + 1$ and the x-axis on the interval $[1, 4]$.

Activity 4.3.2

Use known geometric formulas and the net signed area interpretation of the definite integral to evaluate each of the definite integrals below.

1. $\int_0^1 3x \ dx$
2. $\int_{-1}^4 (2 - 2x) \ dx$
3. $\int_{-1}^1 \sqrt{1 - x^2} \ dx$
4. $\int_{-3}^4 g(x) \ dx$, where g is the function pictured in Figure 4.3.17. Assume that each portion of g is either part of a line or part of a circle.



1. Figure 4.3.5. The function g for part (d). Note that g is piecewise defined, and each piece of the function is part of a circle or part of a line.

Some properties of the definite integral

Regarding the definite integral of a function f over an interval $[a, b]$ as the net signed area bounded by f and the x-axis, we discover several standard properties of the definite integral. It is helpful to remember that the definite integral is defined in terms of Riemann sums, which consist of the areas of rectangles.

For any real number a and the definite integral $\int_a^a f(x) \ dx$ it is evident that no area is enclosed, because the interval begins and ends with the same point. Hence,

Note

If f is a continuous function and a is a real number, then $\int_a^a f(x)\ dx = 0$.

Next, we consider the result of subdividing the interval of integration. In Figure 4.3.7, we see that

$$\int_a^b f(x)\,dx = A_1, \quad \int_b^c f(x)\,dx = A_2,$$
$$\text{and} \quad \int_a^c f(x)\,dx = A_1 + A_2,$$

which illustrates the following general rule.

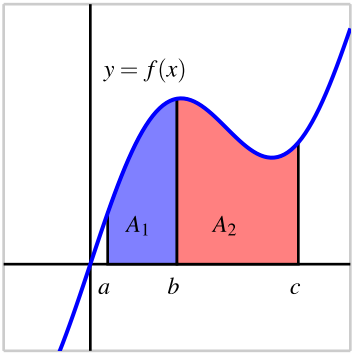


Figure 4.3.7. The area bounded by $y = f(x)$ on the interval $[a, c]$.

Note

If f is a continuous function and a, b, and c are real numbers, then

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx.$$

While this rule is easy to see if $a < b < c$, it in fact holds in general for any values of a, b, and c. Another property of the definite integral states that if we reverse the order of the limits of integration, we change the sign of the integral's value.

Note

If f is a continuous function and a and b are real numbers, then

$$\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx.$$

This result makes sense because if we integrate from a to b, then in the defining Riemann sum we set $\Delta x = \frac{b-a}{n}$, while if we integrate from b to a, we have $\Delta x = \frac{a-b}{n} = -\frac{b-a}{n}$, and this is the only change in the sum used to define the integral.

There are two additional useful properties of the definite integral. When we worked with derivative rules in Chapter 2, we formulated the Constant Multiple Rule and the Sum Rule. Recall that the Constant Multiple Rule says that if f is a differentiable function and k is a constant, then

$$\frac{d}{dx}[kf(x)] = kf'(x),$$

and the Sum Rule says that if f and g are differentiable functions, then

$$\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x).$$

These rules are useful because they allow to deal individually with the simplest parts of certain functions by taking advantage of addition and multiplying by a constant. In other words, the process of taking the derivative respects addition and multiplying by constants in the simplest possible way.

It turns out that similar rules hold for the definite integral. First, let's consider the functions pictured in Figure 4.3.8.

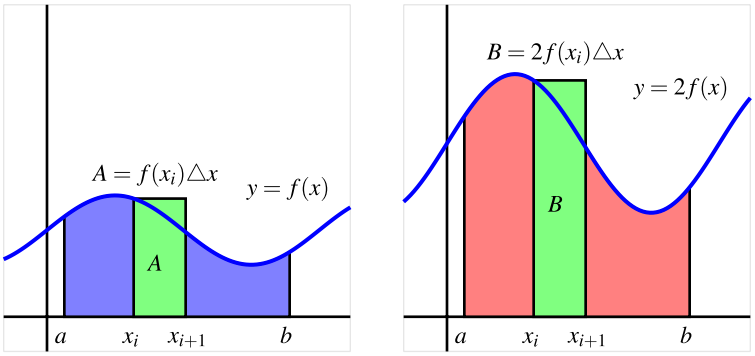


Figure 4.3.8. The areas bounded by $y = f(x)$ and $y = 2f(x)$ on $[a, b]$.

Because multiplying the function by 2 doubles its height at every x-value, we see that the height of each rectangle in a left Riemann sum is doubled, $f(x_i)$ for the original function, versus $2f(x_i)$ in the doubled function. For the areas A and B, it follows $B = 2A$. As this is true regardless of the value of n or the type of sum we use, we see that in the limit, the area of the red region bounded by $y = 2f(x)$ will be twice the area of the blue region bounded by $y = f(x)$. As there is nothing special about the value 2 compared to an arbitrary constant k, the following general principle holds.

Constant Multiple Rule

If f is a continuous function and k is any real number, then

$$\int_a^b k \cdot f(x)\,dx = k\int_a^b f(x)\,dx.$$

We see a similar situation with the sum of two functions f and g.

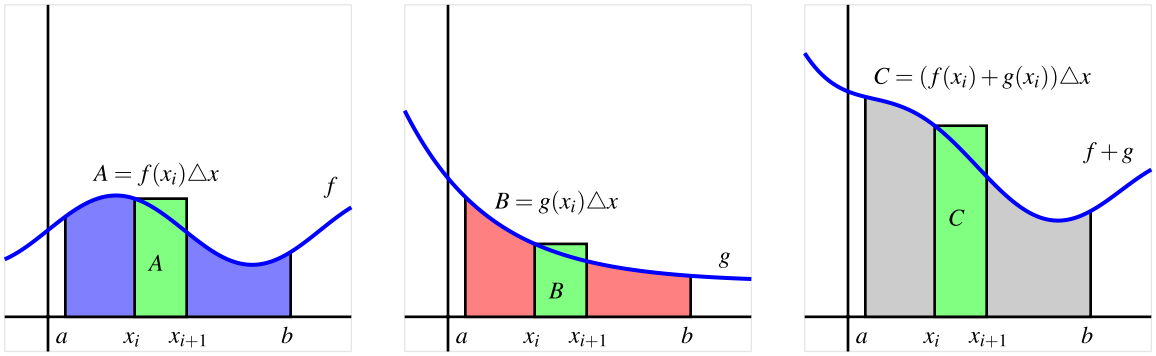


Figure 4.3.9. The areas bounded by $y = f(x)$ and $y = g(x)$ on $[a, b]$, as well as the area bounded by $y = f(x) + g(x)$.

If we take the sum of two functions f and g at every point in the interval, the height of the function $f + g$ is given by $(f + g)(x_i) = f(x_i) + g(x_i)$. Hence, for the pictured rectangles with areas A, B, and C, it follows that $C = A + B$. Because this will occur for every such rectangle, in the limit the area of the gray region will be the sum of the areas of the blue and red regions. In terms of definite integrals, we have the following general rule.

Sum Rule

If f and g are continuous functions, then

$$\int_a^b [f(x) + g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx.$$

The Constant Multiple and Sum Rules can be combined to say that for any continuous functions f and g and any constants c and k,

$$\int_a^b [cf(x) \pm kg(x)]\,dx = c\int_a^b f(x)\,dx \pm k\int_a^b g(x)\,dx.$$

Activity 4.3.3

Suppose that the following information is known about the functions f, g, x^2, and x^3:

- $\int_0^2 f(x)\ dx = -3$; $\int_2^5 f(x)\ dx = 2$
- $\int_0^2 g(x)\ dx = 4$; $\int_2^5 g(x)\ dx = -1$
- $\int_0^2 x^2\ dx = \frac{8}{3}$; $\int_2^5 x^2\ dx = \frac{117}{3}$
- $\int_0^2 x^3\ dx = 4$; $\int_2^5 x^3\ dx = \frac{609}{4}$

Use the provided information and the rules discussed in the preceding section to evaluate each of the following definite integrals.

1. $\int_5^2 f(x)\ dx$
2. $\int_0^5 g(x)\ dx$
3. $\int_0^5 (f(x) + g(x))\ dx$
4. $\int_2^5 (3x^2 - 4x^3)\ dx$
5. $\int_5^0 (2x^3 - 7g(x))\ dx$

How the definite integral is connected to a function's average value

One of the most valuable applications of the definite integral is that it provides a way to discuss the average value of a function, even for a function that takes on infinitely many values. Recall that if we wish to take the average of n numbers $y_1, y_2, \ldots, y_n$, we compute

$$AVG = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

Since integrals arise from Riemann sums in which we add n values of a function, it should not be surprising that evaluating an integral is similar to averaging the output values of a function. Consider, for instance, the right Riemann sum R_n of a function f, which is given by

$$R_n = f(x_1)\Delta x + f(x_2)\Delta x + \cdots + f(x_n)\Delta x = (f(x_1) + f(x_2) + \cdots + f(x_n))\Delta x.$$

Since $\Delta x = \frac{b-a}{n}$, we can thus write

$$\begin{aligned} R_n &= (f(x_1) + f(x_2) + \cdots + f(x_n)) \cdot \frac{b-a}{n} \\ &= (b-a)\frac{f(x_1) + f(x_2) + \cdots + f(x_n)}{n}. \end{aligned} \tag{4.3.11}$$

We see that the right Riemann sum with n subintervals is just the length of the interval $(b-a)$ times the average of the n function values found at the right endpoints. And just as with our efforts to compute area, the larger the value of n we use, the more accurate our average will be. Indeed, we will define the average value of f on $[a,b]$ to be

$$f_{\text{AVG}[a,b]} = \lim_{n\to\infty} \frac{f(x_1) + f(x_2) + \cdots + f(x_n)}{n}.$$

But we also know that for any continuous function f on $[a,b]$, taking the limit of a Riemann sum leads precisely to the definite integral. That is, $\lim_{n\to\infty} R_n = \int_a^b f(x)\ dx$, and thus taking the limit as $n \to \infty$ in Equation (4.3.11), we have that

$$\int_a^b f(x)\,dx = (b-a) \cdot f_{\text{AVG}[a,b]}. \tag{4.3.12}$$

Solving Equation (4.3.12) for $f_{\text{AVG}[a,b]}$, we have the following general principle.

Average value of a function

If f is a continuous function on $[a,b]$, then its **average value** on $[a,b]$ is given by the formula

$$f_{\text{AVG}[a,b]} = \frac{1}{b-a} \cdot \int_a^b f(x)\,dx.$$

Equation (4.3.12) tells us another way to interpret the definite integral: the definite integral of a function f from a to b is the length of the interval $(b-a)$ times the average value of the function on the interval. In addition, when the function f is nonnegative on $[a,b]$, Equation (4.3.12) has a natural visual interpretation.

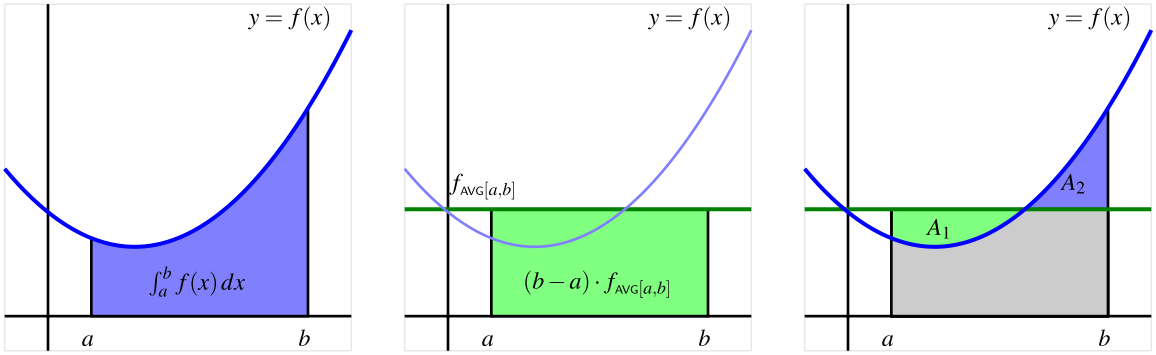


Figure 4.3.10. A function $y = f(x)$, the area it bounds, and its average value on $[a,b]$.

Consider Figure 4.3.10, where we see at left the shaded region whose area is $\int_a^b f(x)\ dx$, at center the shaded rectangle whose dimensions are $(b-a)$ by $f_{\text{AVG}[a,b]}$, and at right these two figures superimposed. Note that in dark green we show the horizontal line $y = f_{\text{AVG}[a,b]}$. Thus, the area of the green rectangle is given by $(b-a) \cdot f_{\text{AVG}[a,b]}$, which is precisely the value of $\int_a^b f(x)\ dx$. The area of the blue region in the left figure is the same as the area of the green rectangle in the center figure. We can also observe that the areas A_1 and A_2 in the rightmost figure appear to be equal. Thus, knowing the average value of a function enables us to construct a rectangle whose area is the same as the value of the definite integral of the function on the interval. The java applet [3] at http://gvsu.edu/s/az provides an opportunity to explore how the average value of the function changes as the interval changes,

through an image similar to that found in Figure 4.3.10.

David Austin, http://gvsu.edu/s/5r.

Activity 4.3.4

Suppose that $v(t) = \sqrt{4-(t-2)^2}$ tells us the instantaneous velocity of a moving object on the interval $0 \le t \le 4$, where t is measured in minutes and v is measured in meters per minute.

1. Sketch an accurate graph of $y = v(t)$. What kind of curve is $y = \sqrt{4-(t-2)^2}$?
2. Evaluate $\int_0^4 v(t)\ dt$ exactly.
3. In terms of the physical problem of the moving object with velocity $v(t)$, what is the meaning of $\int_0^4 v(t)\ dt$? Include units on your answer.
4. Determine the exact average value of $v(t)$ on $[0,4]$. Include units on your answer.
5. Sketch a rectangle whose base is the line segment from $t=0$ to $t=4$ on the t-axis such that the rectangle's area is equal to the value of $\int_0^4 v(t)\ dt$. What is the rectangle's exact height?
6. How can you use the average value you found in (d) to compute the total distance traveled by the moving object over $[0,4]$?

Summary

- Any Riemann sum of a continuous function f on an interval $[a,b]$ provides an estimate of the net signed area bounded by the function and the horizontal axis on the interval. Increasing the number of subintervals in the Riemann sum improves the accuracy of this estimate, and letting the number of subintervals increase without bound results in the values of the corresponding Riemann sums approaching the exact value of the enclosed net signed area.
- When we take the limit of Riemann sums, we arrive at what we call the definite integral of f over the interval $[a,b]$. In particular, the symbol $\int_a^b f(x)\ dx$ denotes the definite integral of f over $[a,b]$, and this quantity is defined by the equation

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x,$$

  where $\Delta x = \frac{b-a}{n}$, $x_i = a + i\Delta x$ (for $i = 0, \ldots, n$), and x_i^* satisfies $x_{i-1} \le x_i^* \le x_i$ (for $i = 1, \ldots, n$).

- The definite integral $\int_a^b f(x)\ dx$ measures the exact net signed area bounded by f and the horizontal axis on $[a,b]$; in addition, the value of the definite integral is related to what we call the average value of the function on $[a,b]$: $f_{\text{AVG}[a,b]} = \frac{1}{b-a} \cdot \int_a^b f(x)\ dx$. In the setting where we consider the integral of a velocity function v, $\int_a^b v(t)\ dt$ measures the exact change in position of the moving object on $[a,b]$; when v is nonnegative, $\int_a^b v(t)\ dt$ is the object's distance traveled on $[a,b]$.
- The definite integral is a sophisticated sum, and thus has some of the same natural properties that finite sums have. Perhaps most important of these is how the definite integral respects sums and constant multiples of functions, which can be summarized by the rule

$$\int_a^b [cf(x) \pm kg(x)]\,dx = c\int_a^b f(x)\,dx \pm k\int_a^b g(x)\,dx$$

  where f and g are continuous functions on $[a,b]$ and c and k are arbitrary constants.

This page titled 4.3: The Definite Integral is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 4.3: The Definite Integral by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

4.4: The Fundamental Theorem of Calculus

Motivating Questions

- How can we find the exact value of a definite integral without taking the limit of a Riemann sum?
- What is the statement of the Fundamental Theorem of Calculus, and how do antiderivatives of functions play a key role in applying the theorem?
- What is the meaning of the definite integral of a rate of change in contexts other than when the rate of change represents velocity?

Much of our work in Chapter 4 has been motivated by the velocity-distance problem: if we know the instantaneous velocity function, $v(t)$, for a moving object on a given time interval $[a, b]$, can we determine the distance it traveled on $[a, b]$? If the velocity function is nonnegative on $[a, b]$, the area bounded by $y = v(t)$ and the t-axis on $[a, b]$ is equal to the distance traveled. This area is also the value of the definite integral $\int_a^b v(t)\ dt$. If the velocity is sometimes negative, the total area bounded by the velocity function still tells us distance traveled, while the net signed area tells us the object's change in position.

For instance, for the velocity function in Figure 4.4.1, the total distance D traveled by the moving object on $[a, b]$ is

$$D = A_1 + A_2 + A_3,$$

and the total change in the object's position is

$$s(b) - s(a) = A_1 - A_2 + A_3.$$

The areas A_1, A_2, and A_3 are each given by definite integrals, which may be computed by limits of Riemann sums (and in special circumstances by geometric formulas).

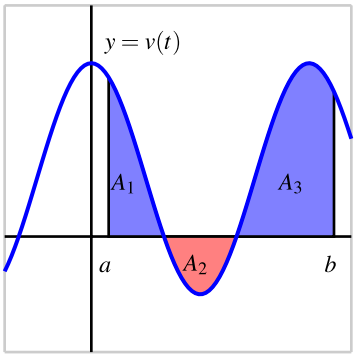


Figure 4.4.1. A velocity function that is sometimes negative.

We turn our attention to an alternate approach.

Preview Activity 4.4.1

A student with a third floor dormitory window 32 feet off the ground tosses a water balloon straight up in the air with an initial velocity of 16 feet per second. It turns out that the instantaneous velocity of the water balloon is given by $v(t) = -32t + 16$, where v is measured in feet per second and t is measured in seconds.

1. Let $s(t)$ represent the height of the water balloon above ground at time t, and note that s is an antiderivative of v. That is, v is the derivative of s: $s'(t) = v(t)$. Find a formula for $s(t)$ that satisfies the initial condition that the balloon is tossed from 32 feet above ground. In other words, make your formula for s satisfy $s(0) = 32$.
2. When does the water balloon reach its maximum height? When does it land?
3. Compute $s(\frac{1}{2}) - s(0)$, $s(2) - s(\frac{1}{2})$, and $s(2) - s(0)$. What do these represent?
4. What is the total vertical distance traveled by the water balloon from the time it is tossed until the time it lands?
5. Sketch a graph of the velocity function $y = v(t)$ on the time interval $[0, 2]$. What is the total net signed area bounded by $y = v(t)$ and the t-axis on $[0, 2]$? Answer this question in two ways: first by using your work above, and then by using a familiar geometric formula to compute areas of certain relevant regions.

The Fundamental Theorem of Calculus

Suppose we know the position function $s(t)$ and the velocity function $v(t)$ of an object moving in a straight line, and for the moment let us assume that $v(t)$ is positive on $[a, b]$. Then, as shown in Figure 4.4.2, we know two different ways to compute the distance, D, the object travels: one is that $D = s(b) - s(a)$, the object's change in position. The other is the area under the velocity curve, which is given by the definite integral, so $D = \int_a^b v(t)\ dt$.

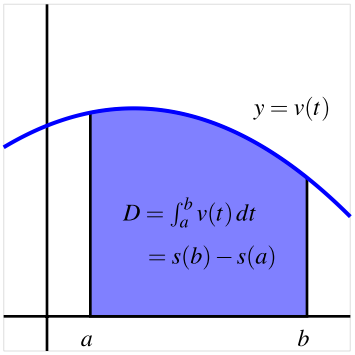


Figure 4.4.2. Finding distance traveled when we know a velocity function v.

Since both of these expressions tell us the distance traveled, it follows that they are equal, so

$$s(b) - s(a) = \int_a^b v(t)\, dt. \tag{4.4.1}$$

Equation (4.4.1) holds even when velocity is sometimes negative, because $s(b) - s(a)$,the object's change in position, is also measured by the net signed area on $[a, b]$ which is given by $\int_a^b v(t)\ dt$.

Perhaps the most powerful fact Equation (4.4.1) reveals is that we can compute the integral's value if we can find a formula for s. Remember, s and v are related by the fact that v is the derivative of s, or equivalently that s is an antiderivative of v.

Example 4.4.1

Determine the exact distance traveled on $[1, 5]$ by an object with velocity function $v(t) = 3t^2 + 40$ feet per second. The distance traveled on the interval $[1, 5]$ is given by

$$D = \int_1^5 v(t)\, dt = \int_1^5 (3t^2 + 40)\, dt = s(5) - s(1),$$

where s is an antiderivative of v. Now, the derivative of t^3 is $3t^2$ and the derivative of $40t$ is 40, so it follows that $s(t) = t^3 + 40t$ is an antiderivative of v. Therefore,

$$D = \int_1^5 3t^2 + 40\,dt = s(5) - s(1)$$
$$= (5^3 + 40 \cdot 5) - (1^3 + 40 \cdot 1) = 284 \text{ feet.}$$

Note the key lesson of Example 4.4.1: to find the distance traveled, we need to compute the area under a curve, which is given by the definite integral. But to evaluate the integral, we can find an antiderivative, s, of the velocity function, and then compute the total change in s on the interval. In particular, we can evaluate the integral without computing the limit of a Riemann sum.

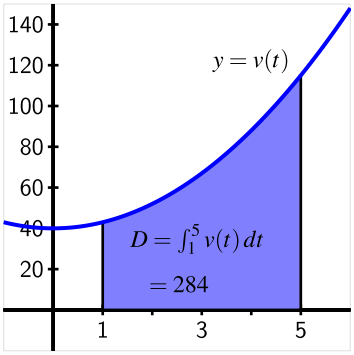


Figure 4.4.3. The exact area of the region enclosed by $v(t) = 3t^2 + 40$ on $[1, 5]$.

It will be convenient to have a shorthand symbol for a function's antiderivative. For a continuous function f, we will often denote an antiderivative of f by F, so that $F'(x) = f(x)$ for all relevant x. Using the notation V in place of s (so that V is an antiderivative of v) in Equation (4.4.11), we can write

$$V(b) - V(a) = \int_a^b v(t)\,dt. \tag{4.4.2}$$

Now, to evaluate the definite integral $\int_a^b f(x)\ dx$ for an arbitrary continuous function f, we could certainly think of f as representing the velocity of some moving object, and x as the variable that represents time. But Equations (4.4.1) and (4.4.2) hold for any continuous velocity function, even when v is sometimes negative. So Equation (4.4.2) offers a shortcut route to evaluating any definite integral, provided that we can find an antiderivative of the integrand. The Fundamental Theorem of Calculus (FTC) summarizes these observations.

Fundamental Theorem of Calculus

If f is a continuous function on $[a, b]$, and F is any antiderivative of f, then $\int_a^b f(x)\ dx = F(b) - F(a)$.

A common alternate notation for $F(b) - F(a)$ is

$$F(b) - F(a) = F(x)\Big|_a^b,$$

where we read the righthand side as “the function F evaluated from a to b.” In this notation, the FTC says that

$$\int_a^b f(x)\,dx = F(x)\Big|_a^b.$$

The FTC opens the door to evaluating a wide range of integrals if we can find an antiderivative F for the integrand f. For instance since $\frac{d}{dx}[\frac{1}{3}x^3] = x^2$, the FTC tells us that

$$\begin{aligned}\int_0^1 x^2\,dx &= \frac{1}{3}x^3\Big|_0^1 \\ &= \frac{1}{3}(1)^3 - \frac{1}{3}(0)^3 \\ &= \frac{1}{3}.\end{aligned}$$

But finding an antiderivative can be far from simple; it is often difficult or even impossible. While we can differentiate just about any function, even some relatively simple functions don't have an elementary antiderivative. A significant portion of integral calculus (which is the main focus of second semester college calculus) is devoted to the problem of finding antiderivatives.

Activity 4.4.2

Use the Fundamental Theorem of Calculus to evaluate each of the following integrals exactly. For each, sketch a graph of the integrand on the relevant interval and write one sentence that explains the meaning of the value of the integral in terms of the (net signed) area bounded by the curve.

1. $\int_{-1}^{4} (2-2x)\ dx$
2. $\int_{0}^{\frac{\pi}{2}} \sin(x)\ dx$
3. $\int_{0}^{1} e^x\ dx$
4. $\int_{-1}^{1} x^5\ dx$
5. $\int_{0}^{2} (3x^3 - 2x^2 - e^x)\ dx$

Basic antiderivatives

The general problem of finding an antiderivative is difficult. In part, this is due to the fact that we are trying to undo the process of differentiating, and the undoing is much more difficult than the doing. For example, while it is evident that an antiderivative of $f(x) = \sin(x)$ is $F(x) = -\cos(x)$ and that an antiderivative of $g(x) = x^2$ is $G(x) = \frac{1}{3}x^3$, combinations of f and g can be far more complicated. Consider the functions

$$5\sin(x) - 4x^2,\ x^2\sin(x),\ \frac{\sin(x)}{x^2},\ \text{and}\ \sin(x^2).$$

What is involved in trying to find an antiderivative for each? From our experience with derivative rules, we know that derivatives of sums and constant multiples of basic functions are simple to execute, but derivatives involving products, quotients, and composites of familiar functions are more complicated. Therefore, it stands to reason that antidifferentiating products, quotients, and composites of basic functions may be even more challenging. We defer our study of all but the most elementary antiderivatives to later in the text.

We do note that whenever we know the derivative of a function, we have a *function-derivative pair*, so we also know the antiderivative of a function. For instance, since we know that

$$\frac{d}{dx}[-\cos(x)] = \sin(x),$$

we also know that $F(x) = -\cos(x)$ is an antiderivative of $f(x) = \sin(x)$. F and f together form a function-derivative pair. Clearly, every basic derivative rule leads us to such a pair, and thus to a known antiderivative.

In Activity 4.4.3, we will construct a list of the basic antiderivatives we know at this time. Those rules will help us antidifferentiate sums and constant multiples of basic functions. For example, since $-\cos(x)$ is an antiderivative of $\sin(x)$ and $\frac{1}{3}x^3$ is an antiderivative of x^2, it follows that

$$F(x) = -5\cos(x) - \frac{4}{3}x^3$$

is an antiderivative of $f(x) = 5\sin(x) - 4x^2$, by the sum and constant multiple rules for differentiation.

Finally, before proceeding to build a list of common functions whose antiderivatives we know, we recall that each function has more than one antiderivative. Because the derivative of any constant is zero, we may add a constant of our choice to any antiderivative. For instance, we know that $G(x) = \frac{1}{3}x^3$ is an antiderivative of $g(x) = x^2$. But we could also have chosen $G(x) = \frac{1}{3}x^3 + 7$, since in this case as well, $G'(x) = x^2$. If $g(x) = x^2$, we say that the *general antiderivative* of g is

$$G(x) = \frac{1}{3}x^3 + C,$$

where C represents an arbitrary real number constant. Regardless of the formula for g, including $+C$ in the formula for its antiderivative G results in the most general possible antiderivative.

Our current interest in antiderivatives is so that we can evaluate definite integrals by the Fundamental Theorem of Calculus. For that task, the constant C is irrelevant, and we usually omit it. To see why, consider the definite integral

$$\int_0^1 x^2\,dx.$$

For the integrand $g(x) = x^2$, suppose we find and use the general antiderivative $G(x) = \frac{1}{3}x^3 + C$. Then, by the FTC,

$$\begin{aligned}\int_0^1 x^2\,dx &= \frac{1}{3}x^3 + C\Big|_0^1 \\ &= \left(\frac{1}{3}(1)^3 + C\right) - \left(\frac{1}{3}(0)^3 + C\right) \\ &= \frac{1}{3} + C - 0 - C \\ &= \frac{1}{3}.\end{aligned}$$

Observe that the C-values appear as opposites in the evaluation of the integral and thus do not affect the definite integral's value.

In the following activity, we work to build a list of basic functions whose antiderivatives we already know.

Activity 4.4.3

Use your knowledge of derivatives of basic functions to complete Table 4.4.1 of antiderivatives. For each entry, your task is to find a function F whose derivative is the given function f. When finished, use the FTC and the results in the table to evaluate the three given definite integrals.

Table 4.4.1. Familiar basic functions and their antiderivatives.

given function, $f(x)$	antiderivative, $F(x)$
k, (k is constant)	
x^n, $n \neq -1$	
$\frac{1}{x}$, $x > 0$	
$\sin(x)$	
$\cos(x)$	
$\sec(x)\tan(x)$	
$\csc(x)\cot(x)$	
$\sec^2(x)$	

given function, $f(x)$	antiderivative, $F(x)$
$\csc^2(x)$	
e^x	
$a^x\ (a>1)$	
$\frac{1}{1+x^2}$	
$\frac{1}{\sqrt{1-x^2}}$	

a. $\int_0^1 \left(x^3 - x - e^x + 2\right)\ dx$
b. $\int_0^{\pi/3} \left(2\sin(t) - 4\cos(t) + \sec^2(t) - \pi\right)\ dt$
c. $\int_0^1 \left(\sqrt{x} - x^2\right)\ dx$

The Total Change Theorem

Let us review three interpretations of the definite integral.

- For a moving object with instantaneous velocity $v(t)$, the object's change in position on the time interval $[a,b]$ is given by $\int_a^b v(t)\ dt$, and whenever $v(t) \geq 0$ on $[a,b]$, $\int_a^b v(t)\ dt$ tells us the total distance traveled by the object on $[a,b]$.
- For any continuous function f, its definite integral $\int_a^b f(x)\ dx$ represents the net signed area bounded by $y = f(x)$ and the x-axis on $[a,b]$, where regions that lie below the x-axis have a minus sign associated with their area.
- The value of a definite integral is linked to the average value of a function: for a continuous function f on $[a,b]$, its average value $f_{\text{AVG}[a,b]}$ is given by

$$f_{\text{AVG}[a,b]} = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

The Fundamental Theorem of Calculus now enables us to evaluate exactly (without taking a limit of Riemann sums) any definite integral for which we are able to find an antiderivative of the integrand.

A slight change in perspective allows us to gain even more insight into the meaning of the definite integral. Recall Equation (4.4.1 2), where we wrote the Fundamental Theorem of Calculus for a velocity function v with antiderivative V as

$$V(b) - V(a) = \int_a^b v(t)\,dt.$$

If we instead replace V with s (which represents position) and replace v with s' (since velocity is the derivative of position), Equation (4.4.12) then reads as

$$s(b) - s(a) = \int_a^b s'(t)\,dt. \tag{4.4.3}$$

In words, this version of the FTC tells us that the total change in an object's position function on a particular interval is given by the definite integral of the position function's derivative over that interval.

Of course, this result is not limited to only the setting of position and velocity. Writing the result in terms of a more general function f, we have the Total Change Theorem.

Total Change Theorem

If f is a continuously differentiable function on $[a,b]$ with derivative f', then $f(b) - f(a) = \int_a^b f'(x)\ dx$. That is, the definite integral of the rate of change of a function on $[a,b]$ is the total change of the function itself on $[a,b]$.

The Total Change Theorem tells us more about the relationship between the graph of a function and that of its derivative. Recall that heights on the graph of the derivative function are equal to slopes on the graph of the function itself. If instead we know f' and are seeking information about f, we can say the following:

> *differences in heights on f correspond to net signed areas bounded by f'.*

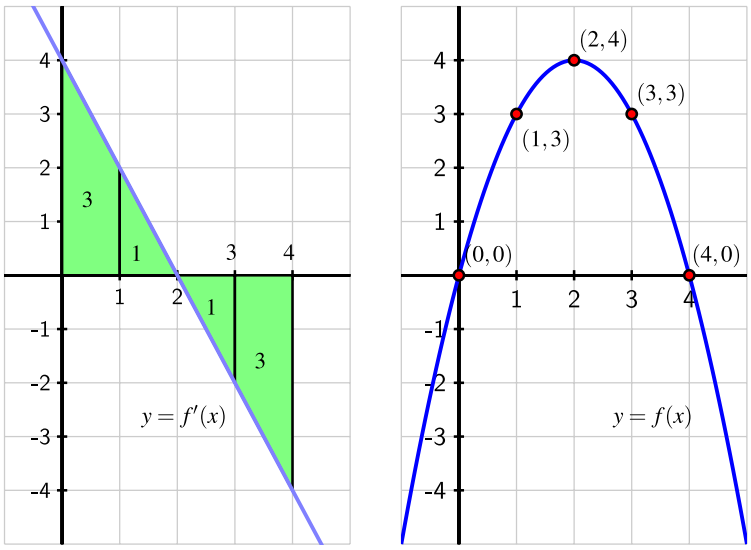


Figure 4.4.4. The graphs of $f'(x) = 4 - 2x$ (at left) and an antiderivative $f(x) = 4x - x^2$ at right. Differences in heights on f correspond to net signed areas bounded by f'.

To see why this is so, consider the difference $f(1) - f(0)$. This value is 3, because $f(1) = 3$ and $f(0) = 0$, but also because the net signed area bounded by $y = f'(x)$ on $[0, 1]$ is 3. That is,

$$f(1) - f(0) = \int_0^1 f'(x)\,dx.$$

In addition to this observation about area, the Total Change Theorem enables us to answer questions about a function whose rate of change we know.

Example 4.4.2

Suppose that pollutants are leaking out of an underground storage tank at a rate of $r(t)$ gallons/day, where t is measured in days. It is conjectured that $r(t)$ is given by the formula $r(t) = 0.0069t^3 - 0.125t^2 + 11.079$ over a certain 12-day period. The graph of $y = r(t)$ is given in Figure 4.4.5. What is the meaning of $\int_4^{10} r(t)\ dt$ and what is its value? What is the average rate at which pollutants are leaving the tank on the time interval $4 \le t \le 10$?

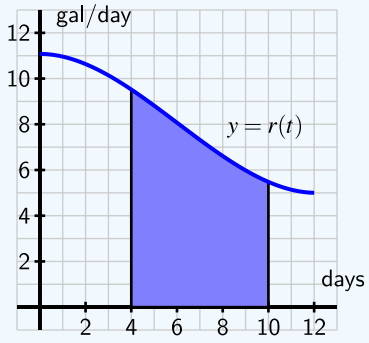


Figure 4.4.5. The rate $r(t)$ of pollution leaking from a tank, measured in gallons per day.

Answer available at the end of this section.

Activity 4.4.4

During a 40-minute workout, a person riding an exercise machine burns calories at a rate of c calories per minute, where the function $y = c(t)$ is given in Figure 4.4.6. On the interval $0 \le t \le 10$, the formula for c is $c(t) = -0.05t^2 + t + 10$, while on $30 \le t \le 40$, its formula is $c(t) = -0.05t^2 + 3t - 30$.

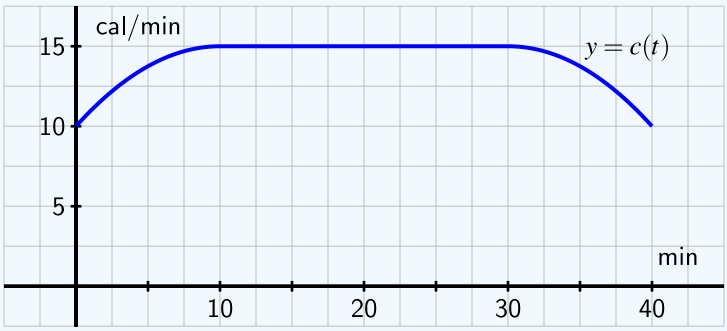


Figure 4.4.6. The rate $c(t)$ at which a person exercising burns calories, measured in calories per minute.

1. What is the exact total number of calories the person burns during the first 10 minutes of her workout?
2. Let $C(t)$ be an antiderivative of $c(t)$. What is the meaning of $C(40) - C(0)$ in the context of the person exercising? Include units on your answer.
3. Determine the exact average rate at which the person burned calories during the 40-minute workout.
4. At what time(s), if any, is the instantaneous rate at which the person is burning calories equal to the average rate at which she burns calories, on the time interval $0 \le t \le 40$?

Summary

- We can find the exact value of a definite integral without taking the limit of a Riemann sum or using a familiar area formula by finding the antiderivative of the integrand, and hence applying the Fundamental Theorem of Calculus.
- The Fundamental Theorem of Calculus says that if f is a continuous function on $[a, b]$ and F is an antiderivative of f, then

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

Hence, if we can find an antiderivative for the integrand f, evaluating the definite integral comes from simply computing the change in F on $[a, b]$.

- A slightly different perspective on the FTC allows us to restate it as the Total Change Theorem, which says that

$$\int_a^b f'(x)\,dx = f(b) - f(a),$$

for any continuously differentiable function f. This means that the definite integral of the instantaneous rate of change of a function f on an interval $[a, b]$ is equal to the total change in the function f on $[a, b]$.

Answers

Example 4.4.2

Since $r(t) \geq 0$, the value of $\int_4^{10} r(t)\ dt$ is the area under the curve on the interval $[4, 10]$. A Riemann sum for this area will have rectangles with heights measured in gallons per day and widths measured in days, so the area of each rectangle will have units of

$$\frac{\text{gallons}}{\text{day}} \cdot \text{days} = \text{gallons}.$$

Thus, the definite integral tells us the total number of gallons of pollutant that leak from the tank from day 4 to day 10. The Total Change Theorem tells us the same thing: if we let $R(t)$ denote the total number of gallons of pollutant that have leaked from the tank up to day t, then $R'(t) = r(t)$, and

$$\int_4^{10} r(t)\,dt = R(10) - R(4),$$

the number of gallons that have leaked from day 4 to day 10.

To compute the exact value of the integral, we use the Fundamental Theorem of Calculus. Antidifferentiating $r(t) = 0.0069t^3 - 0.125t^2 + 11.079$, we find that

$$\int_4^{10} 0.0069t^3 - 0.125t^2 + 11.079\,dt = 0.0069 \cdot \frac{1}{4}\,t^4 - 0.125 \cdot \frac{1}{3}t^3 + 11.079t\Big|_4^{10}$$

$$\approx 44.282.$$

Thus, approximately 44.282 gallons of pollutant leaked over the six day time period.

To find the average rate at which pollutant leaked from the tank over $4 \leq t \leq 10$, we compute the average value of r on $[4, 10]$. Thus,

$$r_{\text{AVG}[4,10]} = \frac{1}{10-4}\int_4^{10} r(t)\,dt \approx \frac{44.282}{6} = 7.380$$

gallons per day.

This page titled 4.4: The Fundamental Theorem of Calculus is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 4.4: The Fundamental Theorem of Calculus by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

4.5: The Definite Integral (Exercises)

4.1: Determining Distance Traveled from Velocity

1. Estimating distance traveled from velocity data

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

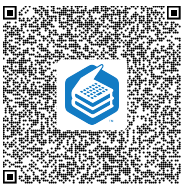
View this Interactive Content online



2. Distance from a linear veloity function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Change in position from a linear velocity function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Comparing distance traveled from velocity graphs

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5. Finding average acceleration from velocity data

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



6. Change in position from a quadratic velocity function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

7

Along the eastern shore of Lake Michigan from Lake Macatawa (near Holland) to Grand Haven, there is a bike path that runs almost directly north-south. For the purposes of this problem, assume the road is completely straight, and that the function $s(t)$ tracks the position of the biker along this path in miles north of Pigeon Lake, which lies roughly halfway between the ends of the bike path.

Suppose that the biker's velocity function is given by the graph in Figure 4.1.11 on the time interval $0 \leq t \leq 4$ (where t is measured in hours), and that $s(0) = 1$.

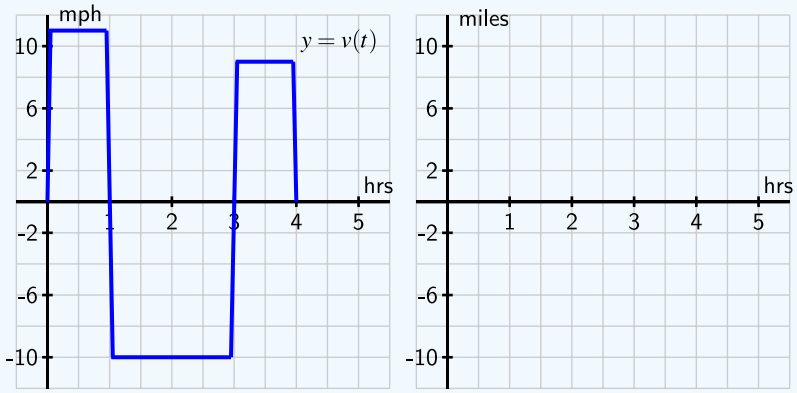


Figure 4.1.11. The graph of the biker's velocity, $y = v(t)$, at left. At right, axes to plot an approximate sketch of $y = s(t)$.

1. Approximately how far north of Pigeon Lake was the cyclist when she was the greatest distance away from Pigeon Lake? At what time did this occur?
2. What is the cyclist's total change in position on the time interval $0 \leq t \leq 2$? At $t = 2$, was she north or south of Pigeon Lake?
3. What is the total distance the biker traveled on $0 \leq t \leq 4$? At the end of the ride, how close was she to the point at which she started?
4. Sketch an approximate graph of $y = s(t)$, the position function of the cyclist, on the interval $0 \leq t \leq 4$. Label at least four important points on the graph of s.

8

A toy rocket is launched vertically from the ground on a day with no wind. The rocket's vertical velocity at time t (in seconds) is given by $v(t) = 500 - 32t$ feet/sec.

1. At what time after the rocket is launched does the rocket's velocity equal zero? Call this time value a. What happens to the rocket at $t = a$?
2. Find the value of the total area enclosed by $y = v(t)$ and the t-axis on the interval $0 \leq t \leq a$. What does this area represent in terms of the physical setting of the problem?
3. Find an antiderivative s of the function v. That is, find a function s such that $s'(t) = v(t)$.

4. Compute the value of $s(a) - s(0)$. What does this number represent in terms of the physical setting of the problem?
5. Compute $s(5) - s(1)$. What does this number tell you about the rocket's flight?

9

An object moving along a horizontal axis has its instantaneous velocity at time t in seconds given by the function v pictured in Figure 4.1.12, where v is measured in feet/sec. Assume that the curves that make up the parts of the graph of $y = v(t)$ are either portions of straight lines or portions of circles.

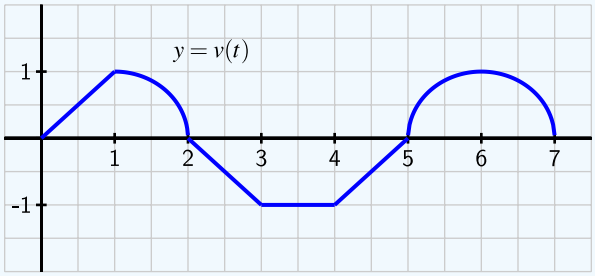


Figure 4.1.12. The graph of $y = v(t)$, the velocity function of a moving object.

1. Determine the exact total distance the object traveled on $0 \leq t \leq 2$.
2. What is the value and meaning of $s(5) - s(2)$, where $y = s(t)$ is the position function of the moving object?
3. On which time interval did the object travel the greatest distance: $[0, 2]$, $[2, 4]$, or $[5, 7]$?
4. On which time interval(s) is the position function s increasing? At which point(s) does s achieve a relative maximum?

10

Filters at a water treatment plant become dirtier over time and thus become less effective; they are replaced every 30 days. During one 30-day period, the rate at which pollution passes through the filters into a nearby lake (in units of particulate matter per day) is measured every 6 days and is given in the following table. The time t is measured in days since the filters were replaced.

Table 4.1.13. Pollution data for the water filters.

Day, t	0	6	12	18	24	30
Rate of pollution in units per day, $p(t)$	7	8	10	13	18	35

1. Plot the given data on a set of axes with time on the horizontal axis and the rate of pollution on the vertical axis.
2. Explain why the amount of pollution that entered the lake during this 30-day period would be given exactly by the area bounded by $y = p(t)$ and the t-axis on the time interval $[0, 30]$.
3. Estimate the total amount of pollution entering the lake during this 30-day period. Carefully explain how you determined your estimate.

4.2: Riemann Sums

1. Evaluating Riemann sums for a quadratic function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Estimating distance traveled with a Riemann sum from data

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Writing basic Riemann sums

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4

Consider the function $f(x) = 3x + 4$.

1. Compute M_4 for $y = f(x)$ on the interval $[2, 5]$. Be sure to clearly identify the value of Δx, as well as the locations of $x_0, x_1, \ldots, x_4$. Include a careful sketch of the function and the corresponding rectangles being used in the sum.
2. Use a familiar geometric formula to determine the exact value of the area of the region bounded by $y = f(x)$ and the x -axis on $[2, 5]$.
3. Explain why the values you computed in (a) and (b) turn out to be the same. Will this be true if we use a number different than $n = 4$ and compute M_n? Will L_4 or R_4 have the same value as the exact area of the region found in (b)?
4. Describe the collection of functions g for which it will always be the case that M_n, regardless of the value of n, gives the exact net signed area bounded between the function g and the x-axis on the interval $[a, b]$.

5

Let S be the sum given by

$$S = ((1.4)^2 + 1) \cdot 0.4 + ((1.8)^2 + 1) \cdot 0.4 + ((2.2)^2 + 1) \cdot 0.4 + ((2.6)^2 + 1) \cdot 0.4 + ((3.0)^2 + 1) \cdot 0.4.$$

1. Assume that S is a right Riemann sum. For what function f and what interval $[a, b]$ is S this function's Riemann sum? Why?
2. How does your answer to (a) change if S is a left Riemann sum? a middle Riemann sum?
3. Suppose that S really is a right Riemann sum. What is geometric quantity does S approximate?

4. Use sigma notation to write a new sum R that is the right Riemann sum for the same function, but that uses twice as many subintervals as S.

6

A car traveling along a straight road is braking and its velocity is measured at several different points in time, as given in the following table.

Table 4.2.8. Data for the braking car.

seconds, t	0	0.3	0.6	0.9	1.2	1.5	1.8
Velocity in ft/sec, $v(t)$	100	88	74	59	40	19	0

1. Plot the given data on a set of axes with time on the horizontal axis and the velocity on the vertical axis.
2. Estimate the total distance traveled during the car the time brakes using a middle Riemann sum with 3 subintervals.
3. Estimate the total distance traveled on $[0, 1.8]$ by computing L_6, R_6, and $\frac{1}{2}(L_6 + R_6)$.
4. Assuming that $v(t)$ is always decreasing on $[0, 1.8]$, what is the maximum possible distance the car traveled before it stopped? Why?

7

The rate at which pollution escapes a scrubbing process at a manufacturing plant increases over time as filters and other technologies become less effective. For this particular example, assume that the rate of pollution (in tons per week) is given by the function r that is pictured in Figure 4.2.9.

1. Use the graph to estimate the value of M_4 on the interval $[0, 4]$.
2. What is the meaning of M_4 in terms of the pollution discharged by the plant?
3. Suppose that $r(t) = 0.5e^{0.5t}$. Use this formula for r to compute L_5 on $[0, 4]$.
4. Determine an upper bound on the total amount of pollution that can escape the plant during the pictured four week time period that is accurate within an error of at most one ton of pollution.

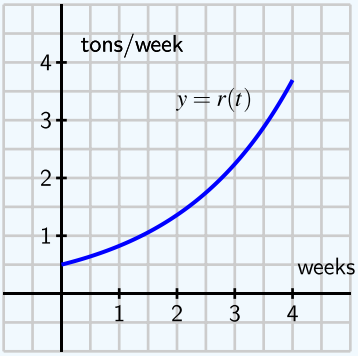


Figure 4.2.9. The rate, $r(t)$, of pollution in tons per week.

4.3: The Definite Integral

1. Evaluating definite integrals from graphical information

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Estimating definite integrals from a graph

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Finding the average value of a linear function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Finding the average value of a function given graphically

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Estimating a definite integral and average value from a graph

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6. Using rules to combine known integral values

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

7

The velocity of an object moving along an axis is given by the piecewise linear function v that is pictured in Figure 4.3.12. Assume that the object is moving to the right when its velocity is positive, and moving to the left when its velocity is negative. Assume that the given velocity function is valid for $t = 0$ to $t = 4$.

1. Write an expression involving definite integrals whose value is the total change in position of the object on the interval $[0, 4]$.
2. Use the provided graph of v to determine the value of the total change in position on $[0, 4]$.
3. Write an expression involving definite integrals whose value is the total distance traveled by the object on $[0, 4]$. What is the exact value of the total distance traveled on $[0, 4]$?
4. What is the object's exact average velocity on $[0, 4]$?
5. Find an algebraic formula for the object's position function on $[0, 1.5]$ that satisfies $s(0) = 0$.

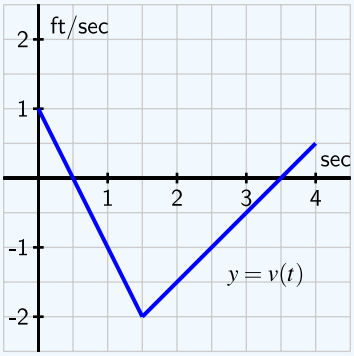


Figure 4.3.12. The velocity function of a moving object.

8

Suppose that the velocity of a moving object is given by $v(t) = t(t - 1)(t - 3)$, measured in feet per second, and that this function is valid for $0 \leq t \leq 4$.

1. Write an expression involving definite integrals whose value is the total change in position of the object on the interval $[0, 4]$.
2. Use appropriate technology (such as http://gvsu.edu/s/a9 [4]) to compute Riemann sums to estimate the object's total change in position on $[0, 4]$. Work to ensure that your estimate is accurate to two decimal places, and explain how you know this to be the case.
3. Write an expression involving definite integrals whose value is the total distance traveled by the object on $[0, 4]$.
4. Use appropriate technology to compute Riemann sums to estimate the object's total distance travelled on $[0, 4]$. Work to ensure that your estimate is accurate to two decimal places, and explain how you know this to be the case.
5. What is the object's average velocity on $[0, 4]$, accurate to two decimal places?

Marc Renault, Shippensburg University.

9

Consider the graphs of two functions f and g that are provided in Figure 4.3.13. Each piece of f and g is either part of a straight line or part of a circle.

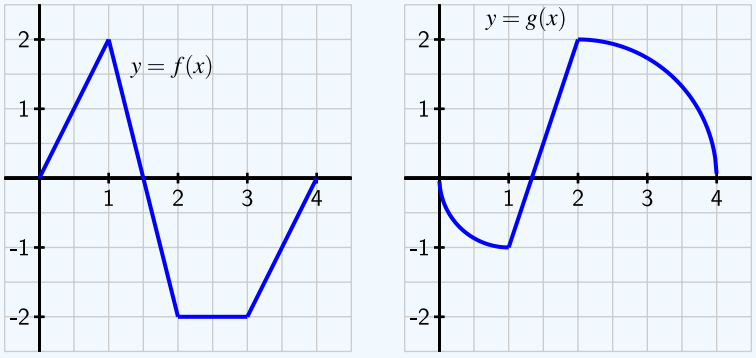


Figure 4.3.13. Two functions f and g.

1. Determine the exact value of $\int_0^1 [f(x) + g(x)]\ dx$.
2. Determine the exact value of $\int_1^4 [2f(x) - 3g(x)]\ dx$.
3. Find the exact average value of $h(x) = g(x) - f(x)$ on $[0, 4]$.
4. For what constant c does the following equation hold?

$$\int_0^4 c\,dx = \int_0^4 [f(x) + g(x)]\,dx$$

10

Let $f(x) = 3 - x^2$ and $g(x) = 2x^2$.

1. On the interval $[-1, 1]$, sketch a labeled graph of $y = f(x)$ and write a definite integral whose value is the exact area bounded by $y = f(x)$ on $[-1, 1]$.
2. On the interval $[-1, 1]$, sketch a labeled graph of $y = g(x)$ and write a definite integral whose value is the exact area bounded by $y = g(x)$ on $[-1, 1]$.
3. Write an expression involving a difference of definite integrals whose value is the exact area that lies between $y = f(x)$ and $y = g(x)$ on $[-1, 1]$.
4. Explain why your expression in (c) has the same value as the single integral $\int_{-1}^1 [f(x) - g(x)]\ dx$.
5. Explain why, in general, if $p(x) \geq q(x)$ for all x in $[a, b]$, the exact area between $y = p(x)$ and $y = q(x)$ is given by

$$\int_a^b [p(x) - q(x)]\,dx.$$

4.4: The Fundamental Theorem of Calculus

Finding exact displacement

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Evaluating the definite integral of a rational function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Evaluating the definite integral of a linear function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Evaluating the definite integral of a quadratic function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Simplifying an integrand before integrating

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Evaluating the definite integral of a trigonometric function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

7

The instantaneous velocity (in meters per minute) of a moving object is given by the function v as pictured in Figure $4.5.10$. Assume that on the interval $0 \le t \le 4$, $v(t)$ is given by $v(t) = -\frac{1}{4}t^3 + \frac{3}{2}t^2 + 1$, and that on every other interval v is piecewise linear, as shown.

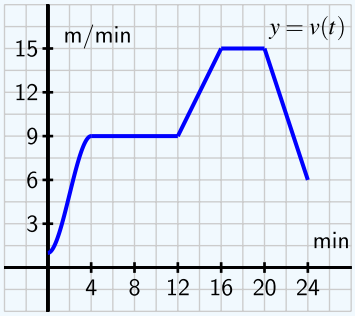


Figure 4.4.10. The velocity function of a moving body.

1. Determine the exact distance traveled by the object on the time interval $0 \le t \le 4$.
2. What is the object's average velocity on $[12, 24]$?
3. At what time is the object's acceleration greatest?
4. Suppose that the velocity of the object is increased by a constant value c for all values of t. What value of c will make the object's total distance traveled on $[12, 24]$ be 210 meters?

8

A function f is given piecewise by the formula

$$f(x) = \begin{cases} -x^2 + 2x + 1, & \text{if } 0 \le x < 2 \\ -x + 3, & \text{if } 2 \le x < 3 \\ x^2 - 8x + 15, & \text{if } 3 \le x \le 5 \end{cases}.$$

1. Determine the exact value of the net signed area enclosed by f and the x-axis on the interval $[2, 5]$.
2. Compute the exact average value of f on $[0, 5]$.
3. Find a formula for a function g on $5 \le x \le 7$ so that if we extend the above definition of f so that $f(x) = g(x)$ if $5 \le x \le 7$, it follows that $\int_0^7 f(x)\ dx = 0$.

9

When an aircraft attempts to climb as rapidly as possible, its climb rate (in feet per minute) decreases as altitude increases, because the air is less dense at higher altitudes. Given below is a table showing performance data for a certain single engine aircraft, giving its climb rate at various altitudes, where $c(h)$ denotes the climb rate of the airplane at an altitude h.

h (feet)	0	1000	2000	3000	4000	5000	6000	7000	8000	9000	10,000
c (ft/min)	925	875	830	780	730	685	635	585	535	490	440

Let a new function called $m(h)$ measure the number of minutes required for a plane at altitude h to climb the next foot of altitude.

1. Determine a similar table of values for $m(h)$ and explain how it is related to the table above. Be sure to explain the units.
2. Give a careful interpretation of a function whose derivative is $m(h)$. Describe what the input is and what the output is. Also, explain in plain English what the function tells us.
3. Determine a definite integral whose value tells us exactly the number of minutes required for the airplane to ascend to 10,000 feet of altitude. Clearly explain why the value of this integral has the required meaning.
4. Use the Riemann sum M_5 to estimate the value of the integral you found in (c). Include units on your result.

10

In Chapter 1, we showed that for an object moving along a straight line with position function $s(t)$, the object's "average velocity on the interval $[a, b]$" is given by

$$AV_{[a,b]} = \frac{s(b) - s(a)}{b - a}.$$

More recently in Chapter 4, we found that for an object moving along a straight line with velocity function $v(t)$, the object's "average value of its velocity function on $[a, b]$" is

$$v_{\text{AVG}[a,b]} = \frac{1}{b-a} \int_a^b v(t)\, dt.$$

Are the "average velocity on the interval $[a, b]$" and the "average value of the velocity function on $[a, b]$" the same thing? Why or why not? Explain.

11

In Table $4.5.5$ in Activity 4.4.3, we noted that for $x > 0$, the antiderivative of $f(x) = \frac{1}{x}$ is $F(x) = \ln(x)$. Here we observe that a key difference between $f(x)$ and $F(x)$ is that f is defined for all $x \neq 0$, while F is only defined for $x > 0$, and see how we can actually define the antiderivative of f for all values of x.

1. Suppose that $x < 0$, and let $G(x) = \ln(-x)$. Compute $G'(x)$.
2. Explain why G is an antiderivative of f for $x < 0$.
3. Let $H(x) = \ln(|x|)$, and recall that

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

   Explain why $H(x) = G(x)$ for $x < 0$ and $H(x) = F(x)$ for $x > 0$.

4. Now discuss why we say that the antiderivative of $\frac{1}{x}$ is $\ln(|x|)$ for all $x \neq 0$.

This page titled 4.5: The Definite Integral (Exercises) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 4.E: The Definite Integral (Exercises) has no license indicated.
- 4.1: Determining Distance Traveled from Velocity by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 4.2: Riemann Sums by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 4.3: The Definite Integral by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 4.4: The Fundamental Theorem of Calculus by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

5: FINDING ANTIDERIVATIVES AND EVALUATING INTEGRALS

5.1: Construction Accurate Graphs of Antiderivatives

5.2: The Second Fundamental Theorem of Calculus

5.3: Integration by Substitution

5.4: Integration by Parts

5.5: Other Options for Finding Algebraic Derivatives

5.6: Numerical Integration

5.7: Finding Antiderivatives and Evaluating Integrals (Exercises)

This page titled 5: Finding Antiderivatives and Evaluating Integrals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 5: Finding Antiderivatives and Evaluating Integrals by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

5.1: Construction Accurate Graphs of Antiderivatives

Motivating Questions

- Given the graph of a function's derivative, how can we construct a completely accurate graph of the original function?
- How many antiderivatives does a given function have? What do those antiderivatives all have in common?
- Given a function f, how does the rule $A(x) = \int_0^x f(t)\ dt$ define a new function A?

A recurring theme in our discussion of differential calculus has been the question "Given information about the derivative of an unknown function f, how much information can we obtain about f itself?" In Activity 1.8.3, the graph of $y = f'(x)$ was known (along with the value of f at a single point) and we endeavored to sketch a possible graph of f near the known point. In Example 3.1.2 — we investigated how the first derivative test enables us to use information about f' to determine where the original function f is increasing and decreasing, as well as where f has relative extreme values. If we know a formula or graph of f', by computing f'' we can find where the original function f is concave up and concave down. Thus, knowing f' and f'' enables us to understand the shape of the graph of f.

We returned to this question in even more detail in Section 4.1. In that setting, we knew the instantaneous velocity of a moving object and worked to determine as much as possible about the object's position function. We found connections between the net signed area under the velocity function and the corresponding change in position of the function, and the Total Change Theorem further illuminated these connections between f' and f, showing that the total change in the value of f over an interval $[a, b]$ is determined by the net signed area bounded by f' and the x-axis on the same interval.

In what follows, we explore the situation where we possess an accurate graph of the derivative function along with a single value of the function f. From that information, we'd like to determine a graph of f that shows where f is increasing, decreasing, concave up, and concave down, and also provides an accurate function value at any point.

Preview Activity 5.1.1

Suppose that the following information is known about a function f: the graph of its derivative, $y = f'(x)$, is given in Figure 5.1.1. Further, assume that f' is piecewise linear (as pictured) and that for $x \leq 0$ and $x \geq 6$, $f'(x) = 0$. Finally, it is given that $f(0) = 1$.

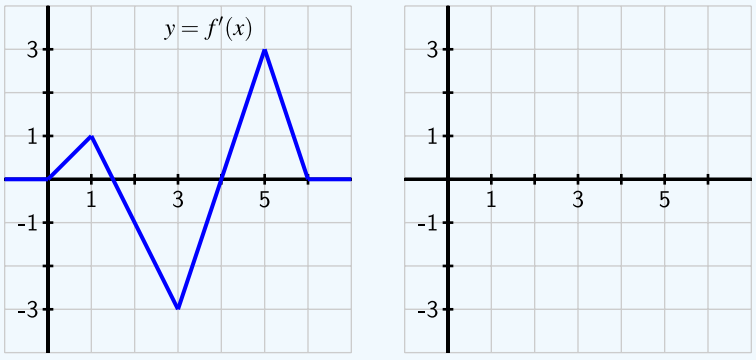


Figure 5.1.1. At left, the graph of $y = f'(x)$; at right, axes for plotting $y = f(x)$.

1. On what interval(s) is f an increasing function? On what intervals is f decreasing?
2. On what interval(s) is f concave up? concave down?
3. At what point(s) does f have a relative minimum? a relative maximum?
4. Recall that the Total Change Theorem tells us that

$$f(1) - f(0) = \int_0^1 f'(x)\,dx.$$

What is the exact value of $f(1)$?

5. Use the given information and similar reasoning to that in (d) to determine the exact value of $f(2)$, $f(3)$, $f(4)$, $f(5)$, and $f(6)$.
6. Based on your responses to all of the preceding questions, sketch a complete and accurate graph of $y = f(x)$ on the axes provided, being sure to indicate the behavior of f for $x < 0$ and $x > 6$.

Constructing the graph of an antiderivative

Preview Activity 5.1.1 demonstrates that when we can find the exact area under the graph of a function on any given interval, it is possible to construct a graph of the function's antiderivative. That is, we can find a function whose derivative is given. We can now determine not only the overall shape of the antiderivative graph, but also the actual *height* of the graph at any point of interest.

This is a consequence of the Fundamental Theorem of Calculus: if we know a function f and the value of the antiderivative F at some starting point a, we can determine the value of $F(b)$ via the definite integral. Since $F(b) - F(a) = \int_a^b f(x)\ dx$, it follows that

$$F(b) = F(a) + \int_a^b f(x)\,dx. \tag{5.1.1}$$

We can also interpret the equation $F(b) - F(a) = \int_a^b f(x)\ dx$ in terms of the graphs of f and F as follows. On an interval $[a, b]$,

> *differences in heights on the graph of the antiderivative given by $F(b) - F(a)$ correspond to the net signed area bounded by the original function on the interval $[a, b]$, which is given by $\int_a^b f(x)\ dx$.*

Activity 5.1.2

Suppose that the function $y = f(x)$ is given by the graph shown in Figure 5.1.2, and that the pieces of f are either portions of lines or portions of circles. In addition, let F be an antiderivative of f and say that $F(0) = -1$. Finally, assume that for $x \le 0$ and $x \ge 7$, $f(x) = 0$.

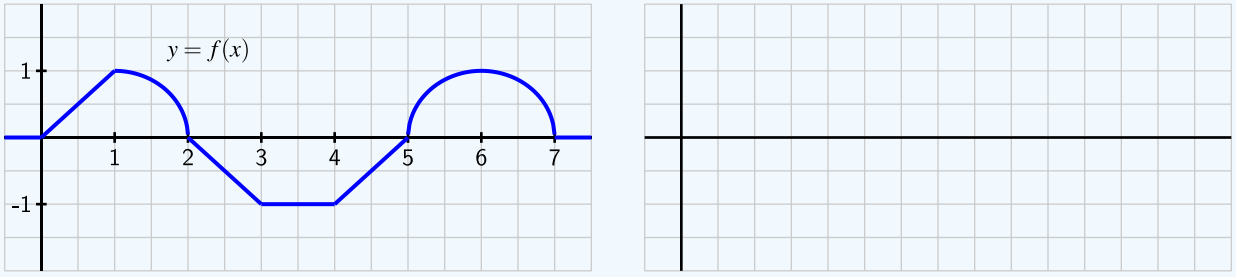


Figure 5.1.2. At left, the graph of $y = f(x)$.

1. On what interval(s) is F an increasing function? On what intervals is F decreasing?
2. On what interval(s) is F concave up? concave down? neither?
3. At what point(s) does F have a relative minimum? a relative maximum?
4. Use the given information to determine the exact value of $F(x)$ for $x = 1, 2, \ldots, 7$. In addition, what are the values of $F(-1)$ and $F(8)$?
5. Based on your responses to all of the preceding questions, sketch a complete and accurate graph of $y = F(x)$ on the axes provided, being sure to indicate the behavior of F for $x < 0$ and $x > 7$. Clearly indicate the scale on the vertical and horizontal axes of your graph.

6. What happens if we change one key piece of information: in particular, say that G is an antiderivative of f and $G(0) = 0$. How (if at all) would your answers to the preceding questions change? Sketch a graph of G on the same axes as the graph of F you constructed in (e).

Multiple antiderivatives of a single function

In the final question of Activity 5.1.2, we encountered a very important idea: a function f has more than one antiderivative. Each antiderivative of f is determined uniquely by its value at a single point. For example, suppose that f is the function given at left in Figure 5.1.3, and suppose further that F is an antiderivative of f that satisfies $F(0) = 1$.

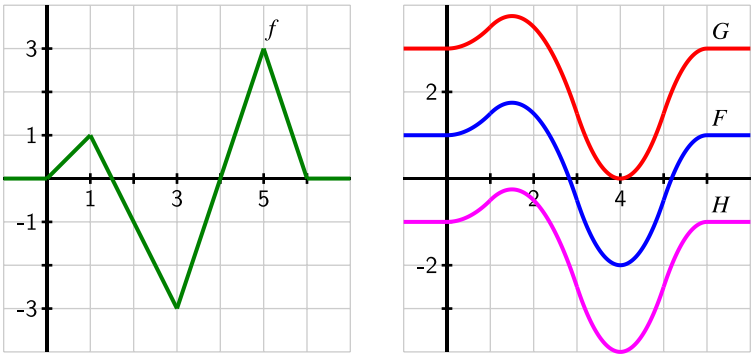


Figure 5.1.3. At left, the graph of $y = f(x)$. At right, three different antiderivatives of f.

Then, using Equation (5.1.1), we can compute

$$\begin{aligned} F(1) &= F(0) + \int_0^1 f(x)\,dx \\ &= 1 + 0.5 \\ &= 1.5. \end{aligned}$$

Similarly, $F(2) = 1.5$, $F(3) = -0.5$, $F(4) = -2$, $F(5) = -0.5$, and $F(6) = 1$. In addition, we can use the fact that $F' = f$ to ascertain where F is increasing and decreasing, concave up and concave down, and has relative extremes and inflection points. We ultimately find that the graph of F is the one given in blue in Figure 5.1.3.

If we want an antiderivative G for which $G(0) = 3$, then G will have the exact same shape as F (since both share the derivative f), but G will be shifted vertically from the graph of F, as pictured in red in Figure 5.1.3. Note that $G(1) - G(0) = \int_0^1 f(x)\ dx = 0.5$, just as $F(1) - F(0) = 0.5$, but since $G(0) = 3$, $G(1) = G(0) + 0.5 = 3.5$, whereas $F(1) = 1.5$. In the same way, if we assigned a different initial value to the antiderivative, say $H(0) = -1$, we would get still another antiderivative, as shown in magenta in Figure 5.1.13.

This example demonstrates an important fact that holds more generally:

Note

If G and H are both antiderivatives of a function f, then the function $G - H$ must be constant.

To see why this result holds, observe that if G and H are both antiderivatives of f, then $G' = f$ and $H' = f$. Hence,

$$\frac{d}{dx}[G(x) - H(x)] = G'(x) - H'(x) = f(x) - f(x) = 0.$$

Since the only way a function can have derivative zero is by being a constant function, it follows that the function $G - H$ must be constant.

We now see that if a function has at least one antiderivative, it must have infinitely many: we can add any constant of our choice to the antiderivative and get another antiderivative. For this reason, we sometimes refer to the *general antiderivative* of a function f.

To identify a particular antiderivative of f, we must know a single value of the antiderivative F (this value is often called an *initial condition*). For example, if $f(x) = x^2$, its general antiderivative is $F(x) = \frac{1}{3}x^3 + C$, where we include the "$+C$" to indicate that F includes *all* of the possible antiderivatives of f. If we know that $F(2) = 3$, we substitute 2 for x in $F(x) = \frac{1}{3}x^3 + C$, and find that

$$3 = \frac{1}{3}(2)^3 + C,$$

or $C = 3 - \frac{8}{3} = \frac{1}{3}$. Therefore, the particular antiderivative in this case is $F(x) = \frac{1}{3}x^3 + \frac{1}{3}$.

Activity 5.1.3

For each of the following functions, sketch an accurate graph of the antiderivative that satisfies the given initial condition. In addition, sketch the graph of two additional antiderivatives of the given function, and state the corresponding initial conditions that each of them satisfy. If possible, find an algebraic formula for the antiderivative that satisfies the initial condition.

1. original function: $g(x) = |x| - 1$; initial condition: $G(-1) = 0$; interval for sketch: $[-2, 2]$
2. original function: $h(x) = \sin(x)$; initial condition: $H(0) = 1$; interval for sketch: $[0, 4\pi]$
3. original function: $p(x) = \begin{cases} x^2, & \text{if } 0 < x < 1 \\ -(x-2)^2, & \text{if } 1 < x < 2 \\ 0 & \text{otherwise} \end{cases}$; initial condition: $P(0) = 1$; interval for sketch: $[-1, 3]$

Functions defined by integrals

Equation (5.1.1) allows us to compute the value of the antiderivative F at a point b, provided that we know $F(a)$ and can evaluate the definite integral from a to b of f. That is,

$$F(b) = F(a) + \int_a^b f(x)\,dx.$$

In several situations, we have used this formula to compute $F(b)$ for several different values of b, and then plotted the points $(b, F(b))$ to help us draw an accurate graph of F. This suggests that we may want to think of b, the upper limit of integration, as a variable itself. To that end, we introduce the idea of an *integral function*, a function whose formula involves a definite integral.

Definition 5.1.1

If f is a continuous function, we define the corresponding ***integral function*** A according to the rule

$$A(x) = \int_a^x f(t)\,dt. \tag{5.1.12}$$

Note that because x is the independent variable in the function A, and determines the endpoint of the interval of integration, we need to use a different variable as the variable of integration. A standard choice is t, but any variable other than x is acceptable.

One way to think of the function A is as the "net signed area from a up to x" function, where we consider the region bounded by $y = f(t)$. For example, in Figure 5.1.4, we see a function f pictured at left, and its corresponding area function (choosing $a = 0$), $A(x) = \int_0^x f(t)\ dt$ shown at right.

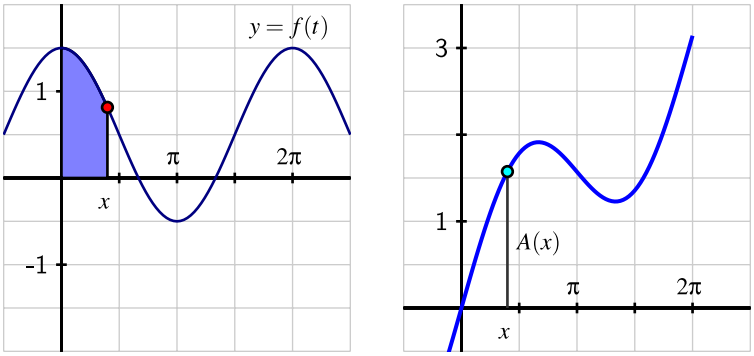


Figure 5.1.4. At left, the graph of the given function f. At right, the area function $A(x) = \int_0^x f(t)\ dt$.

The function A measures the net signed area from $t = 0$ to $t = x$ bounded by the curve $y = f(t)$; this value is then reported as the corresponding height on the graph of $y = A(x)$. At http://gvsu.edu/s/cz, we find a java applet[1] that brings the static picture in Figure 5.1.4 to life. There, the user can move the red point on the function f and see how the corresponding height changes at the light blue point on the graph of A.

David Austin, Grand Valley State University

The choice of a is somewhat arbitrary. In the activity that follows, we explore how the value of a affects the graph of the integral function.

Activity 5.1.4

Suppose that g is given by the graph at left in Figure 5.1.5 and that A is the corresponding integral function defined by $A(x) = \int_1^x g(t)\ dt$.

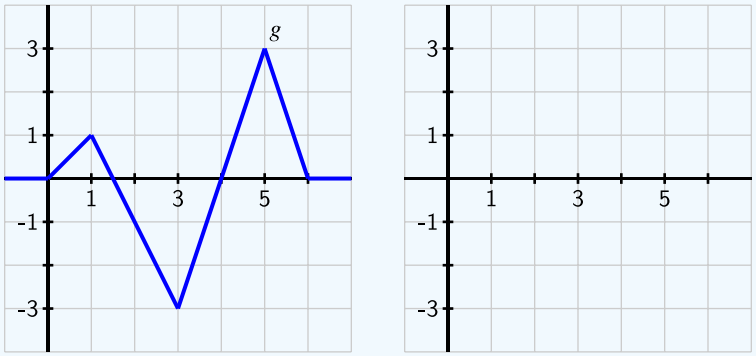


Figure 5.1.5. At left, the graph of $y = g(t)$; at right, axes for plotting $y = A(x)$, where A is defined by the formula $A(x) = \int_1^x g(t)\ dt$.

1. On what interval(s) is A an increasing function? On what intervals is A decreasing? Why?
2. On what interval(s) do you think A is concave up? concave down? Why?
3. At what point(s) does A have a relative minimum? a relative maximum?
4. Use the given information to determine the exact values of $A(0)$, $A(1)$, $A(2)$, $A(3)$, $A(4)$, $A(5)$, and $A(6)$.
5. Based on your responses to all of the preceding questions, sketch a complete and accurate graph of $y = A(x)$ on the axes provided, being sure to indicate the behavior of A for $x < 0$ and $x > 6$.
6. How does the graph of B compare to A if B is instead defined by $B(x) = \int_0^x g(t)\ dt$?

Summary

- Given the graph of a function f, we can construct the graph of its antiderivative F provided that (a) we know a starting value of F, say $F(a)$, and (b) we can evaluate the integral $\int_a^b f(x)\ dx$ exactly for relevant choices of a and b. For instance, if we wish to know $F(3)$, we can compute $F(3) = F(a) + \int_a^3 f(x)\ dx$. When we combine this information about the function values of F together with our understanding of how the behavior of $F' = f$ affects the overall shape of F, we can develop a completely accurate graph of the antiderivative F.
- Because the derivative of a constant is zero, if F is an antiderivative of f, it follows that $G(x) = F(x) + C$ will also be an antiderivative of f. Moreover, any two antiderivatives of a function f differ precisely by a constant. Thus, any function with at least one antiderivative in fact has infinitely many, and the graphs of any two antiderivatives will differ only by a vertical translation.
- Given a function f, the rule $A(x) = \int_a^x f(t)\ dt$ defines a new function A that measures the net-signed area bounded by f on the interval $[a, x]$. We call the function A the integral function corresponding to f.

This page titled 5.1: Construction Accurate Graphs of Antiderivatives is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.1: Construction Accurate Graphs of Antiderivatives by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

5.2: The Second Fundamental Theorem of Calculus

Motivating Questions

- How does the integral function $A(x) = \int_1^x f(t)\ dt$ define an antiderivative of f?
- What is the statement of the Second Fundamental Theorem of Calculus?
- How do the First and Second Fundamental Theorems of Calculus enable us to formally see how differentiation and integration are almost inverse processes?

In Section 4.4, we learned the Fundamental Theorem of Calculus (FTC), which from here forward will be referred to as the *First* Fundamental Theorem of Calculus, as in this section we develop a corresponding result that follows it. Recall that the First FTC tells us that if f is a continuous function on $[a, b]$ and F is any antiderivative of f (that is, $F' = f$), then

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

We have used this result in two settings:

1. If we have a graph of f and we can compute the exact area bounded by f on an interval $[a, b]$, we can compute the change in an antiderivative F over the interval.
2. If we can find an algebraic formula for an antiderivative of f, we can evaluate the integral to find the net signed area bounded by the function on the interval.

For the former, see Preview Activity 5.1.1 or Activity 5.1.2. For the latter, we can easily evaluate exactly integrals such as

$$\int_1^4 x^2\,dx,$$

since we know that the function $F(x) = \frac{1}{3}x^3$ is an antiderivative of $f(x) = x^2$. Thus,

$$\begin{aligned}\int_1^4 x^2\,dx &= \frac{1}{3}x^3\Big|_1^4 \\ &= \frac{1}{3}(4)^3 - \frac{1}{3}(1)^3 \\ &= 21.\end{aligned}$$

Thus, the First FTC can used in two ways. First, to find the difference $F(b) - F(a)$ for an antiderivative F of the integrand f, even if we may not have a formula for F itself. To do this, we must know the value of the integral $\int_a^b f(x)\ dx$ exactly, perhaps through known geometric formulas for area. In addition, the First FTC provides a way to find the exact value of a definite integral, and hence a certain net signed area exactly, by finding an antiderivative of the integrand and evaluating its total change over the interval. In this case, we need to know a formula for the antiderivative F. Both of these perspectives are reflected in Figure 5.2.1.

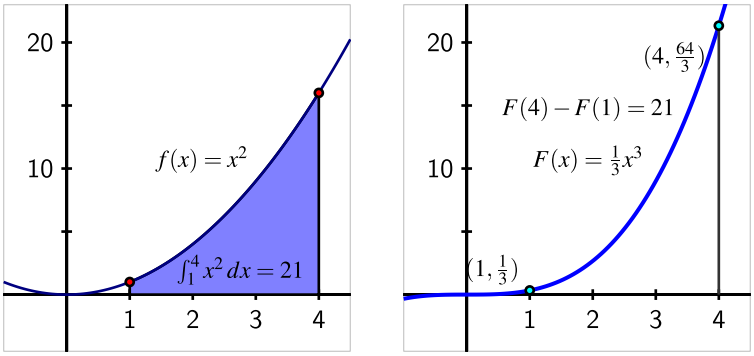


Figure 5.2.1. At left, the graph of $f(x) = x^2$ on the interval $[1, 4]$ and the area it bounds. At right, the antiderivative function $F(x) = \frac{1}{3}x^3$, whose total change on $[1, 4]$ is the value of the definite integral at left.

The value of a definite integral may have additional meaning depending on context: as the change in position when the integrand is a velocity function, the total amount of pollutant leaked from a tank when the integrand is the rate at which pollution is leaking, or other total changes if the integrand is a rate function. Also, the value of the definite integral is connected to the average value of a continuous function on a given interval: $f_{\text{AVG}[a,b]} = \frac{1}{b-a}\int_a^b f(x)\ dx$.

In the last part of Section 5.1, we studied integral functions of the form $A(x) = \int_c^x f(t)\ dt$. Figure 5.1.5 is a particularly important image to keep in mind as we work with integral functions, and the corresponding java applet at gvsu.edu/s/cz can help us understand the function A. In what follows, we use the First FTC to gain additional understanding of the function $A(x) = \int_c^x f(t)\ dt$, where the integrand f is given (either through a graph or a formula), and c is a constant.

Preview Activity 5.2.1

Consider the function A defined by the rule

$$A(x) = \int_1^x f(t)\,dt,$$

where $f(t) = 4 - 2t$.

1. Compute $A(1)$ and $A(2)$ exactly.
2. Use the First Fundamental Theorem of Calculus to find a formula for $A(x)$ that does not involve integrals. That is, use the first FTC to evaluate $\int_1^x (4 - 2t)\ dt$.
3. Observe that f is a linear function; what kind of function is A?
4. Using the formula you found in (b) that does not involve integrals, compute $A'(x)$.
5. While we have defined f by the rule $f(t) = 4 - 2t$, it is equivalent to say that f is given by the rule $f(x) = 4 - 2x$. What do you observe about the relationship between A and f?

The Second Fundamental Theorem of Calculus

The result of Preview Activity 5.2.1 is not particular to the function $f(t) = 4 - 2t$, nor to the choice of "1" as the lower bound in the integral that defines the function A. For instance, if we let $f(t) = \cos(t) - t$ and set $A(x) = \int_2^x f(t)\ dt$, we can determine a formula for A by the First FTC. Specifically,

$$\begin{aligned} A(x) &= \int_2^x (\cos(t) - t)\,dt \\ &= \sin(t) - \frac{1}{2}t^2\Big|_2^x \\ &= \sin(x) - \frac{1}{2}x^2 - (\sin(2) - 2)\,. \end{aligned}$$

Differentiating $A(x)$, since $(\sin(2) - 2)$ is constant, it follows that

$$A'(x) = \cos(x) - x,$$

and thus we see that $A'(x) = f(x)$, so A is an antiderivative of f. And since $A(2) = \int_2^2 f(t)\ dt = 0$, A is the only antiderivative of f for which $A(2) = 0$.

In general, if f is any continuous function, and we define the function A by the rule

$$A(x) = \int_c^x f(t)\,dt,$$

where c is an arbitrary constant, then we can show that A is an antiderivative of f. To see why, let's demonstrate that $A'(x) = f(x)$ by using the limit definition of the derivative. Doing so, we observe that

$$\begin{aligned} A'(x) &= \lim_{h\to 0} \frac{A(x+h) - A(x)}{h} \\ &= \lim_{h\to 0} \frac{\int_c^{x+h} f(t)\,dt - \int_c^x f(t)\,dt}{h} \\ &= \lim_{h\to 0} \frac{\int_x^{x+h} f(t)\,dt}{h}, \end{aligned} \tag{5.2.1}$$

where Equation (5.2.1) follows from the fact that $\int_c^x f(t)\ dt + \int_x^{x+h} f(t)\ dt = \int_c^{x+h} f(t)\ dt$. Now, observe that for small values of h,

$$\int_x^{x+h} f(t)\,dt \approx f(x) \cdot h,$$

by a simple left-hand approximation of the integral. Thus, as we take the limit in Equation (5.2.1), it follows that

$$A'(x) = \lim_{h\to 0} \frac{\int_x^{x+h} f(t)\,dt}{h} = \lim_{h\to 0} \frac{f(x)\cdot h}{h} = f(x).$$

Hence, A is indeed an antiderivative of f. In addition, $A(c) = \int_c^c f(t)\ dt = 0$. The preceding argument demonstrates the truth of the Second Fundamental Theorem of Calculus, which we state as follows.

The Second Fundamental Theorem of Calculus

If f is a continuous function and c is any constant, then f has a unique antiderivative A that satisfies $A(c) = 0$, and that antiderivative is given by the rule $A(x) = \int_c^x f(t)\ dt$.

Activity 5.2.2

Suppose that f is the function given in Figure 5.2.2 and that f is a piecewise function whose parts are either portions of lines or portions of circles, as pictured.

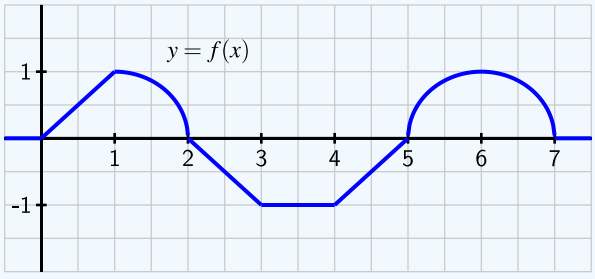




Figure 5.2.2. At left, the graph of $y = f(x)$. At right, axes for sketching $y = A(x)$.

In addition, let A be the function defined by the rule $A(x) = \int_2^x f(t)\ dt$.

1. What does the Second FTC tell us about the relationship between A and f?
2. Compute $A(1)$ and $A(3)$ exactly.
3. Sketch a precise graph of $y = A(x)$ on the axes at right that accurately reflects where A is increasing and decreasing, where A is concave up and concave down, and the exact values of A at $x = 0, 1, \ldots, 7$.
4. How is A similar to, but different from, the function F that you found in Activity 5.1.2?
5. With as little additional work as possible, sketch precise graphs of the functions $B(x) = \int_3^x f(t)\ dt$ and $C(x) = \int_1^x f(t)\ dt$. Justify your results with at least one sentence of explanation.

Understanding Integral Functions

The Second FTC provides us with a way to construct an antiderivative of any continuous function. In particular, if we are given a continuous function g and wish to find an antiderivative G, we can now say that

$$G(x) = \int_c^x g(t)\,dt$$

provides the rule for such an antiderivative, and moreover that $G(c) = 0$. Note especially that we know that $G'(x) = g(x)$, or

$$\frac{d}{dx}\left[\int_c^x g(t)\,dt\right] = g(x). \tag{5.2.2}$$

This result is useful for understanding the graph of G.

Example 5.2.1

Investigate the behavior of the integral function

$$E(x) = \int_0^x e^{-t^2}\,dt.$$

Answer available at the end of this section.

Activity 5.2.3

Suppose that $f(t) = \frac{t}{1+t^2}$ and $F(x) = \int_0^x f(t)\ dt$.

1. On the axes at left in Figure 5.2.4, plot a graph of $f(t) = \frac{t}{1+t^2}$ on the interval $-10 \le t \le 10$. Clearly label the vertical axes with appropriate scale.
2. What is the key relationship between F and f, according to the Second FTC?
3. Use the first derivative test to determine the intervals on which F is increasing and decreasing.
4. Use the second derivative test to determine the intervals on which F is concave up and concave down. Note that $f'(t)$ can be simplified to be written in the form $f'(t) = \frac{1-t^2}{(1+t^2)^2}$.
5. Using technology appropriately, estimate the values of $F(5)$ and $F(10)$ through appropriate Riemann sums.
6. Sketch an accurate graph of $y = F(x)$ on the righthand axes provided, and clearly label the vertical axes with appropriate scale.



Figure 5.2.4. Axes for plotting f and F.

Differentiating an Integral Function

We have seen that the Second FTC enables us to construct an antiderivative F for any continuous function f as the integral function $F(x) = \int_c^x f(t)\ dt$. If we have a function of the form $F(x) = \int_c^x f(t)\ dt$, then we know that $F'(x) = \frac{d}{dx}\left[\int_c^x f(t)\,dt\right] = f(x)$. This shows that integral functions, while perhaps having the most complicated formulas of any functions we have encountered, are nonetheless particularly simple to differentiate. For instance, if

$$F(x) = \int_\pi^x \sin(t^2)\,dt,$$

then by the Second FTC, we know immediately that

$$F'(x) = \sin(x^2).$$

In general, we know by the Second FTC that

$$\frac{d}{dx}\left[\int_a^x f(t)\,dt\right] = f(x).$$

This equation says that "the derivative of the integral function whose integrand is f, is f." We see that if we first integrate the function f from $t = a$ to $t = x$, and then differentiate with respect to x, these two processes "undo" each other.

What happens if we differentiate a function $f(t)$ and then integrate the result from $t = a$ to $t = x$? That is, what can we say about the quantity

$$\int_a^x \frac{d}{dt}\left[f(t)\right]\,dt?$$

We note that $f(t)$ is an antiderivative of $\frac{d}{dt}\left[f(t)\right]$ and apply the First FTC. We see that

$$\begin{aligned}\int_a^x \frac{d}{dt}\left[f(t)\right]\,dt &= f(t)\Big|_a^x \\ &= f(x) - f(a).\end{aligned}$$

Thus, we see that if we first differentiate f and then integrate the result from a to x, we return to the function f, minus the constant value $f(a)$. So the two processes almost undo each other, up to the constant $f(a)$.

The observations made in the preceding two paragraphs demonstrate that differentiating and integrating (where we integrate from a constant up to a variable) are almost inverse processes. This should not be surprising: integrating involves antidifferentiating, which reverses the process of differentiating. On the other hand, we see that there is some subtlety involved, because integrating the derivative of a function does not quite produce the function itself. This is because every function has an entire family of antiderivatives, and any two of those antiderivatives differ only by a constant.

Activity 5.2.4

Evaluate each of the following derivatives and definite integrals. Clearly cite whether you use the First or Second FTC in so doing.

1. $\frac{d}{dx}\left[\int_4^x e^{t^2}\,dt\right]$
2. $\int_{-2}^{x} \frac{d}{dt}\left[\frac{t^4}{1+t^4}\right]\,dt$
3. $\frac{d}{dx}\left[\int_x^1 \cos(t^3)\,dt\right]$
4. $\int_3^x \frac{d}{dt}\left[\ln(1+t^2)\right]\,dt$
5. $\frac{d}{dx}\left[\int_4^{x^3} \sin(t^2)\,dt\right]$.

Summary

- For a continuous function f, the integral function $A(x) = \int_1^x f(t)\,dt$ defines an antiderivative of f.
- The Second Fundamental Theorem of Calculus is the formal, more general statement of the preceding fact: if f is a continuous function and c is any constant, then $A(x) = \int_c^x f(t)\,dt$ is the unique antiderivative of f that satisfies $A(c) = 0$.
- Together, the First and Second FTC enable us to formally see how differentiation and integration are almost inverse processes through the observations that

$$\int_c^x \frac{d}{dt}\left[f(t)\right]\,dt = f(x) - f(c)$$

and

$$\frac{d}{dx}\left[\int_c^x f(t)\,dt\right] = f(x).$$

Answers

Example 5.2.1

E is closely related to the well known *error function* [1] in probability and statistics. It turns out that the function e^{-t^2} does not have an elementary antiderivative. The error function is defined by the rule $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt$ and has the key property that $0 \le \operatorname{erf}(x) < 1$ for all $x \ge 0$ and moreover that $\lim_{x\to\infty} \operatorname{erf}(x) = 1$.

While we cannot evaluate E exactly for any value other than $x = 0$, we still can gain a tremendous amount of information about the function E. By applying the rule in Equation (5.2.2) to E, it follows that

$$E'(x) = \frac{d}{dx}\left[\int_0^x e^{-t^2}\,dt\right] = e^{-x^2},$$

so we know a formula for the derivative of E, and we know that $E(0) = 0$. This information is precisely the type we were given in Activity 3.1.2, where we were given information about the derivative of a function, but lacked a formula for the function itself.

Using the first and second derivatives of E, along with the fact that $E(0) = 0$, we can determine more information about the behavior of E. First, we note that for all real numbers x, $e^{-x^2} > 0$, and thus $E'(x) > 0$ for all x. Thus E is an always increasing function. Further, as $x \to \infty$, $E'(x) = e^{-x^2} \to 0$, so the slope of the function E tends to zero as $x \to \infty$ (and similarly as $x \to -\infty$). Indeed, it turns out that E has horizontal asymptotes as x increases or decreases without bound.

In addition, we can observe that $E''(x) = -2xe^{-x^2}$, and that $E''(0) = 0$, while $E''(x) < 0$ for $x > 0$ and $E''(x) > 0$ for $x < 0$. This information tells us that E is concave up for $x < 0$ and concave down for $x > 0$ with a point of inflection at $x = 0$.

The only thing we lack at this point is a sense of how big E can get as x increases. If we use a midpoint Riemann sum with 10 subintervals to estimate $E(2)$, we see that $E(2) \approx 0.8822$; a similar calculation to estimate $E(3)$ shows little change ($E(3) \approx 0.8862$), so it appears that as x increases without bound, E approaches a value just larger than 0.886, which aligns with the fact that E has horizontal asymptotes. Putting all of this information together (and using the symmetry of $f(t) = e^{-t^2}$), we see the results shown in Figure 5.2.3.

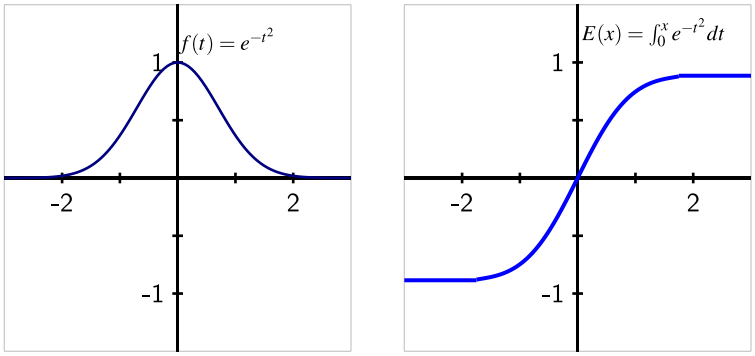


Figure 5.2.3. At left, the graph of $f(t) = e^{-t^2}$. At right, the integral function $E(x) = \int_0^x e^{-t^2}\, dt$, which is the unique antiderivative of f that satisfies $E(0) = 0$.

Because E is the antiderivative of $f(t) = e^{-t^2}$ that satisfies $E(0) = 0$, values on the graph of $y = E(x)$ represent the net signed area of the region bounded by $f(t) = e^{-t^2}$ from 0 up to x. We see that the value of E increases rapidly near zero but then levels off as x increases, since there is less and less additional accumulated area bounded by $f(t) = e^{-t^2}$ as x increases.

This page titled 5.2: The Second Fundamental Theorem of Calculus is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.2: The Second Fundamental Theorem of Calculus by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

5.3: Integration by Substitution

Motivating Questions

- How can we begin to find algebraic formulas for antiderivatives of more complicated algebraic functions?
- What is an indefinite integral and how is its notation used in discussing antiderivatives?
- How does the technique of u-substitution work to help us evaluate certain indefinite integrals, and how does this process rely on identifying function-derivative pairs?

In Section 4.4, we learned the key role that antiderivatives play in the process of evaluating definite integrals exactly. The Fundamental Theorem of Calculus tells us that if F is any antiderivative of f, then

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

Furthermore, we realized that each elementary derivative rule developed in Chapter 2 leads to a corresponding elementary antiderivative, as summarized in Table 4.4.5. Thus, if we wish to evaluate an integral such as

$$\int_0^1 \left(x^3 - \sqrt{x} + 5^x\right)\,dx,$$

it is straightforward to do so, since we can easily antidifferentiate $f(x) = x^3 - \sqrt{x} + 5^x$. Because one antiderivative of f is $F(x) = \frac{1}{4}x^4 - \frac{2}{3}x^{3/2} + \frac{1}{\ln(5)}5^x$, the Fundamental Theorem of Calculus tells us that

$$\begin{aligned}\int_0^1 \left(x^3 - \sqrt{x} + 5^x\right)\,dx &= \frac{1}{4}x^4 - \frac{2}{3}x^{3/2} + \frac{1}{\ln(5)}5^x\Big|_0^1 \\ &= \left(\frac{1}{4}(1)^4 - \frac{2}{3}(1)^{3/2} + \frac{1}{\ln(5)}5^1\right) - \left(0 - 0 + \frac{1}{\ln(5)}5^0\right) \\ &= -\frac{5}{12} + \frac{4}{\ln(5)}.\end{aligned}$$

We see that we have a natural interest in being able to find such algebraic antiderivatives. We emphasize *algebraic* antiderivatives, as opposed to any antiderivative, since we know by the Second Fundamental Theorem of Calculus that $G(x) = \int_a^x f(t)\,dt$ is indeed an antiderivative of the given function f, but one that still involves a definite integral. Our goal in this section is to "undo" the process of differentiation to find an algebraic antiderivative for a given function.

Preview Activity 5.3.1

In Section 2.5, we learned the Chain Rule and how it can be applied to find the derivative of a composite function. In particular, if u is a differentiable function of x, and f is a differentiable function of $u(x)$, then

$$\frac{d}{dx}\left[f(u(x))\right] = f'(u(x)) \cdot u'(x).$$

In words, we say that the derivative of a composite function $c(x) = f(u(x))$, where f is considered the "outer" function and u the "inner" function, is "the derivative of the outer function, evaluated at the inner function, times the derivative of the inner function."

1. For each of the following functions, use the Chain Rule to find the function's derivative. Be sure to label each derivative by name (e.g., the derivative of $g(x)$ should be labeled $g'(x)$).
   1. $g(x) = e^{3x}$
   2. $h(x) = \sin(5x + 1)$
   3. $p(x) = \arctan(2x)$

4. $q(x) = (2-7x)^4$
5. $r(x) = 3^{4-11x}$

2. For each of the following functions, use your work in (a) to help you determine the general antiderivative [1] of the function. Label each antiderivative by name (e.g., the antiderivative of m should be called M). In addition, check your work by computing the derivative of each proposed antiderivative.
   1. $m(x) = e^{3x}$
   2. $n(x) = \cos(5x+1)$
   3. $s(x) = \dfrac{1}{1+4x^2}$
   4. $v(x) = (2-7x)^3$
   5. $w(x) = 3^{4-11x}$
3. Based on your experience in parts (a) and (b), conjecture an antiderivative for each of the following functions. Test your conjectures by computing the derivative of each proposed antiderivative.
   1. $a(x) = \cos(\pi x)$
   2. $b(x) = (4x+7)^{11}$
   3. $c(x) = xe^{x^2}$

Recall that the general antiderivative of a function includes "$+C$" to reflect the entire family of functions that share the same derivative.

Reversing the Chain Rule: First Steps

Whenever f is a familiar function whose antiderivative is known and $u(x)$ is a linear function, it is straightforward to antidifferentiate a function of the form

$$h(x) = f(u(x)).$$

Example 5.3.1

Determine the general antiderivative of

$$h(x) = (5x-3)^6.$$

Check the result by differentiating.

For this composite function, the outer function f is $f(u) = u^6$, while the inner function is $u(x) = 5x-3$. Since the antiderivative of f is $F(u) = \frac{1}{7}u^7 + C$, we see that the antiderivative of h is

$$H(x) = \frac{1}{7}(5x-3)^7 \cdot \frac{1}{5} + C = \frac{1}{35}(5x-3)^7 + C.$$

The inclusion of the constant $\frac{1}{5}$ is essential precisely because the derivative of the inner function is $u'(x) = 5$. Indeed, if we now compute $H'(x)$, we find by the Chain Rule (and Constant Multiple Rule) that

$$H'(x) = \frac{1}{35} \cdot 7(5x-3)^6 \cdot 5 = (5x-3)^6 = h(x),$$

and thus H is indeed the general antiderivative of h.

Hence, in the special case where the outer function is familiar and the inner function is linear, we can antidifferentiate composite functions according to the following rule.

Note

If $h(x) = f(ax+b)$ and F is a known algebraic antiderivative of f, then the general antiderivative of h is given by

$$H(x) = \frac{1}{a}F(ax+b) + C.$$

It is useful to have shorthand notation that indicates the instruction to find an antiderivative. Thus, in a similar way to how the notation

$$\frac{d}{dx}\left[f(x)\right]$$

represents the derivative of $f(x)$ with respect to x, we use the notation of the *indefinite integral*,

$$\int f(x)\,dx$$

to represent the general antiderivative of f with respect to x. Returning to the earlier example with $h(x) = (5x-3)^6$, we can rephrase the relationship between h and its antiderivative H through the notation

$$\int (5x-3)^6\,dx = \frac{1}{35}(5x-6)^7 + C.$$

When we find an antiderivative, we will often say that we *evaluate an indefinite integral*. Just as the notation $\frac{d}{dx}[\square]$ means "find the derivative with respect to x of $\square$," the notation $\int \square\ dx$ means "find a function of x whose derivative is $\square$."

Activity 5.3.2

Evaluate each of the following indefinite integrals. Check each antiderivative that you find by differentiating.

1. $\int \sin(8-3x)\ dx$
2. $\int \sec^2(4x)\ dx$
3. $\int \frac{1}{11x-9}\ dx$
4. $\int \csc(2x+1)\cot(2x+1)\ dx$
5. $\int \frac{1}{\sqrt{1-16x^2}}\ dx$
6. $\int 5^{-x}\ dx$

Reversing the Chain Rule: u-substitution

A natural question arises from our recent work: what happens when the inner function is not linear? For example, can we find antiderivatives of such functions as

$$g(x) = xe^{x^2} \text{ and } h(x) = e^{x^2}?$$

It is important to remember that differentiation and antidifferentiation are almost inverse processes (that they are not is due to the $+C$ that arises when antidifferentiating). This almost-inverse relationship enables us to take any known derivative rule and rewrite it as a corresponding rule for an indefinite integral. For example, since

$$\frac{d}{dx}\left[x^5\right] = 5x^4,$$

we can equivalently write

$$\int 5x^4\,dx = x^5 + C.$$

Recall that the Chain Rule states that

$$\frac{d}{dx}\left[f(g(x))\right] = f'(g(x)) \cdot g'(x).$$

Restating this relationship in terms of an indefinite integral,

$$\int f'(g(x))g'(x)\,dx = f(g(x)) + C. \tag{5.3.1}$$

Equation (5.3.1) tells us that if we can view a given function as $f'(g(x))g'(x)$ for some appropriate choices of f and g, then we can antidifferentiate the function by reversing the Chain Rule. Note that both $g(x)$ and $g'(x)$ appear in the form of $f'(g(x))g'(x)$; we will sometimes say that we seek to *identify a function-derivative pair* ($g(x)$ and $g'(x)$) when trying to apply the rule in Equation (5.3.1).

If we can identify a function-derivative pair, we will introduce a new variable u to represent the function $g(x)$. With $u = g(x)$, it follows in Leibniz notation that $\frac{du}{dx} = g'(x)$, so that in terms of differentials[2], $du = g'(x)\ dx$. Now converting the indefinite integral to a new one in terms of u, we have

$$\int f'(g(x))g'(x)\,dx = \int f'(u)\,du.$$

If we recall from the definition of the derivative that $\frac{du}{dx} \approx \frac{\Delta u}{\Delta x}$ and use the fact that $\frac{du}{dx} = g'(x)$, then we see that $g'(x) \approx \frac{\Delta u}{\Delta x}$. Solving for Δu, $\Delta u \approx g'(x)\Delta x$. It is this last relationship that, when expressed in "differential" notation enables us to write $du = g'(x)\ dx$ in the change of variable formula.

Provided that f' is an elementary function whose antiderivative is known, we can easily evaluate the indefinite integral in u, and then go on to determine the desired overall antiderivative of $f'(g(x))g'(x)$. We call this process *u-substitution*, and summarize the rule as follows:

Note

With the substitution $u = g(x)$,

$$\int f'(g(x))g'(x)\,dx = \int f'(u)\,du = f(u) + C = f(g(x)) + C.$$

To see u-substitution at work, we consider the following example.

Example 5.3.2

Evaluate the indefinite integral

$$\int x^3 \cdot \sin(7x^4 + 3)\,dx$$

and check the result by differentiating.

Answer available at the end of this section.

The u-substitution worked because the function multiplying $\sin(7x^4 + 3)$ was x^3. If instead that function was x^2 or x^4, the substitution process would not have worked. This is one of the primary challenges of antidifferentiation: slight changes in the integrand make tremendous differences. For instance, we can use u-substitution with $u = x^2$ and $du = 2xdx$ to find that

$$\begin{aligned}\int xe^{x^2}\,dx &= \int e^u \cdot \frac{1}{2}\,du \\ &= \frac{1}{2}\int e^u\,du \\ &= \frac{1}{2}e^u + C \\ &= \frac{1}{2}e^{x^2} + C.\end{aligned}$$

However, for the similar indefinite integral

$$\int e^{x^2}\,dx,$$

the u-substitution $u = x^2$ is no longer possible because the factor of x is missing. Hence, part of the lesson of u-substitution is just how specialized the process is: it only applies to situations where, up to a missing constant, the integrand is the result of applying the Chain Rule to a different, related function.

Activity 5.3.3

Evaluate each of the following indefinite integrals by using these steps:

- Find two functions within the integrand that form (up to a possible missing constant) a function-derivative pair;
- Make a substitution and convert the integral to one involving u and du;
- Evaluate the new integral in u;
- Convert the resulting function of u back to a function of x by using your earlier substitution;
- Check your work by differentiating the function of x. You should come up with the integrand originally given.

1. $\displaystyle\int \frac{x^2}{5x^3+1}\,dx$
2. $\displaystyle\int e^x \sin(e^x)\,dx$
3. $\displaystyle\int \frac{\cos(\sqrt{x})}{\sqrt{x}}\,dx$

Evaluating Definite Integrals via u-substitution

We have introduced u-substitution as a means to evaluate indefinite integrals of functions that can be written, up to a constant multiple, in the form $f(g(x))g'(x)$. This same technique can be used to evaluate definite integrals involving such functions, though we need to be careful with the corresponding limits of integration. Consider, for instance, the definite integral

$$\int_2^5 xe^{x^2}\,dx.$$

Whenever we write a definite integral, it is implicit that the limits of integration correspond to the variable of integration. To be more explicit, observe that

$$\int_2^5 xe^{x^2}\,dx = \int_{x=2}^{x=5} xe^{x^2}\,dx.$$

When we execute a u-substitution, we change the *variable* of integration; it is essential to note that this also changes the *limits* of integration. For instance, with the substitution $u = x^2$ and $du = 2x\ dx$, it also follows that when $x = 2$, $u = 2^2 = 4$, and when $x = 5$, $u = 5^2 = 25$. Thus, under the change of variables of u-substitution, we now have

$$\begin{aligned}\int_{x=2}^{x=5} xe^{x^2}\,dx &= \int_{u=4}^{u=25} e^u \cdot \frac{1}{2}\,du \\ &= \frac{1}{2}e^u\Big|_{u=4}^{u=25} \\ &= \frac{1}{2}e^{25} - \frac{1}{2}e^4.\end{aligned}$$

Alternatively, we could consider the related indefinite integral $\int xe^{x^2}\,dx$, find the antiderivative $\frac{1}{2}e^{x^2}$ through u-substitution, and then evaluate the original definite integral. With that method, we'd have

$$\begin{aligned}\int_2^5 xe^{x^2}\,dx &= \frac{1}{2}e^{x^2}\Big|_2^5 \\ &= \frac{1}{2}e^{25} - \frac{1}{2}e^4,\end{aligned}$$

which is, of course, the same result.

Activity 5.3.4

Evaluate each of the following definite integrals exactly through an appropriate u-substitution.

1. $\displaystyle\int_1^2 \frac{x}{1+4x^2}\,dx$
2. $\displaystyle\int_0^1 e^{-x}(2e^{-x}+3)^9\,dx$
3. $\displaystyle\int_{2/\pi}^{4/\pi} \frac{\cos\left(\frac{1}{x}\right)}{x^2}\,dx$

Summary

- To find algebraic formulas for antiderivatives of more complicated algebraic functions, we need to think carefully about how we can reverse known differentiation rules. To that end, it is essential that we understand and recall known derivatives of basic functions, as well as the standard derivative rules.
- The indefinite integral provides notation for antiderivatives. When we write "$\int f(x)\,dx$," we mean "the general antiderivative of f." In particular, if we have functions f and F such that $F' = f$, the following two statements say the exact thing:

  $$\frac{d}{dx}[F(x)] = f(x) \text{ and } \int f(x)\,dx = F(x) + C.$$

  That is, f is the derivative of F, and F is an antiderivative of f.

- The technique of u-substitution helps us to evaluate indefinite integrals of the form $\int f(g(x))g'(x)\,dx$ through the substitutions $u = g(x)$ and $du = g'(x)\,dx$, so that

  $$\int f(g(x))g'(x)\,dx = \int f(u)\,du.$$

  A key part of choosing the expression in x to be represented by u is the identification of a function-derivative pair. To do so, we often look for an "inner" function $g(x)$ that is part of a composite function, while investigating whether $g'(x)$ (or a constant multiple of $g'(x)$) is present as a multiplying factor of the integrand.

Answers

Example 5.3.2

We can make two algebraic observations regarding the integrand, $x^3 \cdot \sin(7x^4+3)$. First, $\sin(7x^4+3)$ is a composite function; as such, we know we'll need a more sophisticated approach to antidifferentiating. Second, x^3 is almost the derivative of $(7x^4+3)$; the only issue is a missing constant. Thus, x^3 and $(7x^4+3)$ are nearly a function-derivative pair. Furthermore, we know the antiderivative of $f(u) = \sin(u)$. The combination of these observations suggests that we can evaluate the given indefinite integral by reversing the chain rule through u-substitution.

Letting u represent the inner function of the composite function $\sin(7x^4+3)$, we have $u = 7x^4+3$, and thus $\frac{du}{dx} = 28x^3$. In differential notation, it follows that $du = 28x^3\ dx$, and thus $x^3\ dx = \frac{1}{28}\ du$. The original indefinite integral may be slightly rewritten as

$$\int \sin(7x^4+3)\cdot x^3\,dx,$$

and so by substituting u for $7x^4+3$ and $\frac{1}{28}\ du$ for $x^3\ dx$, it follows that

$$\int \sin(7x^4+3)\cdot x^3\,dx = \int \sin(u)\cdot\frac{1}{28}\,du.$$

Now we may evaluate the easier integral in u, and then replace u by the expression $7x^4+3$. Doing so, we find

$$\begin{aligned}\int \sin(7x^4+3)\cdot x^3\,dx &= \int \sin(u)\cdot\frac{1}{28}\,du\\ &= \frac{1}{28}\int \sin(u)\,du\\ &= \frac{1}{28}(-\cos(u)) + C\\ &= -\frac{1}{28}\cos(7x^4+3) + C.\end{aligned}$$

To check our work, we observe by the Chain Rule that

$$\frac{d}{dx}\left[-\frac{1}{28}\cos(7x^4+3)\right] = -\frac{1}{28}\cdot(-1)\sin(7x^4+3)\cdot 28x^3 = \sin(7x^4+3)\cdot x^3,$$

which is indeed the original integrand.

This page titled 5.3: Integration by Substitution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.3: Integration by Substitution by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

5.4: Integration by Parts

Motivating Questions

- How do we evaluate indefinite integrals that involve products of basic functions such as $\int x\sin(x)\ dx$ and $\int xe^x\ dx$?
- What is the method of integration by parts and how can we consistently apply it to integrate products of basic functions?
- How does the algebraic structure of functions guide us in identifying u and dv in using integration by parts?

In Section 5.3, we learned the technique of u-substitution for evaluating indefinite integrals. For example, the indefinite integral $\int x^3\sin(x^4)\ dx$ is perfectly suited to u-substitution, because one factor is a composite function and the other factor is the derivative (up to a constant) of the inner function. Recognizing the algebraic structure of a function can help us to find its antiderivative.

Next we consider integrands with a different elementary algebraic structure: a product of basic functions. For instance, suppose we are interested in evaluating the indefinite integral

$$\int x\sin(x)\,dx.$$

The integrand is the product of the basic functions $f(x)=x$ and $g(x)=\sin(x)$. We know that it is relatively complicated to compute the derivative of the product of two functions, so we should expect that antidifferentiating a product should be similarly involved. Intuitively, we expect that evaluating $\int x\sin(x)\ dx$ will involve somehow reversing the Product Rule.

To that end, in Preview Activity 5.4.1 we refresh our understanding of the Product Rule and then investigate some indefinite integrals that involve products of basic functions.

Preview Activity 5.4.1

In Section 2.3, we developed the Product Rule and studied how it is employed to differentiate a product of two functions. In particular, recall that if f and g are differentiable functions of x, then

$$\frac{d}{dx}\left[f(x)\cdot g(x)\right]=f(x)\cdot g'(x)+g(x)\cdot f'(x).$$

1. For each of the following functions, use the Product Rule to find the function's derivative. Be sure to label each derivative by name (e.g., the derivative of $g(x)$ should be labeled $g'(x)$).
   1. $g(x)=x\sin(x)$
   2. $h(x)=xe^x$
   3. $p(x)=x\ln(x)$
   4. $q(x)=x^2\cos(x)$
   5. $r(x)=e^x\sin(x)$
2. Use your work in (a) to help you evaluate the following indefinite integrals. Use differentiation to check your work.
   1. $\int xe^x+e^x\ dx$
   2. $\int e^x(\sin(x)+\cos(x))\ dx$
   3. $\int 2x\cos(x)-x^2\sin(x)\ dx$
   4. $\int x\cos(x)+\sin(x)\ dx$
   5. $\int 1+\ln(x)\ dx$

3. Observe that the examples in (b) work nicely because of the derivatives you were asked to calculate in (a). Each integrand in (b) is precisely the result of differentiating one of the products of basic functions found in (a). To see what happens when an integrand is still a product but not necessarily the result of differentiating an elementary product, we consider how to evaluate

$$\int x\cos(x)\,dx.$$

   1. First, observe that

$$\frac{d}{dx}\left[x\sin(x)\right] = x\cos(x) + \sin(x).$$

      Integrating both sides indefinitely and using the fact that the integral of a sum is the sum of the integrals, we find that

$$\int \left(\frac{d}{dx}\left[x\sin(x)\right]\right)\,dx = \int x\cos(x)\,dx + \int \sin(x)\,dx.$$

      In this last equation, evaluate the indefinite integral on the left side as well as the rightmost indefinite integral on the right.

   2. In the most recent equation from (i.), solve the equation for the expression $\int x\cos(x)\ dx$.
   3. For which product of basic functions have you now found the antiderivative?

Reversing the Product Rule: Integration by Parts

Problem (c) in Preview Activity $5.4.1$ provides a clue to the general technique known as Integration by Parts, which comes from reversing the Product Rule. Recall that the Product Rule states that

$$\frac{d}{dx}\left[f(x)g(x)\right] = f(x)g'(x) + g(x)f'(x).$$

Integrating both sides of this equation indefinitely with respect to x, we find

$$\int \frac{d}{dx}\left[f(x)g(x)\right]\,dx = \int f(x)g'(x)\,dx + \int g(x)f'(x)\,dx. \tag{5.4.1}$$

On the left side of Equation $(5.4.1)$, we have the indefinite integral of the derivative of a function. Temporarily omitting the constant that may arise, we have

$$f(x)g(x) = \int f(x)g'(x)\,dx + \int g(x)f'(x)\,dx. \tag{5.4.2}$$

We solve for the first indefinite integral on the left to generate the rule

$$\int f(x)g'(x)\,dx = f(x)g(x) - \int g(x)f'(x)\,dx. \tag{5.4.3}$$

Often we express Equation $(5.4.3)$ in terms of the variables u and v, where $u = f(x)$ and $v = g(x)$. In differential notation, $du = f'(x)\ dx$ and $dv = g'(x)\ dx$, so we can state the rule for Integration by Parts in its most common form as follows:

Note

$$\int u\,dv = uv - \int v\,du.$$

To apply integration by parts, we look for a product of basic functions that we can identify as u and dv. If we can antidifferentiate dv to find v, and evaluating $\int v\ du$ is not more difficult than evaluating $\int u\ dv$, then this substitution usually proves to be fruitful. To demonstrate, we consider the following example.

Example 5.4.1

Evaluate the indefinite integral

$$\int x\cos(x)\,dx$$

using integration by parts.

Answer available at the end of this section.

The general technique of integration by parts involves trading the problem of integrating the product of two functions for the problem of integrating the product of two related functions. That is, we convert the problem of evaluating $\int u\ dv$ to that of evaluating $\int v\ du$. This clearly shapes our choice of u and v. In Example 5.4.1, the original integral to evaluate was $\int x\cos(x)\ dx$, and through the substitution provided by integration by parts, we were instead able to evaluate $\int \sin(x)\cdot 1\ dx$. Note that the original function x was replaced by its derivative, while $\cos(x)$ was replaced by its antiderivative.

Activity 5.4.2

Evaluate each of the following indefinite integrals. Check each antiderivative that you find by differentiating.

1. $\int te^{-t}\ dt$
2. $\int 4x\sin(3x)\ dx$
3. $\int z\sec^2(z)\ dz$
4. $\int x\ln(x)\ dx$

Some Subtleties with Integration by Parts

Sometimes integration by parts is not an obvious choice, but the technique is appropriate nonetheless. Integration by parts allows us to replace one function in a product with its derivative while replacing the other with its antiderivative. For instance, consider evaluating

$$\int \arctan(x)\,dx.$$

Initially, this problem seems ill-suited to integration by parts, since there does not appear to be a product of functions present. But if we note that $\arctan(x) = \arctan(x)\cdot 1$, and realize that we know the derivative of $\arctan(x)$ as well as the antiderivative of 1, we see the possibility for the substitution $u = \arctan(x)$ and $dv = 1\ dx$. We explore this substitution further in Activity 5.4.3.

In a related problem, consider $\int t^3\sin(t^2)\ dt$. Observe that there is a composite function present in $\sin(t^2)$, but there is not an obvious function-derivative pair, as we have t^3 (rather than simply t) multiplying $\sin(t^2)$. In this problem we use both u-substitution and integration by parts. First we write $t^3 = t\cdot t^2$ and consider the indefinite integral

$$\int t\cdot t^2\cdot\sin(t^2)\,dt.$$

We let $z = t^2$ so that $dz = 2t\ dt$, and thus $t\ dt = \frac{1}{2}\ dz$. (We are using the variable z to perform a "z-substitution" first so that we may then apply integration by parts.) Under this z-substitution, we now have

$$\int t\cdot t^2\cdot\sin(t^2)\,dt = \int z\cdot\sin(z)\cdot\frac{1}{2}\,dz.$$

The resulting integral can be evaluated by parts. This, too, is explored further in Activity 5.4.3.

These problems show that we sometimes must think creatively in choosing the variables for substitution in integration by parts, and that we may need to use substitution for an additional change of variables.

Activity 5.4.3

Evaluate each of the following indefinite integrals, using the provided hints.

1. Evaluate $\int \arctan(x)\ dx$ by using Integration by Parts with the substitution $u = \arctan(x)$ and $dv = 1\ dx$.
2. Evaluate $\int \ln(z)\ dz$. Consider a similar substitution to the one in (a).
3. Use the substitution $z = t^2$ to transform the integral $\int t^3 \sin(t^2)\ dt$ to a new integral in the variable z, and evaluate that new integral by parts.
4. Evaluate $\int s^5 e^{s^3}\ ds$ using an approach similar to that described in (c).
5. Evaluate $\int e^{2t} \cos(e^t)\ dt$. You will find it helpful to note that $e^{2t} = e^t \cdot e^t$.

Using Integration by Parts Multiple Times

Integration by parts is well suited to integrating the product of basic functions, allowing us to trade a given integrand for a new one where one function in the product is replaced by its derivative, and the other is replaced by its antiderivative. The goal in this trade of $\int u\ dv$ for $\int v\ du$ is that the new integral be simpler to evaluate than the original one. Sometimes it is necessary to apply integration by parts more than once in order to evaluate a given integral.

Example 5.4.2

Evaluate $\int t^2 e^t\ dt$.

Answer available at the end of this section.

Of course, even more than two applications of integration by parts may be necessary. In the preceding example, if the integrand had been $t^3 e^t$, we would have had to use integration by parts three times.

Next, we consider the slightly different scenario.

Example 5.4.3

Evaluate $\int e^t \cos(t)\ dt$.

Answer available at the end of this section.

Activity 5.4.4

Evaluate each of the following indefinite integrals.

1. $\int x^2 \sin(x)\ dx$
2. $\int t^3 \ln(t)\ dt$
3. $\int e^z \sin(z)\ dz$
4. $\int s^2 e^{3s}\ ds$
5. $\int t \arctan(t)\ dt$ (*Hint:* At a certain point in this problem, it is very helpful to note that $\frac{t^2}{1+t^2} = 1 - \frac{1}{1+t^2}$.)

Evaluating Definite Integrals Using Integration by Parts

We can use the technique of integration by parts to evaluate a definite integral.

Example 5.4.4

Evaluate

$$\int_0^{\pi/2} t\sin(t)\,dt.$$

Answer available at the end of this section.

As with any substitution technique, it is important to use notation carefully and completely, and to ensure that the end result makes sense.

When u-substitution and Integration by Parts Fail to Help

Both integration techniques we have discussed apply in relatively limited circumstances. It is not hard to find examples of functions for which neither technique produces an antiderivative; indeed, there are many, many functions that appear elementary but that do not have an elementary algebraic antiderivative. For instance, neither u-substitution nor integration by parts proves fruitful for the indefinite integrals

$$\int e^{x^2}\,dx \quad \text{and} \quad \int x\tan(x)\,dx.$$

While there are other integration techniques, some of which we will consider briefly, none of them enables us to find an algebraic antiderivative for e^{x^2} or $x\tan(x)$. We do know from the Second Fundamental Theorem of Calculus that we can construct an integral antiderivative for each function; $F(x)=\int_0^x e^{t^2}\,dt$ is an antiderivative of $f(x)=e^{x^2}$, and $G(x)=\int_0^x t\tan(t)\,dt$ is an antiderivative of $g(x)=x\tan(x)$. But finding an elementary algebraic formula that doesn't involve integrals for either F or G turns out not only to be impossible through u-substitution or integration by parts, but indeed impossible altogether. Antidifferentiation is much harder in general than differentiation.

Summary

- Through the method of integration by parts, we can evaluate indefinite integrals that involve products of basic functions such as $\int x\sin(x)\,dx$ and $\int x\ln(x)\,dx$. Using a substitution enables us to trade one of the functions in the product for its derivative, and the other for its antiderivative, in an effort to find a different product of functions that is easier to integrate.
- If the algebraic structure of an integrand is a product of basic functions in the form $\int f(x)g'(x)\,dx$, we can use the substitution $u=f(x)$ and $dv=g'(x)\,dx$ and apply the rule

  $$\int u\,dv = uv - \int v\,du$$

  to evaluate the original integral $\int f(x)g'(x)\,dx$ by instead evaluating

  $$\int v\,du = \int f'(x)g(x)\,dx.$$

- When deciding to integrate by parts, we have to select both u and dv. That selection is guided by the overall principle that the new integral $\int v\,du$ not be more difficult than the original integral $\int u\,dv$. In addition, it is often helpful to recognize if one of the functions present is much easier to differentiate than antidifferentiate (such as $\ln(x)$), in which case that function often is best assigned the variable u. In addition, dv must be a function that we can antidifferentiate.

Answers

Example 5.4.1

When we use integration by parts, we have a choice for u and dv. In this problem, we can either let $u = x$ and $dv = \cos(x)\ dx$, or let $u = \cos(x)$ and $dv = x\ dx$. While there is not a universal rule for how to choose u and dv, a good guideline is this: do so in a way that $\int v\ du$ is at least as simple as the original problem $\int u\ dv$.

This leads us to choose[1] $u = x$ and $dv = \cos(x)\ dx$, from which it follows that $du = 1\ dx$ and $v = \sin(x)$. With this substitution, the rule for integration by parts tells us that

$$\int x\cos(x)\,dx = x\sin(x) - \int \sin(x)\cdot 1\,dx.$$

Observe that if we considered the alternate choice, and let $u = \cos(x)$ and $dv = x\ dx$, then $du = -\sin(x)\ dx$ and $v = \frac{1}{2}x^2$, from which we would write $\int x\cos(x)\ dx = \frac{1}{2}x^2\cos(x) - \int \frac{1}{2}x^2(-\sin(x))\ dx$. Thus we have replaced the problem of integrating $x\cos(x)$ with that of integrating $\frac{1}{2}x^2\sin(x)$; the latter is clearly more complicated, which shows that this alternate choice is not as helpful as the first choice.

All that remains to do is evaluate the (simpler) integral $\int \sin(x)\cdot 1\ dx$. Doing so, we find

$$\int x\cos(x)\,dx = x\sin(x) - (-\cos(x)) + C = x\sin(x) + \cos(x) + C.$$

Observe that when we get to the final stage of evaluating the last remaining antiderivative, it is at this step that we include the integration constant, $+C$.

Example 5.4.2

Let $u = t^2$ and $dv = e^t\ dt$. Then $du = 2t\ dt$ and $v = e^t$, and thus

$$\int t^2e^t\,dt = t^2e^t - \int 2te^t\,dt.$$

The integral on the right side is simpler to evaluate than the one on the left, but it still requires integration by parts. Now letting $u = 2t$ and $dv = e^t\ dt$, we have $du = 2\ dt$ and $v = e^t$, so that

$$\int t^2e^t\,dt = t^2e^t - \left(2te^t - \int 2e^t\,dt\right).$$

(Note the parentheses, which remind us to distribute the minus sign to the entire value of the integral $\int 2te^t\ dt$.) The final integral on the right is a basic one; evaluating that integral and distributing the minus sign, we find

$$\int t^2e^t\,dt = t^2e^t - 2te^t + 2e^t + C.$$

Example 5.4.3

We can choose to let u be either e^t or $\cos(t)$; we pick $u = \cos(t)$, and thus $dv = e^t\ dt$. With $du = -\sin(t)\ dt$ and $v = e^t$, integration by parts tells us that

$$\int e^t\cos(t)\,dt = e^t\cos(t) - \int e^t(-\sin(t))\,dt,$$

or equivalently that

$$\int e^t\cos(t)\,dt = e^t\cos(t) + \int e^t\sin(t)\,dt. \tag{5.4.14}$$

The new integral has the same algebraic structure as the original one. While the overall situation isn't necessarily better than what we started with, it hasn't gotten worse. Thus, we proceed to integrate by parts again. This time we let $u = \sin(t)$ and $dv = e^t\ dt$, so that $du = \cos(t)\ dt$ and $v = e^t$, which implies

$$\int e^t\cos(t)\,dt = e^t\cos(t) + \left(e^t\sin(t) - \int e^t\cos(t)\,dt\right). \tag{5.4.15}$$

We seem to be back where we started, as two applications of integration by parts has led us back to the original problem, $\int e^t\cos(t)\ dt$. But if we look closely at Equation (5.4.15), we see that we can use algebra to solve for the value of the desired integral. Adding $\int e^t\cos(t)\ dt$ to both sides of the equation, we have

$$2\int e^t\cos(t)\,dt = e^t\cos(t) + e^t\sin(t),$$

and therefore

$$\int e^t\cos(t)\,dt = \frac{1}{2}\left(e^t\cos(t) + e^t\sin(t)\right) + C.$$

Note that since we never actually encountered an integral we could evaluate directly, we didn't have the opportunity to add the integration constant C until the final step.

Example 5.4.4

One option is to find an antiderivative (using indefinite integral notation) and then apply the Fundamental Theorem of Calculus to find that

$$\begin{aligned}\int_0^{\pi/2} t\sin(t)\,dt &= (-t\cos(t) + \sin(t))\Big|_0^{\pi/2}\\ &= \left(-\frac{\pi}{2}\cos(\frac{\pi}{2}) + \sin(\frac{\pi}{2})\right) - (-0\cos(0) + \sin(0))\\ &= 1.\end{aligned}$$

Alternatively, we can apply integration by parts and work with definite integrals throughout. With this method, we must remember to evaluate the product uv over the given limits of integration. Using the substitution $u = t$ and $dv = \sin(t)\ dt$, so that $du = dt$ and $v = -\cos(t)$, we write

$$\begin{aligned}\int_0^{\pi/2} t\sin(t)\,dt &= -t\cos(t)\Big|_0^{\pi/2} - \int_0^{\pi/2}(-\cos(t))\,dt\\ &= -t\cos(t)\Big|_0^{\pi/2} + \sin(t)\Big|_0^{\pi/2}\\ &= \left(-\frac{\pi}{2}\cos(\frac{\pi}{2}) + \sin(\frac{\pi}{2})\right) - (-0\cos(0) + \sin(0))\\ &= 1.\end{aligned}$$

This page titled 5.4: Integration by Parts is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.4: Integration by Parts by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

5.5: Other Options for Finding Algebraic Derivatives

Motivating Questions

- How does the method of partial fractions enable any rational function to be antidifferentiated?
- What role have integral tables historically played in the study of calculus and how can a table be used to evaluate integrals such as $\int \sqrt{a^2+u^2}\ du$?
- What role can a computer algebra system play in the process of finding antiderivatives?

We have learned two antidifferentiation techniques: u-substitution and integration by parts. The former is used to reverse the chain rule, while the latter to reverse the product rule. But we have seen that each works only in specialized circumstances. For example, while $\int xe^{x^2}\ dx$ may be evaluated by u-substitution and $\int xe^x\ dx$ by integration by parts, neither method provides a route to evaluate $\int e^{x^2}\ dx$, and in fact an elementary algebraic antiderivative for e^{x^2} does not exist. No antidifferentiation method will provide us with a simple algebraic formula for a function $F(x)$ that satisfies $F'(x) = e^{x^2}$.

In this section of the text, our main goals are to identify some classes of functions that can be antidifferentiated, and to learn some methods to do so. We should also recognize that there are many functions for which an algebraic formula for an antiderivative does not exist, and appreciate the role that computing technology can play in finding antiderivatives of other complicated functions.

Preview Activity 5.5.1

For each of the indefinite integrals below, the main question is to decide whether the integral can be evaluated using u -substitution, integration by parts, a combination of the two, or neither. For integrals for which your answer is affirmative, state the substitution(s) you would use. It is not necessary to actually evaluate any of the integrals completely, unless the integral can be evaluated immediately using a familiar basic antiderivative.

1. $\int x^2\sin(x^3)\ dx$, $\int x^2\sin(x)\ dx$, $\int \sin(x^3)\ dx$, $\int x^5\sin(x^3)\ dx$
2. $\int \frac{1}{1+x^2}\ dx$, $\int \frac{x}{1+x^2}\ dx$, $\int \frac{2x+3}{1+x^2}\ dx$, $\int \frac{e^x}{1+(e^x)^2}\ dx$,
3. $\int x\ln(x)\ dx$, $\int \frac{\ln(x)}{x}\ dx$, $\int \ln(1+x^2)\ dx$, $\int x\ln(1+x^2)\ dx$,
4. $\int x\sqrt{1-x^2}\ dx$, $\int \frac{1}{\sqrt{1-x^2}}\ dx$, $\int \frac{x}{\sqrt{1-x^2}}\ dx$, $\int \frac{1}{x\sqrt{1-x^2}}\ dx$,

The Method of Partial Fractions

The method of partial fractions is used to integrate rational functions. It involves reversing the process of finding a common denominator.

Example 5.5.1

Evaluate

$$\int \frac{5x}{x^2-x-2}\,dx.$$

Answer available at the end of this section.

It turns out that we can use the method of partial fractions to rewrite any ratinal function $R(x) = \frac{P(x)}{Q(x)}$ where the degree of the polynomial P is less than [1] the degree of Q as a sum of simpler rational functions of one of the following forms:

$$\frac{A}{x-c},\ \frac{A}{(x-c)^n},\ \frac{Ax+B}{x^2+k},\ \text{or}\ \frac{Ax+B}{(x^2+k)^n}$$

where A, B, and c are real numbers, and k is a positive real number. Because we can antidifferentiate each of these basic forms, partial fractions enables us to antidifferentiate any rational function.

If the degree of P is greater than or equal to the degree of Q, long division may be used to write R as the sum of a polynomial plus a rational function where the numerator's degree is less than the denominator's.

A computer algebra system such as *Maple*, *Mathematica*, or *WolframAlpha* can be used to find the partial fraction decomposition of any rational function. In *WolframAlpha*, entering

```
partial fraction 5x/(x^2-x-2)
```

results in the output

$$\frac{5x}{x^2-x-2} = \frac{10}{3(x-2)} + \frac{5}{3(x+1)}.$$

We will use technology to generate partial fraction decompositions of rational functions, and then evaluate the integrals using established methods.

Activity 5.5.2

For each of the following problems, evaluate the integral by using the partial fraction decomposition provided.

1. $\int \frac{1}{x^2-2x-3}\,dx$, given that $\frac{1}{x^2-2x-3} = \frac{1/4}{x-3} - \frac{1/4}{x+1}$
2. $\int \frac{x^2+1}{x^3-x^2}\,dx$, given that $\frac{x^2+1}{x^3-x^2} = -\frac{1}{x} - \frac{1}{x^2} + \frac{2}{x-1}$
3. $\int \frac{x-2}{x^4+x^2}\,dx$, given that $\frac{x-2}{x^4+x^2} = \frac{1}{x} - \frac{2}{x^2} + \frac{-x+2}{1+x^2}$

Using an Integral Table

Calculus has a long history, going back to Greek mathematicians in 400-300 BC. Its main foundations were first investigated and understood independently by Isaac Newton and Gottfried Wilhelm Leibniz in the late 1600s, making the modern ideas of calculus well over 300 years old. It is instructive to realize that until the late 1980s, the personal computer did not exist, so calculus (and other mathematics) had to be done by hand for roughly 300 years. In the 21st century, however, computers have revolutionized many aspects of the world we live in, including mathematics. In this section we take a short historical tour to precede discussing the role computer algebra systems can play in evaluating indefinite integrals. In particular, we consider a class of integrals involving certain radical expressions.

As seen in the short table of integrals found in Appendix A, there are many forms of integrals that involve $\sqrt{a^2 \pm w^2}$ and $\sqrt{w^2 - a^2}$. These integral rules can be developed using a technique known as *trigonometric substitution* that we choose to omit; instead, we will simply accept the results presented in the table. To see how these rules are used, consider the differences among

$$\int \frac{1}{\sqrt{1-x^2}}\,dx, \quad \int \frac{x}{\sqrt{1-x^2}}\,dx, \quad \text{and} \quad \int \sqrt{1-x^2}\,dx.$$

The first integral is a familiar basic one, and results in $\arcsin(x) + C$. The second integral can be evaluated using a standard u-substitution with $u = 1 - x^2$. The third, however, is not familiar and does not lend itself to u-substitution.

In Appendix A, we find the rule

$$(h)\ \int \sqrt{a^2-u^2}\,du = \frac{u}{2}\sqrt{a^2-u^2} + \frac{a^2}{2}\arcsin\left(\frac{u}{a}\right) + C.$$

Using the substitutions $a = 1$ and $u = x$ (so that $du = dx$), it follows that

$$\int \sqrt{1-x^2}\,dx = \frac{x}{2}\sqrt{1-x^2} - \frac{1}{2}\arcsin(x) + C.$$

Whenever we are applying a rule in the table, we are doing a u-substitution, especially when the substitution is more complicated than setting $u = x$ as in the last example.

Example 5.5.2

Evaluate the integral

$$\int \sqrt{9+64x^2}\,dx.$$

Answer available at the end of this section.

Whenever we use a u-subsitution in conjunction with Appendix A, it's important that we not forget to address any constants that arise and include them in our computations, such as the $\frac{1}{8}$ that appeared in Example 5.5.12.

Activity 5.5.3

For each of the following integrals, evaluate the integral using u-substitution and/or an entry from the table found in Appendix A.

1. $\displaystyle\int \sqrt{x^2+4}\ dx$
2. $\displaystyle\int \frac{x}{\sqrt{x^2+4}}\ dx$
3. $\displaystyle\int \frac{2}{\sqrt{16+25x^2}}\ dx$
4. $\displaystyle\int \frac{1}{x^2\sqrt{49-36x^2}}\ dx$

Using Computer Algebra Systems

A computer algebra system (CAS) is a computer program that is capable of executing symbolic mathematics. For example, if we ask a CAS to solve the equation $ax^2 + bx + c = 0$ for the variable x, where a, b, and c are arbitrary constants, the program will return $x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$. Research to develop the first CAS dates to the 1960s, and these programs became publicly available in the early 1990s. Two prominent examples are the programs *Maple* and *Mathematica*, which were among the first computer algebra systems to offer a graphical user interface. Today, *Maple* and *Mathematica* are exceptionally powerful professional software packages that can execute an amazing array of sophisticated mathematical computations. They are also very expensive, as each is a proprietary program. The CAS *SAGE* is an open-source, free alternative to *Maple* and *Mathematica*.

For the purposes of this text, when we need to use a CAS, we are going to turn instead to a similar, but somewhat different computational tool, the web-based "computational knowledge engine" called *WolframAlpha*. There are two features of *WolframAlpha* that make it stand out from the CAS options mentioned above: (1) unlike *Maple* and *Mathematica*, *WolframAlpha* is free (provided we are willing to navigate some pop-up advertising); and (2) unlike any of the three, the syntax in *WolframAlpha* is flexible. Think of *WolframAlpha* as being a little bit like doing a Google search: the program will interpret what is input, and then provide a summary of options.

If we want to have *WolframAlpha* evaluate an integral for us, we can provide it syntax such as

```
integrate x^2 dx
```

to which the program responds with

$$\int x^2\,dx = \frac{x^3}{3} + \text{constant}.$$

While there is much to be enthusiastic about regarding CAS programs such as *WolframAlpha*, there are several things we should be cautious about: (1) a CAS only responds to exactly what is input; (2) a CAS can answer using powerful functions from very advanced mathematics; and (3) there are problems that even a CAS cannot do without additional human insight.

Although (1) likely goes without saying, we have to be careful with our input: if we enter syntax that defines the wrong function, the CAS will work with precisely the function we define. For example, if we are interested in evaluating the integral

$$\int \frac{1}{16 - 5x^2}\,dx,$$

and we mistakenly enter

```
integrate 1/16 - 5x^2 dx
```

a CAS will (correctly) reply with

$$\frac{1}{16}x - \frac{5}{3}x^3.$$

But if we are sufficiently well-versed in antidifferentiation, we will recognize that this function cannot be the one that we seek: integrating a rational function such as $\frac{1}{16-5x^2}$, we expect the logarithm function to be present in the result.

Regarding (2), even for a relatively simple integral such as $\int \frac{1}{16-5x^2}\,dx$, some CASs will invoke advanced functions rather than simple ones. For instance, if we use *Maple* to execute the command

```
int(1/(16-5*x^2), x);
```

the program responds with

$$\int \frac{1}{16 - 5x^2}\,dx = \frac{\sqrt{5}}{20} arctanh(\frac{\sqrt{5}}{4}x).$$

While this is correct (save for the missing arbitrary constant, which *Maple* never reports), the inverse hyperbolic tangent function is not a common nor familiar one; a simpler way to express this function can be found by using the partial fractions method, and happens to be the result reported by *WolframAlpha*:

$$\int \frac{1}{16 - 5x^2}\,dx = \frac{1}{8\sqrt{5}}\left(\log(4\sqrt{5} + 5x) - \log(4\sqrt{5} - 5x)\right) + \text{constant}.$$

Using sophisticated functions from more advanced mathematics is sometimes the way a CAS says to the user "I don't know how to do this problem." For example, if we want to evaluate

$$\int e^{-x^2}\,dx,$$

and we ask *WolframAlpha* to do so, the input

```
integrate exp(-x^2) dx
```

results in the output

$$\int e^{-x^2}\,dx = \frac{\sqrt{\pi}}{2} erf(x) + \text{constant}.$$

The function "erf(x)" is the *error function*, which is actually defined by an integral:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,dt.$$

So, in producing output involving an integral, the CAS has basically reported back to us the very question we asked.

Finally, as remarked at (3) above, there are times that a CAS will actually fail without some additional human insight. If we consider the integral

$$\int (1+x)e^x\sqrt{1+x^2e^{2x}}\,dx$$

and ask *WolframAlpha* to evaluate

```
int (1+x) * exp(x) * sqrt(1+x^2 * exp(2x)) dx,
```

the program thinks for a moment and then reports

(*no result found in terms of standard mathematical functions*)

But in fact this integral is not that difficult to evaluate. If we let $u = xe^x$, then $du = (1+x)e^x\ dx$, which means that the preceding integral has form

$$\int (1+x)e^x\sqrt{1+x^2e^{2x}}\,dx = \int \sqrt{1+u^2}\,du,$$

which is a straightforward one for any CAS to evaluate.

So, we should proceed with some caution: while any CAS is capable of evaluating a wide range of integrals (both definite and indefinite), there are times when the result can mislead us. We must think carefully about the meaning of the output, whether it is consistent with what we expect, and whether or not it makes sense to proceed.

Summary

- We can antidifferentiate any rational function with the method of partial fractions. Any polynomial function can be factored into a product of linear and irreducible quadratic terms, so any rational function may be written as the sum of a polynomial plus rational terms of the form $\frac{A}{(x-c)^n}$ (where n is a natural number) and $\frac{Bx+C}{x^2+k}$ (where k is a positive real number).
- Until the development of compute algebra systems, integral tables enabled students of calculus to evaluate integrals such as $\int \sqrt{a^2+u^2}\ du$, where a is a positive real number. A short table of integrals may be found in Appendix A.
- Computer algebra systems can play an important role in finding antiderivatives, though we must be cautious to use correct input, to watch for unusual or unfamiliar advanced functions that the CAS may cite in its result, and to consider the possibility that a CAS may need further assistance or insight from us in order to answer a particular question.

Answers

Example 5.5.1

If we factor the denominator, we can see how R might be the sum of two fractions of the form $\frac{A}{x-2} + \frac{B}{x+1}$, so we suppose that

$$\frac{5x}{(x-2)(x+1)} = \frac{A}{x-2} + \frac{B}{x+1}$$

and look for the constants A and B.

Multiplying both sides of this equation by $(x-2)(x+1)$, we find that

$$5x = A(x+1) + B(x-2).$$

Since we want this equation to hold for every value of x, we can use insightful choices of specific x-values to help us find A and B. Taking $x = -1$, we have

$$5(-1) = A(0) + B(-3),$$

so $B = \frac{5}{3}$. Choosing $x = 2$, it follows

$$5(2) = A(3) + B(0),$$

so $A = \frac{10}{3}$. Thus,

$$\int \frac{5x}{x^2-x-2}\,dx = \int \frac{10/3}{x-2} + \frac{5/3}{x+1}\,dx.$$

This integral is straightforward to evaluate, and hence we find that

$$\int \frac{5x}{x^2-x-2}\,dx = \frac{10}{3}\ln|x-2| + \frac{5}{3}\ln|x+1| + C.$$

Example 5.5.2

Here, we want to use Rule (c) from the table, but we now set $a = 3$ and $u = 8x$. We also choose the "+" option in the rule. With this substitution, it follows that $du = 8dx$, so $dx = \frac{1}{8}du$. Applying the substitution,

$$\int \sqrt{9+64x^2}\,dx = \int \sqrt{9+u^2}\cdot\frac{1}{8}\,du = \frac{1}{8}\int \sqrt{9+u^2}\,du.$$

By Rule (c), we now find that

$$\begin{aligned}\int \sqrt{9+64x^2}\,dx &= \frac{1}{8}\left(\frac{u}{2}\sqrt{u^2+9}+\frac{9}{2}\ln|u+\sqrt{u^2+9}|+C\right)\\ &= \frac{1}{8}\left(\frac{8x}{2}\sqrt{64x^2+9}+\frac{9}{2}\ln|8x+\sqrt{64x^2+9}|+C\right).\end{aligned}$$

This page titled 5.5: Other Options for Finding Algebraic Derivatives is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.5: Other Options for Finding Algebraic Derivatives by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

5.6: Numerical Integration

Motivating Questions

- How do we accurately evaluate a definite integral such as $\int_0^1 e^{-x^2}\,dx$ when we cannot use the First Fundamental Theorem of Calculus because the integrand lacks an elementary algebraic antiderivative? Are there ways to generate accurate estimates without using extremely large values of n in Riemann sums?
- What is the Trapezoid Rule, and how is it related to left, right, and middle Riemann sums?
- How are the errors in the Trapezoid Rule and Midpoint Rule related, and how can they be used to develop an even more accurate rule?

When we first explored finding the net signed area bounded by a curve, we developed the concept of a Riemann sum as a helpful estimation tool and a key step in the definition of the definite integral. Recall that the left, right, and middle Riemann sums of a function f on an interval $[a, b]$ are given by

$$L_n = f(x_0)\Delta x + f(x_1)\Delta x + \cdots + f(x_{n-1})\Delta x = \sum_{i=0}^{n-1} f(x_i)\Delta x, \tag{5.6.1}$$

$$R_n = f(x_1)\Delta x + f(x_2)\Delta x + \cdots + f(x_n)\Delta x = \sum_{i=1}^{n} f(x_i)\Delta x, \tag{5.6.2}$$

$$M_n = f(\overline{x}_1)\Delta x + f(\overline{x}_2)\Delta x + \cdots + f(\overline{x}_n)\Delta x = \sum_{i=1}^{n} f(\overline{x}_i)\Delta x, \tag{5.6.3}$$

where $x_0 = a$, $x_i = a + i\Delta x$, $x_n = b$, and $\Delta x = \frac{b-a}{n}$. For the middle sum, we defined $\overline{x}_i = (x_{i-1} + x_i)/2$.

A Riemann sum is a sum of (possibly signed) areas of rectangles. The value of n determines the number of rectangles, and our choice of left endpoints, right endpoints, or midpoints determines the heights of the rectangles. We can see the similarities and differences among these three options in Figure 5.6.1, where we consider the function $f(x) = \frac{1}{20}(x-4)^3 + 7$ on the interval $[1, 8]$, and use 5 rectangles for each of the Riemann sums.

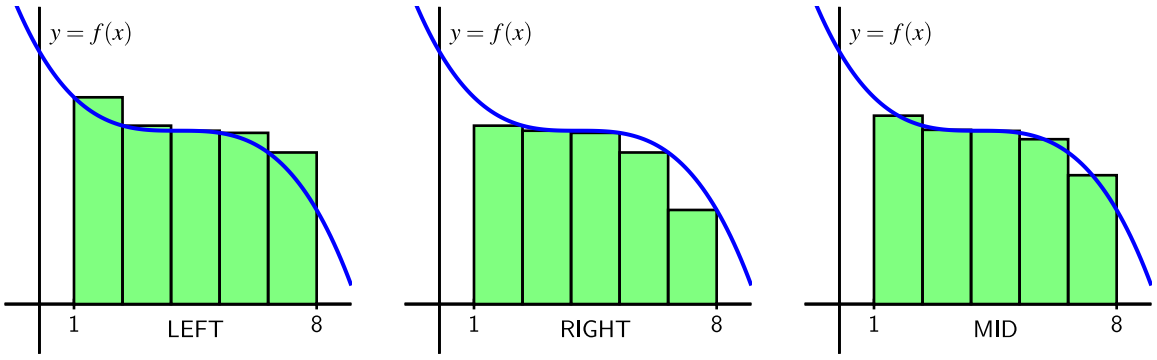


Figure 5.6.1: Left, right, and middle Riemann sums for $y = f(x)$ on $[1, 8]$ with 5 subintervals.

While it is a good exercise to compute a few Riemann sums by hand, just to ensure that we understand how they work and how varying the function, the number of subintervals, and the choice of endpoints or midpoints affects the result, using computing technology is the best way to determine L_n, R_n, and M_n. Any computer algebra system will offer this capability; as we saw in Preview Activity 4.3.1, a straightforward option that is freely available online is the applet[1] at http://gvsu.edu/s/a9. Note that we can adjust the formula for $f(x)$, the window of x- and y-values of interest, the number of subintervals, and the method. (See Preview Activity 4.3.1 for any needed reminders on how the applet works.)

Marc Renault, Shippensburg University

In this section we explore several different alternatives for estimating definite integrals. Our main goal is to develop formulas to estimate definite integrals accurately without using a large numbers of rectangles.

Preview Activity 5.6.1

As we begin to investigate ways to approximate definite integrals, it will be insightful to compare results to integrals whose exact values we know. To that end, the following sequence of questions centers on $\int_0^3 x^2 \, dx$.

1. Use the applet at http://gvsu.edu/s/a9 with the function $f(x) = x^2$ on the window of x values from 0 to 3 to compute L_3, the left Riemann sum with three subintervals.
2. Likewise, use the applet to compute R_3 and M_3, the right and middle Riemann sums with three subintervals, respectively.
3. Use the Fundamental Theorem of Calculus to compute the exact value of $I = \int_0^3 x^2 \, dx$.
4. We define the *error* that results from an approximation of a definite integral to be the approximation's value minus the integral's exact value. What is the error that results from using L_3? From R_3? From M_3?
5. In what follows in this section, we will learn a new approach to estimating the value of a definite integral known as the Trapezoid Rule. The basic idea is to use trapezoids, rather than rectangles, to estimate the area under a curve. What is the formula for the area of a trapezoid with bases of length b_1 and b_2 and height h?
6. Working by hand, estimate the area under $f(x) = x^2$ on $[0, 3]$ using three subintervals and three corresponding trapezoids. What is the error in this approximation? How does it compare to the errors you calculated in (d)?

The Trapezoid Rule

So far, we have used the simplest possible quadrilaterals (that is, rectangles) to estimate areas. It is natural, however, to wonder if other familiar shapes might serve us even better.

An alternative to L_n, R_n, and M_n is called the *Trapezoid Rule*. Rather than using a rectangle to estimate the (signed) area bounded by $y = f(x)$ on a small interval, we use a trapezoid. For example, in Figure 5.6.2, we estimate the area under the curve using three subintervals and the trapezoids that result from connecting the corresponding points on the curve with straight lines.

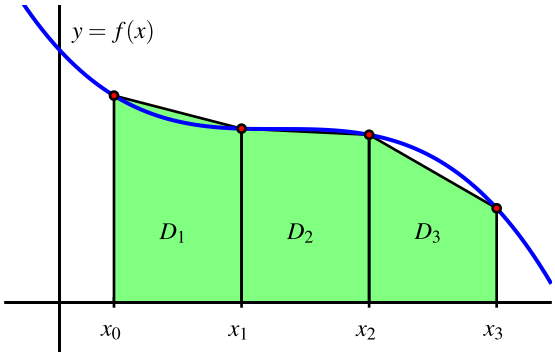


Figure 5.6.2: Estimating $\int_a^b f(x) \, dx$ using three subintervals and trapezoids, rather than rectangles, where $a = x_0$ and $b = x_3$.

The biggest difference between the Trapezoid Rule and a Riemann sum is that on each subinterval, the Trapezoid Rule uses two function values, rather than one, to estimate the (signed) area bounded by the curve. For instance, to compute D_1, the area of the trapezoid on $[x_0, x_1]$, we observe that the left base has length $f(x_0)$, while the right base has length $f(x_1)$. The height of the trapezoid is $x_1 - x_0 = \Delta x = \frac{b-a}{3}$. The area of a trapezoid is the average of the bases times the height, so we have

$$D_1 = \frac{1}{2}(f(x_0) + f(x_1)) \cdot \Delta x.$$

Using similar computations for D_2 and D_3, we find that T_3, the trapezoidal approximation to $\int_a^b f(x) \, dx$ is given by

$$\begin{aligned} T_3 &= D_1 + D_2 + D_3 \\ &= \frac{1}{2}(f(x_0) + f(x_1)) \cdot \Delta x + \frac{1}{2}(f(x_1) + f(x_2)) \cdot \Delta x + \frac{1}{2}(f(x_2) + f(x_3)) \cdot \Delta x. \end{aligned}$$

Because both left and right endpoints are being used, we recognize within the trapezoidal approximation the use of both left and right Riemann sums. Rearranging the expression for T_3 by removing factors of $\frac{1}{2}$ and Δx, grouping the left endpoint and right endpoint evaluations of f, we see that

$$T_3 = \frac{1}{2}\left[f(x_0) + f(x_1) + f(x_2)\right]\Delta x + \frac{1}{2}\left[f(x_1) + f(x_2) + f(x_3)\right]\Delta x. \tag{5.6.4}$$

We now observe that two familiar sums have arisen. The left Riemann sum L_3 is $L_3 = f(x_0)\Delta x + f(x_1)\Delta x + f(x_2)\Delta x$, and the right Riemann sum is $R_3 = f(x_1)\Delta x + f(x_2)\Delta x + f(x_3)\Delta x$. Substituting L_3 and R_3 for the corresponding expressions in Equation 5.6.4, it follows that $T_3 = \frac{1}{2}\left[L_3 + R_3\right]$. We have thus seen a very important result: using trapezoids to estimate the (signed) area bounded by a curve is the same as averaging the estimates generated by using left and right endpoints.

The Trapezoid Rule

The trapezoidal approximation, T_n, of the definite integral $\int_a^b f(x)\ dx$ using n subintervals is given by the rule

$$\begin{aligned} T_n &= \left[\frac{1}{2}(f(x_0) + f(x_1)) + \frac{1}{2}(f(x_1) + f(x_2)) + \cdots + \frac{1}{2}(f(x_{n-1}) + f(x_n))\right]\Delta x. \\ &= \sum_{i=0}^{n-1} \frac{1}{2}(f(x_i) + f(x_{i+1}))\Delta x. \end{aligned}$$

Moreover, $T_n = \frac{1}{2}\left[L_n + R_n\right]$.

Activity 5.6.2

In this activity, we explore the relationships among the errors generated by left, right, midpoint, and trapezoid approximations to the definite integral $\int_1^2 \frac{1}{x^2}\ dx$.

1. Use the First FTC to evaluate $\int_1^2 \frac{1}{x^2}\ dx$ exactly.
2. Use appropriate computing technology to compute the following approximations for $\int_1^2 \frac{1}{x^2}\ dx$: T_4, M_4, T_8, and M_8.
3. Let the *error* that results from an approximation be the approximation's value minus the exact value of the definite integral. For instance, if we let $E_{T,4}$ represent the error that results from using the trapezoid rule with 4 subintervals to estimate the integral, we have

   $$E_{T,4} = T_4 - \int_1^2 \frac{1}{x^2}\,dx.$$

   Similarly, we compute the error of the midpoint rule approximation with 8 subintervals by the formula

   $$E_{M,8} = M_8 - \int_1^2 \frac{1}{x^2}\,dx.$$

   Based on your work in (a) and (b) above, compute $E_{T,4}$, $E_{T,8}$, $E_{M,4}$, $E_{M,8}$.
4. Which rule consistently over-estimates the exact value of the definite integral? Which rule consistently under-estimates the definite integral?
5. What behavior(s) of the function $f(x) = \frac{1}{x^2}$ lead to your observations in (d)?

Comparing the Midpoint and Trapezoid Rules

We know from the definition of the definite integral that if we let n be large enough, we can make any of the approximations L_n, R_n, and M_n as close as we'd like (in theory) to the exact value of $\int_a^b f(x)\ dx$. Thus, it may be natural to wonder why we ever use

any rule other than L_n or R_n (with a sufficiently large n value) to estimate a definite integral. One of the primary reasons is that as $n \to \infty$, $\Delta x = \frac{b-a}{n} \to 0$, and thus in a Riemann sum calculation with a large n value, we end up multiplying by a number that is very close to zero. Doing so often generates roundoff error, because representing numbers close to zero accurately is a persistent challenge for computers.

Hence, we explore ways to estimate definite integrals to high levels of precision, but without using extremely large values of n. Paying close attention to patterns in errors, such as those observed in Activity 5.6.2, is one way to begin to see some alternate approaches.

To begin, we compare the errors in the Midpoint and Trapezoid rules. First, consider a function that is concave up on a given interval, and picture approximating the area bounded on that interval by both the Midpoint and Trapezoid rules using a single subinterval.

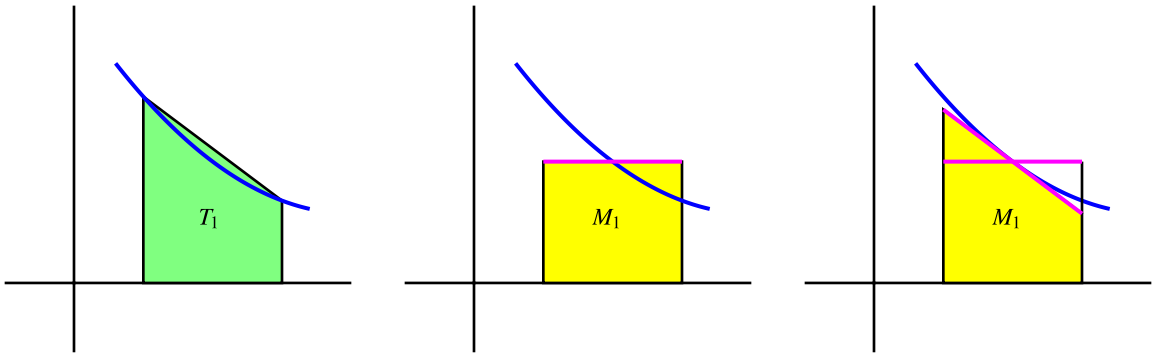


Figure 5.6.3: Estimating $\int_a^b f(x)\ dx$ using a single subinterval: at left, the trapezoid rule; in the middle, the midpoint rule; at right, a modified way to think about the midpoint rule.

As seen in Figure 5.6.3, it is evident that whenever the function is concave up on an interval, the Trapezoid Rule with one subinterval, T_1, will overestimate the exact value of the definite integral on that interval. From a careful analysis of the line that bounds the top of the rectangle for the Midpoint Rule (shown in magenta), we see that if we rotate this line segment until it is tangent to the curve at the midpoint of the interval (as shown at right in Figure 5.6.3), the resulting trapezoid has the same area as M_1, and this value is less than the exact value of the definite integral. Thus, when the function is concave up on the interval, M_1 underestimates the integral's true value.

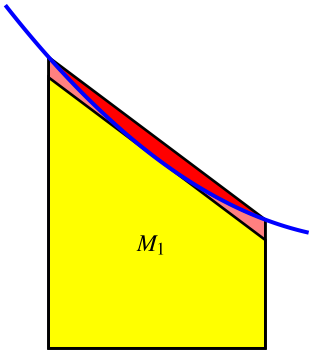


Figure 5.6.4: Comparing the error in estimating $\int_a^b f(x)\ dx$ using a single subinterval: in red, the error from the Trapezoid rule; in light red, the error from the Midpoint rule.

These observations extend easily to the situation where the function's concavity remains consistent but we use larger values of n in the Midpoint and Trapezoid Rules. Hence, whenever f is concave up on $[a, b]$, T_n will overestimate the value of $\int_a^b f(x)\ dx$, while M_n will underestimate $\int_a^b f(x)\ dx$. The reverse observations are true in the situation where f is concave down.

Next, we compare the size of the errors between M_n and T_n. Again, we focus on M_1 and T_1 on an interval where the concavity of f is consistent. In Figure 5.6.4, where the error of the Trapezoid Rule is shaded in red, while the error of the Midpoint Rule is shaded lighter red, it is visually apparent that the error in the Trapezoid Rule is more significant. To see how much more significant, let's consider two examples and some particular computations.

If we let $f(x) = 1 - x^2$ and consider $\int_0^1 f(x)\ dx$, we know by the First FTC that the exact value of the integral is

$$\int_0^1 (1 - x^2)\, dx = x - \frac{x^3}{3}\Big|_0^1 = \frac{2}{3}.$$

Using appropriate technology to compute M_4, M_8, T_4, and T_8, as well as the corresponding errors $E_{M,4}$, $E_{M,8}$, $E_{T,4}$, and $E_{T,8}$, as we did in Activity 5.6.2, we find the results summarized in Table 5.6.5. We also include the approximations and their errors for the example $\int_1^2 \frac{1}{x^2}\ dx$ from Activity 5.6.2.

Table 5.6.5: Calculations of T_4, M_4, T_8, and M_8, along with corresponding errors, for the definite integrals $\int_0^1 (1 - x^2)\ dx$ and $\int_1^2 \frac{1}{x^2}\ dx$.

Rule	$\int_0^1 (1 - x^2)\ dx = 0.\overline{6}$	error	$\int_1^2 \frac{1}{x^2}\ dx = 0.5$	error
T_4	0.65625	−0.0104166667	0.5089937642	0.0089937642
M_4	0.671875	0.0052083333	0.4955479365	−0.0044520635
T_8	0.6640625	−0.0026041667	0.5022708502	0.0022708502
M_8	0.66796875	0.0013020833	0.4988674899	−0.0011325101

For a given function f and interval $[a, b]$, $E_{T,4} = T_4 - \int_a^b f(x)\ dx$ calculates the difference between the approximation generated by the Trapezoid Rule with $n = 4$ and the exact value of the definite integral. If we look at not only $E_{T,4}$, but also the other errors generated by using T_n and M_n with $n = 4$ and $n = 8$ in the two examples noted in Table 5.6.5, we see an evident pattern. Not only is the sign of the error (which measures whether the rule generates an over- or under-estimate) tied to the rule used and the function's concavity, but the magnitude of the errors generated by T_n and M_n seems closely connected. In particular, the errors generated by the Midpoint Rule seem to be about half the size (in absolute value) of those generated by the Trapezoid Rule.

That is, we can observe in both examples that $E_{M,4} \approx -\frac{1}{2}E_{T,4}$ and $E_{M,8} \approx -\frac{1}{2}E_{T,8}$. This property of the Midpoint and Trapezoid Rules turns out to hold in general: for a function of consistent concavity, the error in the Midpoint Rule has the opposite sign and approximately half the magnitude of the error of the Trapezoid Rule. Written symbolically,

$$E_{M,n} \approx -\frac{1}{2}E_{T,n}.$$

This important relationship suggests a way to combine the Midpoint and Trapezoid Rules to create an even more accurate approximation to a definite integral.

Simpson's Rule

When we first developed the Trapezoid Rule, we observed that it is an average of the Left and Right Riemann sums:

$$T_n = \frac{1}{2}(L_n + R_n).$$

If a function is always increasing or always decreasing on the interval $[a, b]$, one of L_n and R_n will over-estimate the true value of $\int_a^b f(x)\ dx$, while the other will under-estimate the integral. Thus, the errors that result from L_n and R_n will have opposite signs; so averaging L_n and R_n eliminates a considerable amount of the error present in the respective approximations. In a similar way, it makes sense to think about averaging M_n and T_n in order to generate a still more accurate approximation.

We've just observed that M_n is typically about twice as accurate as T_n. So we use the weighted average

$$S_{2n} = \frac{2M_n + T_n}{3}. \tag{5.6.5}$$

The rule for S_{2n} giving by Equation 5.6.5 is usually known as *Simpson's Rule*.[2] Note that we use "S_{2n}" rather that "S_n" since the n points the Midpoint Rule uses are different from the n points the Trapezoid Rule uses, and thus Simpson's Rule is using $2n$ points at which to evaluate the function. We build upon the results in Table 5.6.5 to see the approximations generated by Simpson's Rule. In particular, in Table 5.6.6, we include all of the results in Table 5.6.5, but include additional results for $S_8 = \frac{2M_4+T_4}{3}$ and $S_{16} = \frac{2M_8+T_8}{3}$.

Thomas Simpson was an 18th century mathematician; his idea was to extend the Trapezoid rule, but rather than using straight lines to build trapezoids, to use quadratic functions to build regions whose area was bounded by parabolas (whose areas he could find exactly). Simpson's Rule is often developed from the more sophisticated perspective of using interpolation by quadratic functions.

Table 5.6.6: Table 5.6.5 updated to include S_8, S_{16}, and the corresponding errors.

Rule	$\int_0^1 (1-x^2)\ dx = 0.\overline{6}$	error	$\int_1^2 \frac{1}{x^2}\ dx = 0.5$	error
T_4	0.65625	-0.0104166667	0.5089937642	0.0089937642
M_4	0.671875	0.0052083333	0.4955479365	-0.0044520635
S_8	0.6666666667	0	0.5000298792	0.0000298792
T_8	0.6640625	-0.0026041667	0.5022708502	0.0022708502
M_8	0.66796875	0.0013020833	0.4988674899	-0.0011325101
S_{16}	0.6666666667	0	0.5000019434	0.0000019434

The results seen in Table 5.6.6 are striking. If we consider the S_{16} approximation of $\int_1^2 \frac{1}{x^2}\ dx$, the error is only $E_{S,16} = 0.0000019434$. By contrast, $L_8 = 0.5491458502$, so the error of that estimate is $E_{L,8} = 0.0491458502$. Moreover, we observe that generating the approximations for Simpson's Rule is almost no additional work: once we have L_n, R_n, and M_n for a given value of n, it is a simple exercise to generate T_n, and from there to calculate S_{2n}. Finally, note that the error in the Simpson's Rule approximations of $\int_0^1 (1-x^2)\ dx$ is zero![3]

Similar to how the Midpoint and Trapezoid approximations are exact for linear functions, Simpson's Rule approximations are exact for quadratic and cubic functions. See additional discussion on this issue later in the section and in the exercises.

These rules are not only useful for approximating definite integrals such as $\int_0^1 e^{-x^2}\ dx$, for which we cannot find an elementary antiderivative of e^{-x^2}, but also for approximating definite integrals when we are given a function through a table of data.

Activity 5.6.3

A car traveling along a straight road is braking and its velocity is measured at several different points in time, as given in the following table. Assume that v is continuous, always decreasing, and always decreasing at a decreasing rate, as is suggested by the data.

Table 5.6.7 Data for the braking car.

seconds, t	Velocity in ft/sec, $v(t)$
0	100
0.3	99
0.6	96
0.9	90
1.2	80
1.5	50
1.8	0

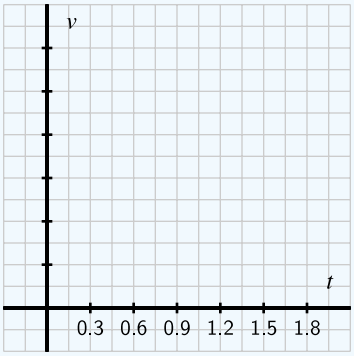


Figure 5.6.8: Axes for plotting the data in Activity 5.6.3.

1. Plot the given data on the set of axes provided in Figure 5.6.8 with time on the horizontal axis and the velocity on the vertical axis.
2. What definite integral will give you the exact distance the car traveled on $[0, 1.8]$?
3. Estimate the total distance traveled on $[0, 1.8]$ by computing L_3, R_3, and T_3. Which of these under-estimates the true distance traveled?
4. Estimate the total distance traveled on $[0, 1.8]$ by computing M_3. Is this an over- or under-estimate? Why?
5. Using your results from (c) and (d), improve your estimate further by using Simpson's Rule.
6. What is your best estimate of the average velocity of the car on $[0, 1.8]$? Why? What are the units on this quantity?

Overall observations regarding L_n, R_n, T_n, M_n, and S_{2n}.

As we conclude our discussion of numerical approximation of definite integrals, it is important to summarize general trends in how the various rules over- or under-estimate the true value of a definite integral, and by how much. To revisit some past observations and see some new ones, we consider the following activity.

Activity 5.6.4

Consider the functions $f(x) = 2 - x^2$, $g(x) = 2 - x^3$, and $h(x) = 2 - x^4$, all on the interval $[0, 1]$. For each of the questions that require a numerical answer in what follows, write your answer exactly in fraction form.

1. On the three sets of axes provided in Figure 5.6.9, sketch a graph of each function on the interval $[0, 1]$, and compute L_1 and R_1 for each. What do you observe?
2. Compute M_1 for each function to approximate $\int_0^1 f(x)\ dx$, $\int_0^1 g(x)\ dx$, and $\int_0^1 h(x)\ dx$, respectively.
3. Compute T_1 for each of the three functions, and hence compute S_2 for each of the three functions.
4. Evaluate each of the integrals $\int_0^1 f(x)\ dx$, $\int_0^1 g(x)\ dx$, and $\int_0^1 h(x)\ dx$ exactly using the First FTC.
5. For each of the three functions f, g, and h, compare the results of L_1, R_1, M_1, T_1, and S_2 to the true value of the corresponding definite integral. What patterns do you observe?

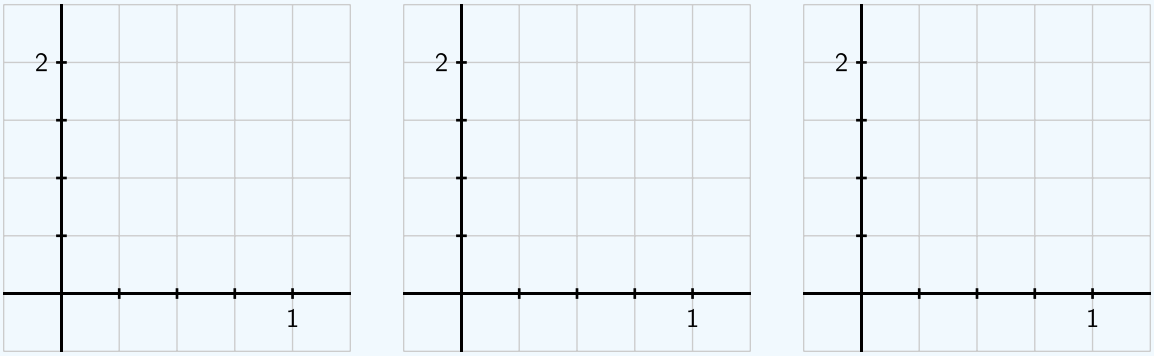


Figure 5.6.9: Axes for plotting the functions in Activity 5.6.4.

The results seen in Activity 5.6.4 generalize nicely. For instance, if f is decreasing on $[a, b]$, L_n will over-estimate the exact value of $\int_a^b f(x)\ dx$, and if f is concave down on $[a, b]$, M_n will over-estimate the exact value of the integral. An excellent exercise is to write a collection of scenarios of possible function behavior, and then categorize whether each of L_n, R_n, T_n, and M_n is an over- or under-estimate.

Finally, we make two important notes about Simpson's Rule. When T. Simpson first developed this rule, his idea was to replace the function f on a given interval with a quadratic function that shared three values with the function f. In so doing, he guaranteed that this new approximation rule would be exact for the definite integral of any quadratic polynomial. In one of the pleasant surprises of numerical analysis, it turns out that even though it was designed to be exact for quadratic polynomials, Simpson's Rule is exact for any cubic polynomial: that is, if we are interested in an integral such as $\int_2^5 (5x^3 - 2x^2 + 7x - 4)\ dx$, S_{2n} will always be exact, regardless of the value of n. This is just one more piece of evidence that shows how effective Simpson's Rule is as an approximation tool for estimating definite integrals. [4]

One reason that Simpson's Rule is so effective is that S_{2n} benefits from using $2n + 1$ points of data. Because it combines M_n, which uses n midpoints, and T_n, which uses the $n + 1$ endpoints of the chosen subintervals, S_{2n} takes advantage of the maximum amount of information we have when we know function values at the endpoints and midpoints of n subintervals.

Summary

- For a definite integral such as $\int_0^1 e^{-x^2}\ dx$ when we cannot use the First Fundamental Theorem of Calculus because the integrand lacks an elementary algebraic antiderivative, we can estimate the integral's value by using a sequence of Riemann sum approximations. Typically, we start by computing L_n, R_n, and M_n for one or more chosen values of n.
- The Trapezoid Rule, which estimates $\int_a^b f(x)\ dx$ by using trapezoids, rather than rectangles, can also be viewed as the average of Left and Right Riemann sums. That is, $T_n = \frac{1}{2}(L_n + R_n)$.
- The Midpoint Rule is typically twice as accurate as the Trapezoid Rule, and the signs of the respective errors of these rules are opposites. Hence, by taking the weighted average $S_{2n} = \frac{2M_n + T_n}{3}$, we can build a much more accurate approximation to $\int_a^b f(x)\ dx$ by using approximations we have already computed. The rule for S_{2n} is known as Simpson's Rule, which can also be developed by approximating a given continuous function with pieces of quadratic polynomials.

This page titled 5.6: Numerical Integration is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.6: Numerical Integration by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

5.7: Finding Antiderivatives and Evaluating Integrals (Exercises)

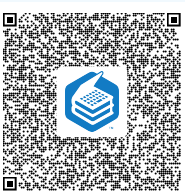
5.1: Construction Accurate Graphs of Antiderivatives

Exercises 5.1.5 Exercises

1. Definite integral of a piecewise linear function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. A smooth function that starts out at 0

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. A piecewise constant function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Another piecewise linear function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5

A moving particle has its velocity given by the quadratic function v pictured in Figure 5.1.7. In addition, it is given that $A_1 = \frac{7}{6}$ and $A_2 = \frac{8}{3}$, as well as that for the corresponding position function s, $s(0) = 0.5$.

1. Use the given information to determine $s(1)$, $s(3)$, $s(5)$, and $s(6)$.
2. On what interval(s) is s increasing? On what interval(s) is s decreasing?
3. On what interval(s) is s concave up? On what interval(s) is s concave down?
4. Sketch an accurate, labeled graph of s on the axes at right in Figure 5.1.7.
5. Note that $v(t) = -2 + \frac{1}{2}(t-3)^2$. Find a formula for s.

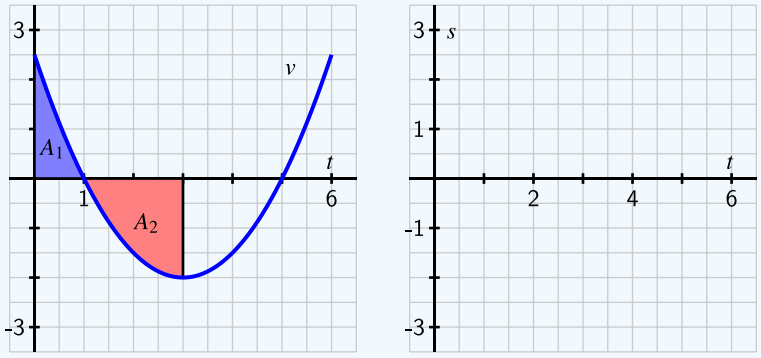


Figure 5.1.7. At left, the given graph of v. At right, axes for plotting s.

6

A person exercising on a treadmill experiences different levels of resistance and thus burns calories at different rates, depending on the treadmill's setting. In a particular workout, the rate at which a person is burning calories is given by the piecewise constant function c pictured in Figure 5.1.8. Note that the units on c are “calories per minute.”

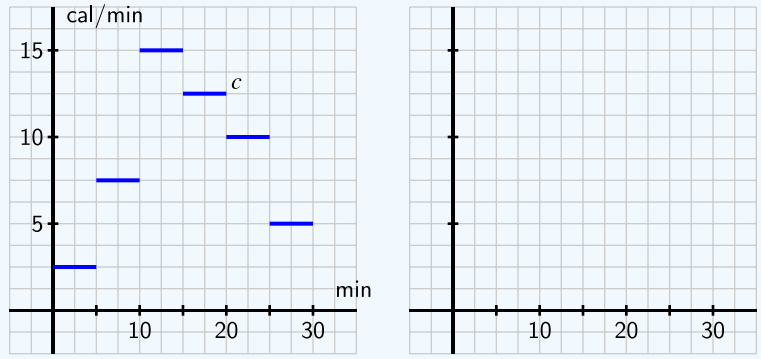


Figure 5.1.8. At left, the given graph of c. At right, axes for plotting C.

1. Let C be an antiderivative of c. What does the function C measure? What are its units?
2. Assume that $C(0) = 0$. Determine the exact value of $C(t)$ at the values $t = 5, 10, 15, 20, 25, 30$.
3. Sketch an accurate graph of C on the axes provided at right in Figure 5.1.8. Be certain to label the scale on the vertical axis.
4. Determine a formula for C that does not involve an integral and is valid for $5 \leq t \leq 10$.

7.

Consider the piecewise linear function f given in Figure 5.1.9. Let the functions A, B, and C be defined by the rules $A(x) = \int_{-1}^{x} f(t)\ dt$, $B(x) = \int_{0}^{x} f(t)\ dt$, and $C(x) = \int_{1}^{x} f(t)\ dt$.

1. For the values $x = -1, 0, 1, \ldots, 6$, make a table that lists corresponding values of $A(x)$, $B(x)$, and $C(x)$.
2. On the axes provided in Figure 5.1.9, sketch the graphs of A, B, and C.
3. How are the graphs of A, B, and C related?
4. How would you best describe the relationship between the function A and the function f?

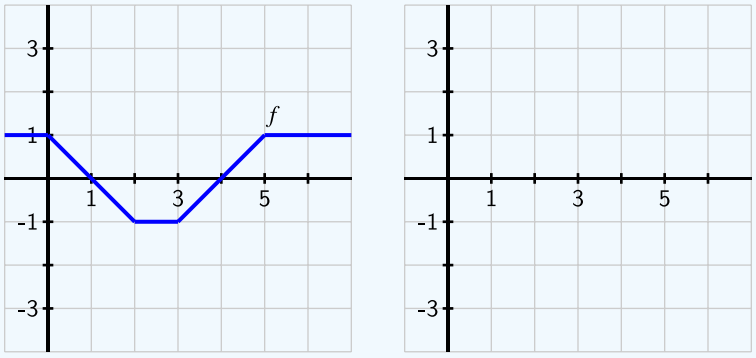


Figure 5.1.9. At left, the given graph of f. At right, axes for plotting A, B, and C.

5.2: The Second Fundamental Theorem of Calculus

1. A definite integral starting at 3

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Variable in the lower limit

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Approximating a function with derivative $e^{-x^2/5}$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4

Let g be the function pictured at left in Figure 5.2.6, and let F be defined by $F(x) = \int_2^x g(t)\ dt$. Assume that the shaded areas have values $A_1 = 4.29$, $A_2 = 12.75$, $A_3 = 0.36$, and $A_4 = 1.79$. Assume further that the portion of A_2 that lies between $x = 0.5$ and $x = 2$ is 6.06.

Sketch a carefully labeled graph of F on the axes provided, and include a written analysis of how you know where F is zero, increasing, decreasing, CCU, and CCD.

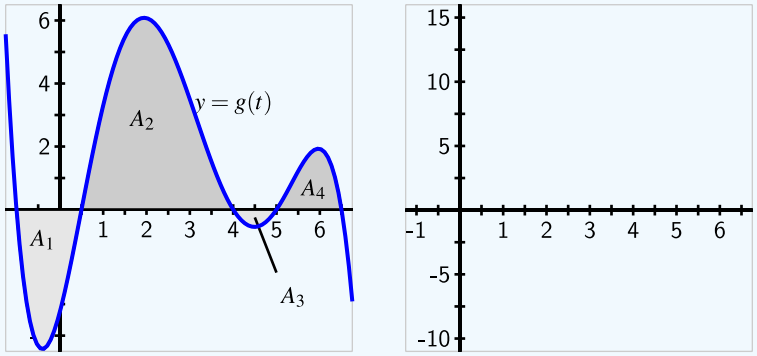


Figure 5.2.6. At left, the graph of g. At right, axes for plotting F.

5

The tide removes sand from the beach at a small ocean park at a rate modeled by the function

$$R(t) = 2 + 5\sin\left(\frac{4\pi t}{25}\right)$$

A pumping station adds sand to the beach at rate modeled by the function

$$S(t) = \frac{15t}{1+3t}$$

Both $R(t)$ and $S(t)$ are measured in cubic yards of sand per hour, t is measured in hours, and the valid times are $0 \le t \le 6$. At time $t = 0$, the beach holds 2500 cubic yards of sand.

1. What definite integral measures how much sand the tide will remove during the time period $0 \le t \le 6$? Why?
2. Write an expression for $Y(x)$, the total number of cubic yards of sand on the beach at time x. Carefully explain your thinking and reasoning.
3. At what instantaneous rate is the total number of cubic yards of sand on the beach at time $t = 4$ changing?
4. Over the time interval $0 \le t \le 6$, at what time t is the amount of sand on the beach least? What is this minimum value? Explain and justify your answers fully.

6

When an aircraft attempts to climb as rapidly as possible, its climb rate (in feet per minute) decreases as altitude increases, because the air is less dense at higher altitudes. Given below is a table showing performance data for a certain single engine aircraft, giving its climb rate at various altitudes, where $c(h)$ denotes the climb rate of the airplane at an altitude h.

Table 5.2.7. Data for the climbing aircraft.

h (feet)	0	1000	2000	3000	4000	5000	6000	7000	8000	9000	10,000
c (ft/min)	925	875	830	780	730	685	635	585	535	490	440

Let a new function m, that also depends on h, (say $y = m(h)$) measure the number of minutes required for a plane at altitude h to climb the next foot of altitude.

1. Determine a similar table of values for $m(h)$ and explain how it is related to the table above. Be sure to discuss the units on m.
2. Give a careful interpretation of a function whose derivative is $m(h)$. Describe what the input is and what the output is. Also, explain in plain English what the function tells us.
3. Determine a definite integral whose value tells us exactly the number of minutes required for the airplane to ascend to 10,000 feet of altitude. Clearly explain why the value of this integral has the required meaning.
4. Determine a formula for a function $M(h)$ whose value tells us the exact number of minutes required for the airplane to ascend to h feet of altitude.
5. Estimate the values of $M(6000)$ and $M(10000)$ as accurately as you can. Include units on your results.

5.3 Integration by Substitution

1. Product involving 4th power of a polynomial

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Product involving $\sin(x^6)$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Fraction involving $\ln^9$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Fraction involving e^{5x}

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Fraction involving $e^{5\sqrt{y}}$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6. Definite integral involving $e^{-cos(q)}$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

7

This problem centers on finding antiderivatives for the basic trigonometric functions other than $\sin(x)$ and $\cos(x)$.

1. Consider the indefinite integral $\int \tan(x)\ dx$. By rewriting the integrand as $\tan(x) = \frac{\sin(x)}{\cos(x)}$ and identifying an appropriate function-derivative pair, make a u-substitution and hence evaluate $\int \tan(x)\ dx$.
2. In a similar way, evaluate $\int \cot(x)\ dx$.
3. Consider the indefinite integral

$$\int \frac{\sec^2(x) + \sec(x)\tan(x)}{\sec(x) + \tan(x)}\, dx.$$

   Evaluate this integral using the substitution $u = \sec(x) + \tan(x)$.

4. Simplify the integrand in (c) by factoring the numerator. What is a far simpler way to write the integrand?
5. Combine your work in (c) and (d) to determine $\int \sec(x)\ dx$.
6. Using (c)-(e) as a guide, evaluate $\int \csc(x)\ dx$.

8

Consider the indefinite integral $\int x\sqrt{x-1}\ dx$.

1. At first glance, this integrand may not seem suited to substitution due to the presence of x in separate locations in the integrand. Nonetheless, using the composite function $\sqrt{x-1}$ as a guide, let $u = x - 1$. Determine expressions for both x and dx in terms of u.

2. Convert the given integral in x to a new integral in u.
3. Evaluate the integral in (b) by noting that $\sqrt{u} = u^{1/2}$ and observing that it is now possible to rewrite the integrand in u by expanding through multiplication.
4. Evaluate each of the integrals $\int x^2\sqrt{x-1}\ dx$ and $\int x\sqrt{x^2-1}\ dx$. Write a paragraph to discuss the similarities among the three indefinite integrals in this problem and the role of substitution and algebraic rearrangement in each.

9

Consider the indefinite integral $\int \sin^3(x)\ dx$.

1. Explain why the substitution $u = \sin(x)$ will not work to help evaluate the given integral.
2. Recall the Fundamental Trigonometric Identity, which states that $\sin^2(x) + \cos^2(x) = 1$. By observing that $\sin^3(x) = \sin(x) \cdot \sin^2(x)$, use the Fundamental Trigonometric Identity to rewrite the integrand as the product of $\sin(x)$ with another function.
3. Explain why the substitution $u = \cos(x)$ now provides a possible way to evaluate the integral in (b).
4. Use your work in (a)-(c) to evaluate the indefinite integral $\int \sin^3(x)\ dx$.
5. Use a similar approach to evaluate $\int \cos^3(x)\ dx$.

10

For the town of Mathland, MI, residential power consumption has shown certain trends over recent years. Based on data reflecting average usage, engineers at the power company have modeled the town's rate of energy consumption by the function

$$r(t) = 4 + \sin(0.263t + 4.7) + \cos(0.526t + 9.4).$$

Here, t measures time in hours after midnight on a typical weekday, and r is the rate of consumption in megawatts at time t. Units are critical throughout this problem.

The unit *megawatt* is itself a rate, which measures energy consumption per unit time. A *megawatt-hour* is the total amount of energy that is equivalent to a constant stream of 1 megawatt of power being sustained for 1 hour.

1. Sketch a carefully labeled graph of $r(t)$ on the interval [0,24] and explain its meaning. Why is this a reasonable model of power consumption?
2. Without calculating its value, explain the meaning of $\int_0^{24} r(t)\ dt$. Include appropriate units on your answer.
3. Determine the exact amount of energy Mathland consumes in a typical day.
4. What is Mathland's average rate of power consumption in a given 24-hour period? What are the units on this quantity?

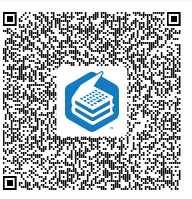
The unit *megawatt* is itself a rate, which measures energy consumption per unit time. A *megawatt-hour* is the total amount of energy that is equivalent to a constant stream of 1 megawatt of power being sustained for 1 hour.

5.4: Integration by Parts

1. Choose which method to use

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. Product involving $\cos(5x)$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Product involving e^{8z}

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Definite integral of te^{-t}

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5

Let $f(t) = te^{-2t}$ and $F(x) = \int_0^x f(t)\ dt$.

1. Determine $F'(x)$.
2. Use the First FTC to find a formula for F that does not involve an integral.
3. Is F an increasing or decreasing function for $x > 0$? Why?

6

Consider the indefinite integral given by $\int e^{2x}\cos(e^x)\ dx$.

1. Noting that $e^{2x} = e^x \cdot e^x$, use the substitution $z = e^x$ to determine a new, equivalent integral in the variable z.
2. Evaluate the integral you found in (a) using an appropriate technique.
3. How is the problem of evaluating $\int e^{2x}\cos(e^{2x})\ dx$ different from evaluating the integral in (a)? Do so.
4. Evaluate each of the following integrals as well, keeping in mind the approach(es) used earlier in this problem:
   - $\int e^{2x}\sin(e^x)\ dx$
   - $\int e^{3x}\sin(e^{3x})\ dx$
   - $\int xe^{x^2}\cos(e^{x^2})\sin(e^{x^2})\ dx$

7

For each of the following indefinite integrals, determine whether you would use u-substitution, integration by parts, neither*, or both to evaluate the integral. In each case, write one sentence to explain your reasoning, and include a statement of any substitutions used. (That is, if you decide in a problem to let $u = e^{3x}$, you should state that, as well as that $du = 3e^{3x}\ dx$.) Finally, use your chosen approach to evaluate each integral. (* one of the following problems does not have an elementary antiderivative and you are not expected to actually evaluate this integral; this will correspond with a choice of "neither" among those given.)

1. $\int x^2 \cos(x^3)\ dx$
2. $\int x^5 \cos(x^3)\ dx$ (Hint: $x^5 = x^2 \cdot x^3$)
3. $\int x \ln(x^2)\ dx$
4. $\int \sin(x^4)\ dx$
5. $\int x^3 \sin(x^4)\ dx$
6. $\int x^7 \sin(x^4)\ dx$

5.5: Other Options for Finding Algebraic Derivatives

1. Partial fractions: linear over difference of squares

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Partial fractions: constant over product

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Partial fractions: linear over quadratic

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Partial fractions: cubic over 4th degree

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Partial fractions: quadratic over factored cubic

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6

For each of the following integrals involving rational functions, (1) use a CAS to find the partial fraction decomposition of the integrand; (2) evaluate the integral of the resulting function without the assistance of technology; (3) use a CAS to evaluate the original integral to test and compare your result in (2).

1. $\displaystyle\int \frac{x^3+x+1}{x^4-1}\,dx$
2. $\displaystyle\int \frac{x^5+x^2+3}{x^3-6x^2+11x-6}\,dx$
3. $\displaystyle\int \frac{x^2-x-1}{(x-3)^3}\,dx$

7

For each of the following integrals involving radical functions, (1) use an appropriate u-substitution along with Appendix A to evaluate the integral without the assistance of technology, and (2) use a CAS to evaluate the original integral to test and compare your result in (1).

1. $\displaystyle\int \frac{1}{x\sqrt{9x^2+25}}\,dx$
2. $\displaystyle\int x\sqrt{1+x^4}\,dx$
3. $\displaystyle\int e^x\sqrt{4+e^{2x}}\,dx$
4. $\displaystyle\int \frac{\tan(x)}{\sqrt{9-\cos^2(x)}}\,dx$

8

Consider the indefinite integral given by

$$\int \frac{\sqrt{x+\sqrt{1+x^2}}}{x}\,dx.$$

1. Explain why u-substitution does not offer a way to simplify this integral by discussing at least two different options you might try for u.
2. Explain why integration by parts does not seem to be a reasonable way to proceed, either, by considering one option for u and dv.
3. Is there any line in the integral table in Appendix A that is helpful for this integral?
4. Evaluate the given integral using *WolframAlpha*. What do you observe?

5.6: Numerical Integration

Two notes about how Exercise 5.6.6.1 is coded: (i) as explained in the problem header, you need responses to every single entry before you can get individual parts marked as correct; if you enter only an answer for (a) and submit, (a) will be marked wrong regardless. And (ii), in this problem, the notation "SIMP(2)" is actually what we have called "SIMP(4)" in our previous work. Different authors use different notation, and the author of this WeBWorK exercise chooses to write "SIMP(n)" where we have written "SIMP(2n)" in Section 5.6.

1. Various methods for e^x numerically

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. Comparison of methods for increasing concave down function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Comparing accuracy for two similar functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Identifying and comparing methods

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5

Consider the definite integral $\int_0^1 x\tan(x)\ dx$.

1. Explain why this integral cannot be evaluated exactly by using either u-substitution or by integrating by parts.
2. Using appropriate subintervals, compute L_4, R_4, M_4, T_4, and S_8.
3. Which of the approximations in (b) is an over-estimate to the true value of $\int_0^1 x\tan(x)\ dx$? Which is an under-estimate? How do you know?

6

For an unknown function $f(x)$, the following information is known.

- f is continuous on $[3,6]$;
- f is either always increasing or always decreasing on $[3,6]$;
- f has the same concavity throughout the interval $[3,6]$;
- As approximations to $\int_3^6 f(x)\ dx$, $L_4 = 7.23$, $R_4 = 6.75$, and $M_4 = 7.05$.

1. Is f increasing or decreasing on $[3,6]$? What data tells you?
2. Is f concave up or concave down on $[3,6]$? Why?
3. Determine the best possible estimate you can for $\int_3^6 f(x)\ dx$, based on the given information.

7

The rate at which water flows through Table Rock Dam on the White River in Branson, MO, is measured in cubic feet per second (CFS). As engineers open the floodgates, flow rates are recorded according to the following chart.

Table 5.6.16. Water flow data.

seconds, t	0	10	20	30	40	50	60
flow in CFS, $r(t)$	2000	2100	2400	3000	3900	5100	6500

1. What definite integral measures the total volume of water to flow through the dam in the 60 second time period provided by the table above?
2. Use the given data to calculate M_n for the largest possible value of n to approximate the integral you stated in (a). Do you think M_n over- or under-estimates the exact value of the integral? Why?
3. Approximate the integral stated in (a) by calculating S_n for the largest possible value of n, based on the given data.
4. Compute $\frac{1}{60}S_n$ and $\frac{2000+2100+2400+3000+3900+5100+6500}{7}$. What quantity do both of these values estimate? Which is a more accurate approximation?

This page titled 5.7: Finding Antiderivatives and Evaluating Integrals (Exercises) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.E: Finding Antiderivatives and Evaluating Integrals (Exercises) has no license indicated.
- 5.1: Construction Accurate Graphs of Antiderivatives by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 5.2: The Second Fundamental Theorem of Calculus by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 5.3: Integration by Substitution by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 5.4: Integration by Parts by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 5.5: Other Options for Finding Algebraic Derivatives by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 5.6: Numerical Integration by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

6: USING DEFINITE INTEGRALS

6.1: Using Definite Integrals to Find Area and Length

6.2: Using Definite Integrals to Find Volume

6.3: Density, Mass, and Center of Mass

6.4: Physics Applications - Work, Force, and Pressure

6.5: Improper Integrals

6.6: Using Definite Integrals (Exercises)

This page titled 6: Using Definite Integrals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 6: Using Definite Integrals by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

6.1: Using Definite Integrals to Find Area and Length

Motivating Questions

- How can we use definite integrals to measure the area between two curves?
- How do we decide whether to integrate with respect to x or with respect to y when we try to find the area of a region?
- How can a definite integral be used to measure the length of a curve?

Early on in our work with the definite integral, we learned that for an object moving along an axis, the area under a non-negative velocity function v between a and b tells us the distance the object traveled on that time interval, and that area is given precisely by the definite integral $\int_a^b v(t)\ dt$. In general, for any nonnegative function f on an interval $[a, b]$, $\int_a^b f(x)\ dx$ measures the area bounded by the curve and the x-axis between $x = a$ and $x = b$.

Next, we will explore how definite integrals can be used to represent other physically important properties. In Preview Activity $6.1.1$, we investigate how a single definite integral may be used to represent the area between two curves.

Preview Activity $6.1.1$

Consider the functions given by $f(x) = 5 - (x - 1)^2$ and $g(x) = 4 - x$.

1. Use algebra to find the points where the graphs of f and g intersect.
2. Sketch an accurate graph of f and g on the axes provided, labeling the curves by name and the intersection points with ordered pairs.
3. Find and evaluate exactly an integral expression that represents the area between $y = f(x)$ and the x-axis on the interval between the intersection points of f and g.
4. Find and evaluate exactly an integral expression that represents the area between $y = g(x)$ and the x-axis on the interval between the intersection points of f and g.
5. What is the exact area between f and g between their intersection points? Why?

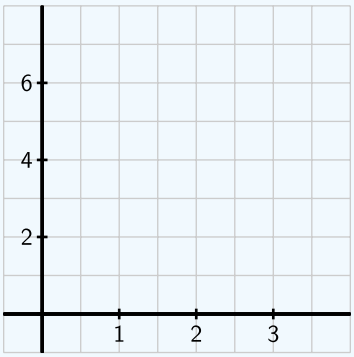


Figure $6.1.1$. Axes for plotting f and g in Preview Activity $6.1.1$.

The Area Between Two Curves

In Preview Activity $6.1.1$, we saw a natural way to think about the area between two curves: it is the area beneath the upper curve minus the area below the lower curve.

Example 6.1.1

Find the area bounded between the graphs of $f(x) = (x-1)^2 + 1$ and $g(x) = x + 2$.

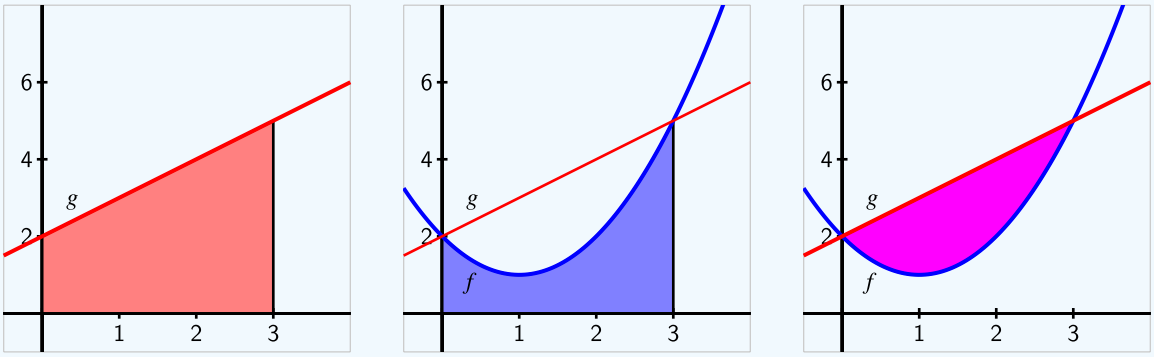


Figure 6.1.3. The areas bounded by the functions $f(x) = (x-1)^2 + 1$ and $g(x) = x + 2$ on the interval $[0, 3]$.

Answer available at the end of this section.

We can also think of the area this way: if we slice up the region between two curves into thin vertical rectangles (in the same spirit as we originally sliced the region between a single curve and the x-axis in Section 4.2), we see (as shown in Figure 6.1.4) that the height of a typical rectangle is given by the difference between the two functions, $g(x) - f(x)$, and its width is Δx. Thus the area of the rectangle is

$$A_{\text{rect}} = (g(x) - f(x))\Delta x.$$

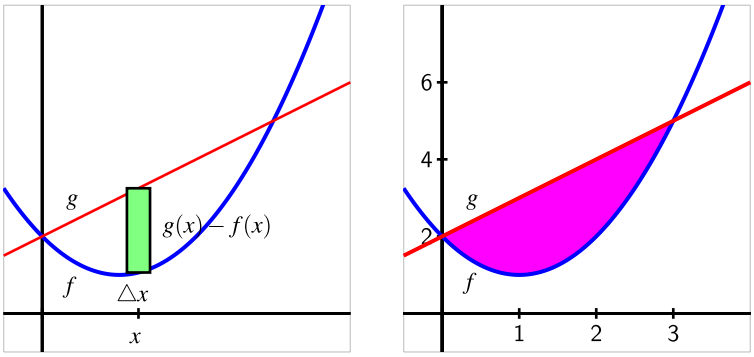


Figure 6.1.4. The area bounded by the functions $f(x) = (x-1)^2 + 1$ and $g(x) = x + 2$ on the interval $[0, 3]$.

The area between the two curves on $[0, 3]$ is thus approximated by the Riemann sum

$$A \approx \sum_{i=1}^{n} (g(x_i) - f(x_i))\Delta x,$$

and as we let $n \to \infty$, it follows that the area is given by the single definite integral

$$A = \int_0^3 (g(x) - f(x))\, dx. \tag{6.1.1}$$

In many applications of the definite integral, we will find it helpful to think of a "representative slice" and use the definite integral to add these slices. Here, the integral sums the areas of thin rectangles.

Finally, it doesn't matter whether we think of the area between two curves as the difference between the area bounded by the individual curves (as in Equation 6.1.2) or as the limit of a Riemann sum of the areas of thin rectangles between the curves (as in Equation 6.1.1). These two results are the same, since the difference of two integrals is the integral of the difference:

$$\int_0^3 g(x)\,dx - \int_0^3 f(x)\,dx = \int_0^3 (g(x) - f(x))\,dx.$$

Our work so far in this section illustrates the following general principle.

Note

If two curves $y = g(x)$ and $y = f(x)$ intersect at $(a, g(a))$ and $(b, g(b))$, and for all x such that $a \le x \le b$, $g(x) \ge f(x)$, then the area between the curves is $A = \int_a^b (g(x) - f(x))\ dx$.

Activity 6.1.2

In each of the following problems, our goal is to determine the area of the region described. For each region, (i) determine the intersection points of the curves, (ii) sketch the region whose area is being found, (iii) draw and label a representative slice, and (iv) state the area of the representative slice. Then, state a definite integral whose value is the exact area of the region, and evaluate the integral to find the numeric value of the region's area.

1. The finite region bounded by $y = \sqrt{x}$ and $y = \frac{1}{4}x$.
2. The finite region bounded by $y = 12 - 2x^2$ and $y = x^2 - 8$.
3. The area bounded by the y-axis, $f(x) = \cos(x)$, and $g(x) = \sin(x)$, where we consider the region formed by the first positive value of x for which f and g intersect.
4. The finite regions between the curves $y = x^3 - x$ and $y = x^2$.

Finding Area with Horizontal Slices

At times, the shape of a region may dictate that we use horizontal rectangular slices, instead of vertical ones.

Example 6.1.2

Find the area of the region bounded by the parabola $x = y^2 - 1$ and the line $y = x - 1$, shown at left in Figure 6.1.6.

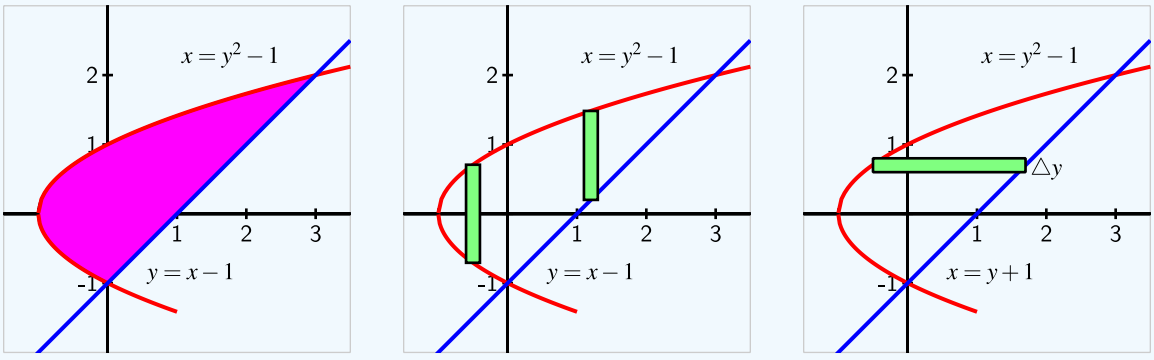


Figure 6.1.6. The area bounded by the functions $x = y^2 - 1$ and $y = x - 1$ (at left), with the region sliced vertically (center) and horizontally (at right).

Answer available at the end of this section.

Just as with the use of vertical rectangles of thickness Δx, we have a general principle for finding the area between two curves, which we state as follows.

Note

If two curves $x = g(y)$ and $x = f(y)$ intersect at $(g(c), c)$ and $(g(d), d)$, and for all y such that $c \le y \le d$, $g(y) \ge f(y)$, then the area between the curves is

$$A = \int_{y=c}^{y=d} (g(y) - f(y))\, dy.$$

Activity 6.1.3

In each of the following problems, our goal is to determine the area of the region described. For each region, (i) determine the intersection points of the curves, (ii) sketch the region whose area is being found, (iii) draw and label a representative slice, and (iv) state the area of the representative slice. Then, state a definite integral whose value is the exact area of the region, and evaluate the integral to find the numeric value of the region's area. *Note well:* At the step where you draw a representative slice, you need to make a choice about whether to slice vertically or horizontally.

1. The finite region bounded by $x = y^2$ and $x = 6 - 2y^2$.
2. The finite region bounded by $x = 1 - y^2$ and $x = 2 - 2y^2$.
3. The area bounded by the x-axis, $y = x^2$, and $y = 2 - x$.
4. The finite regions between the curves $x = y^2 - 2y$ and $y = x$.

Finding the Length of a Curve

We can also use the definite integral to find the length of a portion of a curve. We use the same fundamental principle: we slice the curve up into small pieces whose lengths we can easily approximate. Specifically, we subdivide the curve into small approximating line segments, as shown at left in Figure 6.1.7.

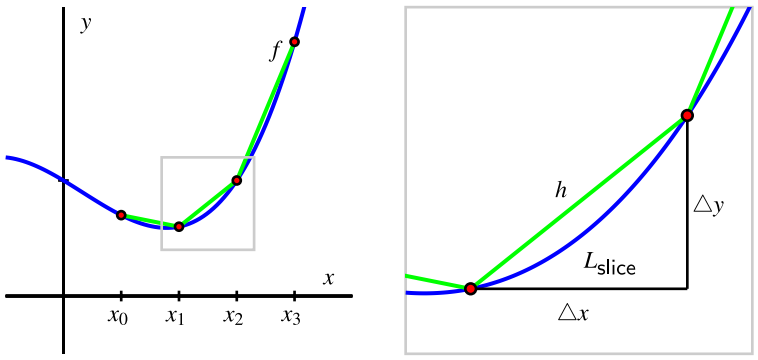


Figure 6.1.7. At left, a continuous function $y = f(x)$ whose length we seek on the interval $a = x_0$ to $b = x_3$. At right, a close up view of a portion of the curve.

We estimate the length L_{slice} of each portion of the curve on a small interval of length Δx. We use the right triangle with legs parallel to the coordinate axes and hypotenuse connecting the endpoints of the slice, as seen at right in Figure 6.1.7. The length, h, of the hypotenuse approximates the length, L_{slice}, of the curve between the two selected points. Thus,

$$L_{\text{slice}} \approx h = \sqrt{(\Delta x)^2 + (\Delta y)^2}.$$

Next we use algebra to rearrange the expression for the length of the hypotenuse into a form that we can integrate. By removing

a factor of $(\Delta x)^2$, we find

$$\begin{aligned} L_{\text{slice}} &\approx \sqrt{(\Delta x)^2 + (\Delta y)^2} \\ &= \sqrt{(\Delta x)^2 \left(1 + \frac{(\Delta y)^2}{(\Delta x)^2}\right)} \\ &= \sqrt{1 + \frac{(\Delta y)^2}{(\Delta x)^2}} \cdot \Delta x. \end{aligned}$$

Then, as $n \to \infty$ and $\Delta x \to 0$, we have that $\frac{\Delta y}{\Delta x} \to \frac{dy}{dx} = f'(x)$. Thus, we can say that

$$L_{\text{slice}} \approx \sqrt{1 + f'(x)^2} \Delta x.$$

Taking a Riemann sum of all of these slices and letting $n \to \infty$, we arrive at the following fact.

Note

Given a differentiable function f on an interval $[a, b]$, the total arc length, L, along the curve $y = f(x)$ from $x = a$ to $x = b$ is given by

$$L = \int_a^b \sqrt{1 + f'(x)^2}\, dx.$$

Activity 6.1.4

Each of the following questions somehow involves the arc length along a curve.

1. Use the definition and appropriate computational technology to determine the arc length along $y = x^2$ from $x = -1$ to $x = 1$.
2. Find the arc length of $y = \sqrt{4 - x^2}$ on the interval $-2 \le x \le 2$. Find this value in two different ways: (a) by using a definite integral, and (b) by using a familiar property of the curve.
3. Determine the arc length of $y = xe^{3x}$ on the interval $[0, 1]$.
4. Will the integrals that arise calculating arc length typically be ones that we can evaluate exactly using the First FTC, or ones that we need to approximate? Why?
5. A moving particle is traveling along the curve given by $y = f(x) = 0.1x^2 + 1$, and does so at a constant rate of 7 cm/sec, where both x and y are measured in cm (that is, the curve $y = f(x)$ is the path along which the object actually travels; the curve is not a "position function"). Find the position of the particle when $t = 4$ sec, assuming that when $t = 0$, the particle's location is $(0, f(0))$.

Summary

- To find the area between two curves, we think about slicing the region into thin rectangles. If, for instance, the area of a typical rectangle on the interval $x = a$ to $x = b$ is given by $A_{\text{rect}} = (g(x) - f(x))\Delta x$, then the exact area of the region is given by the definite integral

$$A = \int_a^b (g(x) - f(x))\, dx.$$

- The shape of the region usually dictates whether we should use vertical rectangles of thickness Δx or horizontal rectangles of thickness Δy. We want the height of the rectangle given by the difference between two curves: if those curves are best thought of as functions of y, we use horizontal rectangles, whereas if those curves are best viewed as functions of x, we use vertical rectangles.
- The arc length, L, along the curve $y = f(x)$ from $x = a$ to $x = b$ is given by

$$L = \int_a^b \sqrt{1 + f'(x)^2}\, dx.$$

Answers

Example 6.1.1

In Figure 6.1.3, we see that the graphs intersect at $(0, 2)$ and $(3, 5)$. We can find these intersection points algebraically by solving the system of equations given by $y = x + 2$ and $y = (x - 1)^2 + 1$: substituting $x + 2$ for y in the second equation yields $x + 2 = (x - 1)^2 + 1$, so $x + 2 = x^2 - 2x + 1 + 1$, and thus

$$x^2 - 3x = x(x - 3) = 0,$$

from which it follows that $x = 0$ or $x = 3$. Using $y = x + 2$, we find the corresponding y-values of the intersection points.

On the interval $[0, 3]$, the area beneath g is

$$\int_0^3 (x + 2)\, dx = \frac{21}{2},$$

while the area under f on the same interval is

$$\int_0^3 [(x - 1)^2 + 1]\, dx = 6.$$

Thus, the area between the curves is

$$A = \int_0^3 (x + 2)\, dx - \int_0^3 [(x - 1)^2 + 1]\, dx = \frac{21}{2} - 6 = \frac{9}{2}. \tag{6.1.2}$$

Example 6.1.2

By solving the second equation for x and writing $x = y + 1$, we find that $y + 1 = y^2 - 1$. Hence the curves intersect where $y^2 - y - 2 = 0$. Thus, we find $y = -1$ or $y = 2$, so the intersection points of the two curves are $(0, -1)$ and $(3, 2)$.

If we attempt to use vertical rectangles to slice up the area (as in the center graph of Figure 6.1.6), we see that from $x = -1$ to $x = 0$ the curves that bound the top and bottom of the rectangle are one and the same. This suggests, as shown in the rightmost graph in the figure, that we try using horizontal rectangles.

Note that the width of a horizontal rectangle depends on y. Between $y = -1$ and $y = 2$, the right end of a representative rectangle is determined by the line $x = y + 1$, and the left end is determined by the parabola, $x = y^2 - 1$. The thickness of the rectangle is Δy.

Therefore, the area of the rectangle is

$$A_{\text{rect}} = [(y + 1) - (y^2 - 1)]\Delta y,$$

and the area between the two curves on the y-interval $[-1, 2]$ is approximated by the Riemann sum

$$A \approx \sum_{i=1}^{n} [(y_i + 1) - (y_i^2 - 1)]\Delta y.$$

Taking the limit of the Riemann sum, it follows that the area of the region is

$$A = \int_{y=-1}^{y=2} [(y + 1) - (y^2 - 1)]\, dy. \tag{6.1.3}$$

We emphasize that we are integrating with respect to y; this is because we chose to use horizontal rectangles whose widths depend on y and whose thickness is denoted Δy. It is a straightforward exercise to evaluate the integral in Equation 6.1.3 and find that $A = \frac{9}{2}$.

This page titled 6.1: Using Definite Integrals to Find Area and Length is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 6.1: Using Definite Integrals to Find Area and Length by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

6.2: Using Definite Integrals to Find Volume

Motivating Questions

- How can we use a definite integral to find the volume of a three-dimensional solid of revolution that results from revolving a two-dimensional region about a particular axis?
- In what circumstances do we integrate with respect to y instead of integrating with respect to x?
- What adjustments do we need to make if we revolve about a line other than the x- or y-axis?

Just as we can use definite integrals to add the areas of rectangular slices to find the exact area that lies between two curves, we can also use integrals to find the volume of regions whose cross-sections have a particular shape.

In particular, we can determine the volume of solids whose cross-sections are all thin cylinders (or washers) by adding up the volumes of these individual slices. We first consider a familiar shape in Preview Activity 6.2.1: a circular cone.

Preview Activity 6.2.1

Consider a circular cone of radius 3 and height 5, which we view horizontally as pictured in Figure 6.2.1. Our goal in this activity is to use a definite integral to determine the volume of the cone.

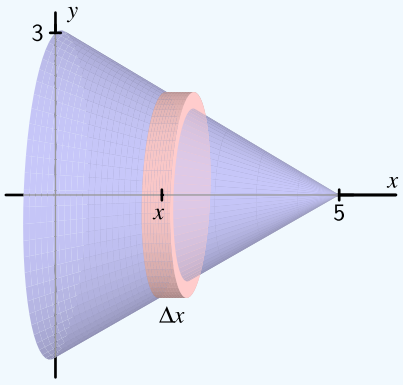


Figure 6.2.1. The circular cone described in Preview Activity 6.2.1.

1. Find a formula for the linear function $y = f(x)$ that is pictured in Figure 6.2.1.
2. For the representative slice of thickness Δx that is located horizontally at a location x (somewhere between $x = 0$ and $x = 5$), what is the radius of the representative slice? Note that the radius depends on the value of x.
3. What is the volume of the representative slice you found in (b)?
4. What definite integral will sum the volumes of the thin slices across the full horizontal span of the cone? What is the exact value of this definite integral?
5. Compare the result of your work in (d) to the volume of the cone that comes from using the formula $V_{\text{cone}} = \frac{1}{3}\pi r^2 h$.

Volume of a Solid of Revolution

A solid of revolution is a three dimensional solid that can be generated by revolving one or more curves around a fixed axis. For example, the circular cone in Figure 6.2.1 is the solid of revolution generated by revolving the portion of the line $y = 3 - \frac{3}{5}x$ from $x = 0$ to $x = 5$ about the x-axis. Notice that if we slice a solid of revolution perpendicular to the axis of revolution, the resulting cross-section is a circle.

We first consider solids whose slices are thin cylinders. Recall that the volume of a cylinder is given by $V = \pi r^2 h$.

Example 6.2.1

Find the volume of the solid of revolution generated when the region R bounded by $y = 4 - x^2$ and the x-axis is revolved about the x-axis.

Answer available at the end of this section.

For a solid such as the one in Example 6.2.1, where each slice is a cylindrical disk, we first find the volume of a typical slice (noting particularly how this volume depends on x), and then integrate over the range of x-values that bound the solid. Often, we will be content with simply finding the integral that represents the volume; if we desire a numeric value for the integral, we typically use a calculator or computer algebra system to find that value.

This method for finding the volume of a solid of revolution is often called the *disk method*.

The Disk Method

If $y = r(x)$ is a nonnegative continuous function on $[a, b]$, then the volume of the solid of revolution generated by revolving the curve about the x-axis over this interval is given by

$$V = \int_a^b \pi r(x)^2 \, dx.$$

A different type of solid can emerge when two curves are involved, as we see in the following example.

Example 6.2.2

Find the volume of the solid of revolution generated when the finite region R that lies between $y = 4 - x^2$ and $y = x + 2$ is revolved about the x-axis.

Answer available at the end of this section.

This method for finding the volume of a solid of revolution generated by two curves is often called the *washer method*.

The Washer Method

If $y = R(x)$ and $y = r(x)$ are nonnegative continuous functions on $[a, b]$ that satisfy $R(x) \geq r(x)$ for all x in $[a, b]$, then the volume of the solid of revolution generated by revolving the region between them about the x-axis over this interval is given by

$$V = \int_a^b \pi[R(x)^2 - r(x)^2] \, dx.$$

Activity 6.2.2

In each of the following questions, draw a careful, labeled sketch of the region described, as well as the resulting solid that results from revolving the region about the stated axis. In addition, draw a representative slice and state the volume of that slice, along with a definite integral whose value is the volume of the entire solid. It is not necessary to evaluate the integrals you find.

1. The region S bounded by the x-axis, the curve $y = \sqrt{x}$, and the line $x = 4$; revolve S about the x-axis.
2. The region S bounded by the y-axis, the curve $y = \sqrt{x}$, and the line $y = 2$; revolve S about the x-axis.

3. The finite region S bounded by the curves $y = \sqrt{x}$ and $y = x^3$; revolve S about the x-axis.
4. The finite region S bounded by the curves $y = 2x^2 + 1$ and $y = x^2 + 4$; revolve S about the x-axis.
5. The region S bounded by the y-axis, the curve $y = \sqrt{x}$, and the line $y = 2$; revolve S about the y-axis. How is this problem different from the one posed in part (b)?

Revolving about the y-axis

When we revolve a given region about the y-axis, the representative slices now have thickness Δy, which means that we must integrate with respect to y.

Example 6.2.3

Find the volume of the solid of revolution generated when the region R that lies between $y = \sqrt{x}$ and $y = x^4$ is revolved about the y-axis.

Answer available at the end of this section.

Activity 6.2.3

In each of the following questions, draw a careful, labeled sketch of the region described, as well as the resulting solid that results from revolving the region about the stated axis. In addition, draw a representative slice and state the volume of that slice, along with a definite integral whose value is the volume of the entire solid. It is not necessary to evaluate the integrals you find.

1. The region S bounded by the y-axis, the curve $y = \sqrt{x}$, and the line $y = 2$; revolve S about the y-axis.
2. The region S bounded by the x-axis, the curve $y = \sqrt{x}$, and the line $x = 4$; revolve S about the y-axis.
3. The finite region S in the first quadrant bounded by the curves $y = 2x$ and $y = x^3$; revolve S about the x-axis.
4. The finite region S in the first quadrant bounded by the curves $y = 2x$ and $y = x^3$; revolve S about the y-axis.
5. The finite region S bounded by the curves $x = (y - 1)^2$ and $y = x - 1$; revolve S about the y-axis

Revolving about Horizontal and Vertical Lines other than the Coordinate Axes

It is possible to revolve a region around any horizontal or vertical line. Doing so adjusts the radii of the cylinders or washers involved by a constant value. A careful, well-labeled plot of the solid of revolution will usually reveal how the different axis of revolution affects the definite integral.

Example 6.2.4

Find the volume of the solid of revolution generated when the finite region S that lies between $y = x^2$ and $y = x$ is revolved about the line $y = -1$.

Answer available at the end of this section.

Activity 6.2.4

In each of the following questions, draw a careful, labeled sketch of the region described, as well as the resulting solid that results from revolving the region about the stated axis. In addition, draw a representative slice and state the volume of that slice, along with a definite integral whose value is the volume of the entire solid. It is not necessary to evaluate the integrals you find. For each prompt, use the finite region S in the first quadrant bounded by the curves $y = 2x$ and $y = x^3$.

1. Revolve S about the line $y = -2$.
2. Revolve S about the line $y = 4$.
3. Revolve S about the line $x = -1$.
4. Revolve S about the line $x = 5$.

Summary

- We can use a definite integral to find the volume of a three-dimensional solid of revolution that results from revolving a two-dimensional region about a particular axis by taking slices perpendicular to the axis of revolution which will then be circular disks or washers.
- If we revolve about a vertical line and slice perpendicular to that line, then our slices are horizontal and of thickness Δy. This leads us to integrate with respect to y, as opposed to with respect to x when we slice a solid vertically.
- If we revolve about a line other than the x- or y-axis, we need to carefully account for the shift that occurs in the radius of a typical slice. Normally, this shift involves taking a sum or difference of the function along with the constant connected to the equation for the horizontal or vertical line; a well-labeled diagram is usually the best way to decide the new expression for the radius.

Answers

Example 6.2.1

First, we observe that $y = 4 - x^2$ intersects the x-axis at the points $(-2, 0)$ and $(2, 0)$. When we revolve the region R about the x-axis, we get the three-dimensional solid pictured in Figure 6.2.1\3.

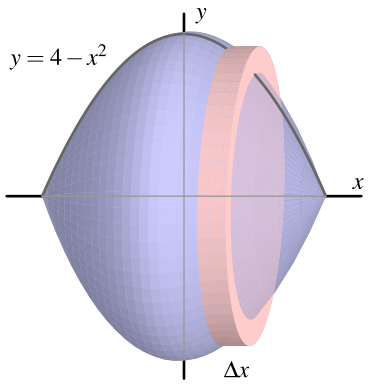


Figure 6.2.3. The solid of revolution in Example 6.2.1.

We slice the solid into vertical slices of thickness Δx between $x = -2$ and $x = 2$. A representative slice is a cylinder of height Δx and radius $4 - x^2$. Hence, the volume of the slice is

$$V_{\text{slice}} = \pi(4 - x^2)^2 \Delta x.$$

Using a definite integral to sum the volumes of the representative slices, it follows that

$$V = \int_{-2}^{2} \pi(4 - x^2)^2 \, dx.$$

It is straightforward to evaluate the integral and find that the volume is $V = \frac{512}{15}\pi$.

Example 6.2.2

First, we must determine where the curves $y = 4 - x^2$ and $y = x + 2$ intersect. Substituting the expression for y from the second equation into the first equation, we find that $x + 2 = 4 - x^2$. Rearranging, it follows that

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

and the solutions to this equation are $x = -2$ and $x = 1$. The curves therefore cross at $(-2, 0)$ and $(1, 1)$.

When we revolve the region R about the x-axis, we get the three-dimensional solid pictured at left in Figure 6.2.5.

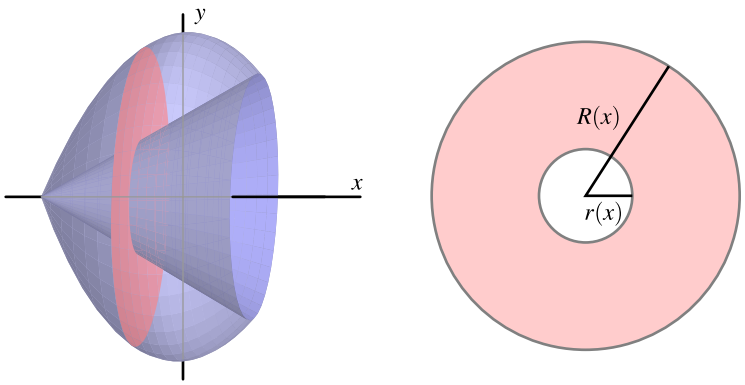


Figure 6.2.5. At left, the solid of revolution in Example 6.2.2. At right, a typical slice with inner radius $r(x)$ and outer radius $R(x)$.

Immediately we see a major difference between the solid in this example and the one in Example 6.2.1: here, the three-dimensional solid of revolution isn't "solid" because it has open space in its center along the axis of revolution. If we slice the solid perpendicular to the axis of revolution, we observe that the resulting slice is not a solid disk, but rather a *washer*, as pictured at right in Figure 6.2.5. At a given location x between $x = -2$ and $x = 1$, the small radius $r(x)$ of the inner circle is determined by the curve $y = x + 2$, so $r(x) = x + 2$. Similarly, the big radius $R(x)$ comes from the function $y = 4 - x^2$, and thus $R(x) = 4 - x^2$.

To find the volume of a representative slice, we compute the volume of the outer disk and subtract the volume of the inner disk. Since

$$\pi R(x)^2 \Delta x - \pi r(x)^2 \Delta x = \pi[R(x)^2 - r(x)^2]\Delta x,$$

it follows that the volume of a typical slice is

$$V_{\text{slice}} = \pi[(4 - x^2)^2 - (x + 2)^2]\Delta x.$$

Using a definite integral to sum the volumes of the respective slices across the integral, we find that

$$V = \int_{-2}^{1} \pi[(4 - x^2)^2 - (x + 2)^2]\, dx.$$

Evaluating the integral, we find that the volume of the solid of revolution is $V = \frac{108}{5}\pi$.

Example 6.2.3

These two curves intersect when $x = 1$, hence at the point $(1, 1)$. When we revolve the region R about the y-axis, we get the three-dimensional solid pictured at left in Figure 6.2.7.

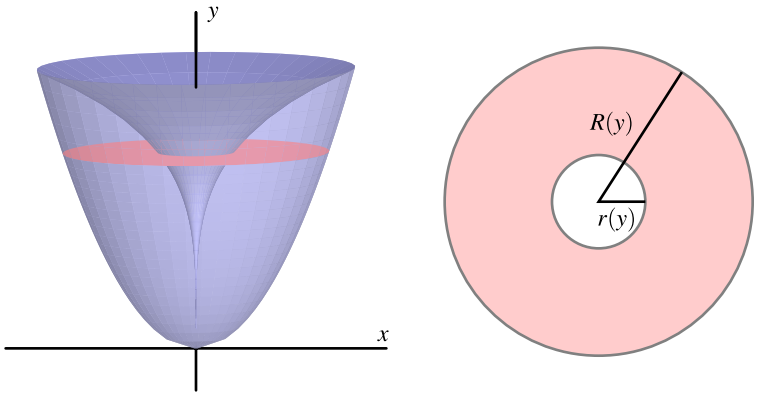


Figure 6.2.7: At left, the solid of revolution in Example 6.2.3. At right, a typical slice with inner radius $r(y)$ and outer radius $R(y)$.

Note that the slices are cylindrical washers only if taken perpendicular to the y-axis. We slice the solid horizontally, starting at $y = 0$ and proceeding up to $y = 1$. The thickness of a representative slice is Δy, so we must express the integrand in terms of y. The inner radius is determined by the curve $y = \sqrt{x}$, so we solve for x and get $x = y^2 = r(y)$. In the same way, we solve the curve $y = x^4$ (which governs the outer radius) for x in terms of y, and hence $x = \sqrt[4]{y}$. Therefore, the volume of a typical slice is

$$V_{\text{slice}} = \pi[R(y)^2 - r(y)^2] = \pi[(\sqrt[4]{y})^2 - (y^2)^2]\Delta y.$$

We use a definite integral to sum the volumes of all the slices from $y = 0$ to $y = 1$. The total volume is

$$V = \int_{y=0}^{y=1} \pi\left[(\sqrt[4]{y})^2 - (y^2)^2\right]\, dy.$$

It is straightforward to evaluate the integral and find that $V = \frac{7}{15}\pi$.

Example 6.2.4

Graphing the region between the two curves in the first quadrant between their points of intersection ($(0, 0)$ and $(1, 1)$) and then revolving the region about the line $y = -1$, we see the solid shown in Figure 6.2.9. Each slice of the solid perpendicular to the axis of revolution is a washer, and the radii of each washer are governed by the curves $y = x^2$ and $y = x$. But we also see that there is one added change: the axis of revolution adds a fixed length to each radius. The inner radius of a typical slice, $r(x)$, is given by $r(x) = x^2 + 1$, while the outer radius is $R(x) = x + 1$.

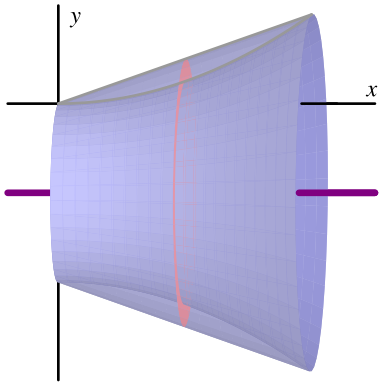


Figure 6.2.9. The solid of revolution described in Example 6.2.4.

Therefore, the volume of a typical slice is

$$V_{\text{slice}} = \pi[R(x)^2 - r(x)^2]\Delta x = \pi\left[(x+1)^2 - (x^2+1)^2\right]\Delta x.$$

Finally, we integrate to find the total volume, and

$$V = \int_0^1 \pi\left[(x+1)^2 - (x^2+1)^2\right]\,dx = \frac{7}{15}\pi.$$

This page titled 6.2: Using Definite Integrals to Find Volume is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 6.2: Using Definite Integrals to Find Volume by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

6.3: Density, Mass, and Center of Mass

Motivating Questions

- How are mass, density, and volume related?
- How is the mass of an object with varying density computed?
- What is is the center of mass of an object, and how are definite integrals used to compute it?

Studying the units on the integrand and variable of integration helps us understand the meaning of a definite integral. For instance, if $v(t)$ is the velocity of an object moving along an axis, measured in feet per second, and t measures time in seconds, then both the definite integral and its Riemann sum approximation,

$$\int_a^b v(t)\,dt \approx \sum_{i=1}^{n} v(t_i)\Delta t,$$

have units given by the product of the units of $v(t)$ and t:

$$(\text{feet/sec}) \cdot (\text{sec}) = \text{feet}.$$

Thus, $\int_a^b v(t)\ dt$ measures the total change in position of the moving object in feet.

Unit analysis will be particularly helpful to us in what follows.

Preview Activity 6.3.1

In each of the following scenarios, we consider the distribution of a quantity along an axis.

1. Suppose that the function $c(x) = 200 + 100e^{-0.1x}$ models the density of traffic on a straight road, measured in cars per mile, where x is number of miles east of a major interchange, and consider the definite integral $\int_0^2 (200 + 100e^{-0.1x})\ dx$.
   1. What are the units on the product $c(x) \cdot \Delta x$?
   2. What are the units on the definite integral and its Riemann sum approximation given by
   $$\int_0^2 c(x)\,dx \approx \sum_{i=1}^{n} c(x_i)\Delta x?$$
   3. Evaluate the definite integral $\int_0^2 c(x)\ dx = \int_0^2 (200 + 100e^{-0.1x})\ dx$ and write one sentence to explain the meaning of the value you find.
2. On a 6 foot long shelf filled with books, the function B models the distribution of the weight of the books, in pounds per inch, where x is the number of inches from the left end of the bookshelf. Let $B(x)$ be given by the rule $B(x) = 0.5 + \frac{1}{(x+1)^2}$.
   1. What are the units on the product $B(x) \cdot \Delta x$?
   2. What are the units on the definite integral and its Riemann sum approximation given by
   $$\int_{12}^{36} B(x)\,dx \approx \sum_{i=1}^{n} B(x_i)\Delta x?$$
   3. Evaluate the definite integral $\int_0^{72} B(x)\ dx = \int_0^{72} \left(0.5 + \frac{1}{(x+1)^2}\right)\ dx$ and write one sentence to explain the meaning of the value you find.

Density

The *mass* of a quantity, typically measured in metric units such as grams or kilograms, is a measure of the amount of the quantity. In a corresponding way, the *density* of an object measures the distribution of mass per unit volume. For instance, if a brick has

mass 3 kg and volume 0.002 m^3, then the density of the brick is

$$\frac{3\text{kg}}{0.002\text{m}^3} = 1500\frac{\text{kg}}{\text{m}^3}.$$

As another example, the mass density of water is 1000 kg/m^3. Each of these relationships demonstrate the following general principle.

Note

For an object of constant density d, with mass m and volume V,

$$d = \frac{m}{V}, \text{ or } m = d \cdot V.$$

But what happens when the density is not constant?

The formula $m = d \cdot V$ is reminiscent of two other equations that we have used in our work: for a body moving in a fixed direction, distance = rate $\cdot$ time, and, for a rectangle, its area is given by $A = l \cdot w$. These formulas hold when the principal quantities involved, such as the rate the body moves and the height of the rectangle, are *constant*. When these quantities are not constant, we have turned to the definite integral for assistance. By working with small slices on which the quantity of interest (such as velocity) is approximately constant, we can use a definite integral to add up the values on the pieces.

For example, if we have a nonnegative velocity function that is not constant, over a short time interval Δt we know that the distance traveled is approximately $v(t)\Delta t$, since $v(t)$ is almost constant on a small interval. Similarly, if we are thinking about the area under a nonnegative function f whose value is changing, on a short interval Δx the area under the curve is approximately the area of the rectangle whose height is $f(x)$ and whose width is Δx: $f(x)\Delta x$. Both of these principles are represented visually in Figure 6.3.1

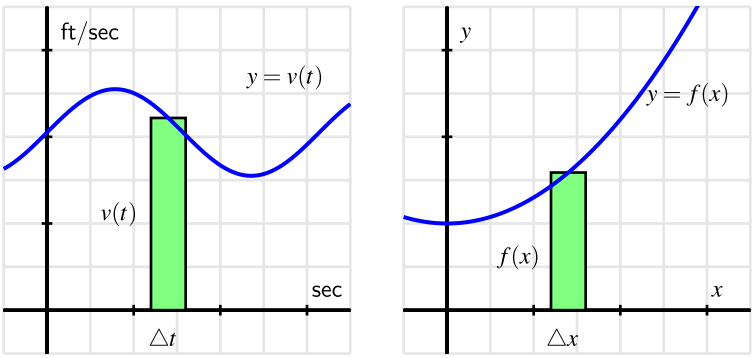


Figure 6.3.1 At left, estimating a small amount of distance traveled, $v(t)\Delta t$, and at right, a small amount of area under the curve, $f(x)\Delta x$.

In a similar way, if the density of some object is not constant, we can use a definite integral to compute the overall mass of the object. We will focus on problems where the density varies in only one dimension, say along a single axis.

Let's consider a thin bar of length b whose left end is at the origin, where $x = 0$, and assume that the bar has constant cross-sectional area of 1 cm^2. We let $\rho(x)$ represent the mass density function of the bar, measured in grams per cubic centimeter. That is, given a location x, $\rho(x)$ tells us approximately how much mass will be found in a one-centimeter wide slice of the bar at x.

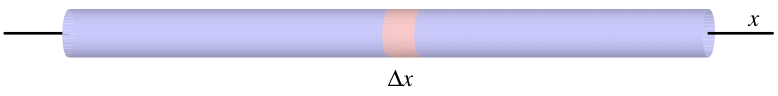


Figure 6.3.2 A thin bar of constant cross-sectional area 1 cm^2 with density function $\rho(x)$ g/cm^3.

The volume of a thin slice of the bar of width Δx, as pictured in Figure 6.3.2, is the cross-sectional area times Δx. Since the cross-sections each have constant area 1 cm^2, it follows that the volume of the slice is $1\Delta x$ cm^3. And because mass is the product of density and volume, we see that the mass of this slice is approximately

$$\text{mass}_{\text{slice}} \approx \rho(x)\ \frac{\text{g}}{\text{cm}^3} \cdot 1\Delta x\ \text{cm}^3 = \rho(x) \cdot \Delta x\ \text{g}.$$

The corresponding Riemann sum (and the integral that it approximates),

$$\sum_{i=1}^{n} \rho(x_i)\Delta x \approx \int_0^b \rho(x)\,dx,$$

therefore measure the mass of the bar between 0 and b. (The Riemann sum is an approximation, while the integral will be the exact mass.)

For objects whose cross-sectional area is constant and whose mass is distributed relative to horizontal location, x, it makes sense to think of the density function $\rho(x)$ with units "mass per unit length," such as g/cm. Thus, when we compute $\rho(x) \cdot \Delta x$ on a small slice Δx, the resulting units are g/cm · cm = g, which thus measures the mass of the slice. The general principle follows.

Note

For an object of constant cross-sectional area whose mass is distributed along a single axis according to the function $\rho(x)$ (whose units are units of mass per unit of length), the total mass, M, of the object between $x = a$ and $x = b$ is given by

$$M = \int_a^b \rho(x)\,dx.$$

Activity 6.3.2

Consider the following situations in which mass is distributed in a non-constant manner.

1. Suppose that a thin rod with constant cross-sectional area of 1 cm^2 has its mass distributed according to the density function $\rho(x) = 2e^{-0.2x}$, where x is the distance in cm from the left end of the rod, and the units on $\rho(x)$ are g/cm. If the rod is 10 cm long, determine the exact mass of the rod.
2. Consider the cone that has a base of radius 4 m and a height of 5 m. Picture the cone lying horizontally with the center of its base at the origin and think of the cone as a solid of revolution.
   1. Write and evaluate a definite integral whose value is the volume of the cone.
   2. Next, suppose that the cone has uniform density of 800 kg/m^3. What is the mass of the solid cone?
   3. Now suppose that the cone's density is not uniform, but rather that the cone is most dense at its base. In particular, assume that the density of the cone is uniform across cross sections parallel to its base, but that in each such cross section that is a distance x units from the origin, the density of the cross section is given by the function $\rho(x) = 400 + \frac{200}{1+x^2}$, measured in kg/m^3. Determine and evaluate a definite integral whose value is the mass of this cone of non-uniform density. Do so by first thinking about the mass of a given slice of the cone x units away from the base; remember that in such a slice, the density will be *essentially constant*.
3. Let a thin rod of constant cross-sectional area 1 cm^2 and length 12 cm have its mass be distributed according to the density function $\rho(x) = \frac{1}{25}(x-15)^2$, measured in g/cm. Find the exact location z at which to cut the bar so that the two pieces will each have identical mass.

Weighted Averages

The concept of an average is a natural one, and one that we have used repeatedly as part of our understanding of the meaning of the definite integral. If we have n values $a_1, a_2, \ldots, a_n$, we know that their average is given by

$$\frac{a_1 + a_2 + \cdots + a_n}{n},$$

and for a quantity being measured by a function f on an interval $[a, b]$, the average value of the quantity on $[a, b]$ is

$$\frac{1}{b-a}\int_a^b f(x)\,dx.$$

As we continue to think about problems involving the distribution of mass, it is natural to consider the idea of a *weighted* average, where certain quantities involved are counted more in the average.

A common use of weighted averages is in the computation of a student's GPA, where grades are weighted according to credit hours. Let's consider the scenario in Table 6.3.1

Table 6.3.1 *A college student's semester grades.*

class	grade	grade points	credits
chemistry	B+	3.3	5
calculus	A-	3.7	4
history	B-	2.7	3
psychology	B-	2.7	3

If all of the classes were of the same weight (i.e., the same number of credits), the student's GPA would simply be calculated by taking the average

$$\frac{3.3+3.7+2.7+2.7}{4} = 3.1.$$

But since the chemistry and calculus courses have higher weights (of 5 and 4 credits respectively), we actually compute the GPA according to the weighted average

$$\frac{3.3\cdot 5+3.7\cdot 4+2.7\cdot 3+2.7\cdot 3}{5+4+3+3} = 3.1\overline{6}.$$

The weighted average reflects the fact that chemistry and calculus, as courses with higher credits, have a greater impact on the students' grade point average. Note particularly that in the weighted average, each grade gets multiplied by its weight, and we divide by the sum of the weights.

In the following activity, we explore further how weighted averages can be used to find the balancing point of a physical system.

Activity 6.3.3

For quantities of equal weight, such as two children on a teeter-totter, the balancing point is found by taking the average of their locations. When the weights of the quantities differ, we use a weighted average of their respective locations to find the balancing point.

1. Suppose that a shelf is 6 feet long, with its left end situated at $x = 0$. If one book of weight 1 lb is placed at $x_1 = 0$, and another book of weight 1 lb is placed at $x_2 = 6$, what is the location of $\overline{x}$, the point at which the shelf would (theoretically) balance on a fulcrum?
2. Now, say that we place four books on the shelf, each weighing 1 lb: at $x_1 = 0$, at $x_2 = 2$, at $x_3 = 4$, and at $x_4 = 6$. Find $\overline{x}$, the balancing point of the shelf.
3. How does $\overline{x}$ change if we change the location of the third book? Say the locations of the 1-lb books are $x_1 = 0$, $x_2 = 2$, $x_3 = 3$, and $x_4 = 6$.
4. Next, suppose that we place four books on the shelf, but of varying weights: at $x_1 = 0$ a 2-lb book, at $x_2 = 2$ a 3-lb book, at $x_3 = 4$ a 1-lb book, and at $x_4 = 6$ a 1-lb book. Use a weighted average of the locations to find $\overline{x}$, the balancing point of the shelf. How does the balancing point in this scenario compare to that found in (b)?
5. What happens if we change the location of one of the books? Say that we keep everything the same in (d), except that $x_3 = 5$. How does $\overline{x}$ change?
6. What happens if we change the weight of one of the books? Say that we keep everything the same in (d), except that the book at $x_3 = 4$ now weighs 2 lbs. How does $\overline{x}$ change?

7. Experiment with a couple of different scenarios of your choosing where you move one of the books to the left, or you decrease the weight of one of the books.
8. Write a couple of sentences to explain how adjusting the location of one of the books or the weight of one of the books affects the location of the balancing point of the shelf. Think carefully here about how your changes should be considered relative to the location of the balancing point $\overline{x}$ of the current scenario.

Center of Mass

In Activity $6.3.3$, we saw that the balancing point of a system of point-masses[1] (such as books on a shelf) is found by taking a weighted average of their respective locations. In the activity, we were computing the *center of mass* of a system of masses distributed along an axis, which is the balancing point of the axis on which the masses rest.

In the activity, we actually used *weight* rather than *mass*. Since weight is proportional to mass, the computations for the balancing point result in the same location regardless of whether we use weight or mass. The gravitational constant is present in both the numerator and denominator of the weighted average.

Center of Mass (point-masses)

For a collection of n masses $m_1, \ldots, m_n$ that are distributed along a single axis at the locations $x_1, \ldots, x_n$, the center of mass is given by

$$\overline{x} = \frac{x_1 m_1 + x_2 m_2 + \cdots + x_n m_n}{m_1 + m_2 + \cdots + m_n}.$$

Now consider a thin bar over which density is distributed continuously. If the density is constant, it is obvious that the balancing point of the bar is its midpoint. But if density is not constant, we must compute a weighted average. Let's say that the function $\rho(x)$ tells us the density distribution along the bar, measured in g/cm. If we slice the bar into small sections, we can think of the bar as holding a collection of adjacent point-masses. The mass m_i of a slice of thickness Δx at location x_i, is $m_i \approx \rho(x_i)\Delta x$.

If we slice the bar into n pieces, we can approximate its center of mass by

$$\overline{x} \approx \frac{x_1 \cdot \rho(x_1)\Delta x + x_2 \cdot \rho(x_2)\Delta x + \cdots + x_n \cdot \rho(x_n)\Delta x}{\rho(x_1)\Delta x + \rho(x_2)\Delta x + \cdots + \rho(x_n)\Delta x}.$$

Rewriting the sums in sigma notation, we have

$$\overline{x} \approx \frac{\sum_{i=1}^{n} x_i \cdot \rho(x_i)\Delta x}{\sum_{i=1}^{n} \rho(x_i)\Delta x}. \tag{6.3.1}$$

The greater the number of slices, the more accurate our estimate of the balancing point will be. The sums in Equation $(6.3.1)$ can be viewed as Riemann sums, so in the limit as $n \to \infty$, we find that the center of mass is given by the quotient of two integrals.

Center of Mass (continuous mass distribution)

For a thin rod of density $\rho(x)$ distributed along an axis from $x = a$ to $x = b$, the center of mass of the rod is given by

$$\overline{x} = \frac{\int_a^b x\rho(x)\,dx}{\int_a^b \rho(x)\,dx}.$$

Note that the denominator of $\overline{x}$ is the mass of the bar, and that this quotient of integrals is simply the continuous version of the weighted average of locations, x, along the bar.

Activity 6.3.4

Consider a thin bar of length 20 cm whose density is distributed according to the function $\rho(x) = 4 + 0.1x$, where $x = 0$ represents the left end of the bar. Assume that ρ is measured in g/cm and x is measured in cm.

1. Find the total mass, M, of the bar.
2. Without doing any calculations, do you expect the center of mass of the bar to be equal to 10, less than 10, or greater than 10? Why?
3. Compute $\overline{x}$, the exact center of mass of the bar.
4. What is the average density of the bar?
5. Now consider a different density function, given by $p(x) = 4e^{0.020732x}$, also for a bar of length 20 cm whose left end is at $x = 0$. Plot both $\rho(x)$ and $p(x)$ on the same axes. Without doing any calculations, which bar do you expect to have the greater center of mass? Why?
6. Compute the exact center of mass of the bar described in (e) whose density function is $p(x) = 4e^{0.020732x}$. Check the result against the prediction you made in (e).

Summary

- For an object of constant density D, with volume V and mass m, we know that $m = D \cdot V$.
- If an object with constant cross-sectional area (such as a thin bar) has its density distributed along an axis according to the function $\rho(x)$, then we can find the mass of the object between $x = a$ and $x = b$ by

$$m = \int_a^b \rho(x)\,dx.$$

- For a system of point-masses distributed along an axis, say $m_1, \ldots, m_n$ at locations $x_1, \ldots, x_n$, the center of mass, $\overline{x}$, is given by the weighted average

$$\overline{x} = \frac{\sum_{i=1}^n x_i m_i}{\sum_{i=1}^n m_i}.$$

If instead we have mass continuously distributed along an axis, such as by a density function $\rho(x)$ for a thin bar of constant cross-sectional area, the center of mass of the portion of the bar between $x = a$ and $x = b$ is given by

$$\overline{x} = \frac{\int_a^b x\rho(x)\,dx}{\int_a^b \rho(x)\,dx}.$$

In each situation, $\overline{x}$ represents the balancing point of the system of masses or of the portion of the bar.

This page titled 6.3: Density, Mass, and Center of Mass is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 6.3: Density, Mass, and Center of Mass by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

6.4: Physics Applications - Work, Force, and Pressure

Motivating 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”?

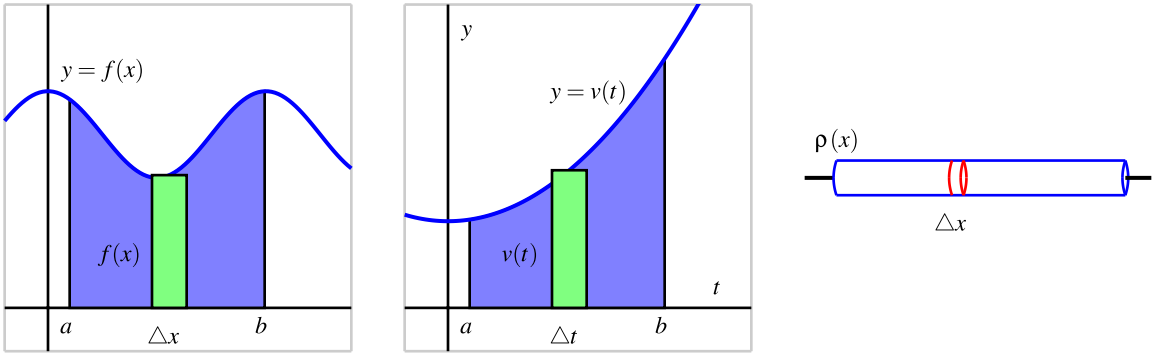


Figure 6.4.1 Three settings where we compute the accumulation of a varying quantity: the area under $y = f(x)$, the distance traveled by an object with velocity $y = v(t)$, and the mass of a bar with density function $y = \rho(x)$.

We have seen several different circumstances where the definite integral enables us to measure the accumulation of a quantity that varies, provided the quantity is approximately constant over small intervals. For instance, to find the area bounded by a nonnegative curve $y = f(x)$ and the x-axis on an interval $[a, b]$, we take a representative slice of width Δx that has area $A_{\text{slice}} = f(x)\Delta x$. As we let the width of the representative slice tend to zero, we find that the exact area of the region is

$$A = \int_a^b f(x)\,dx.$$

In a similar way, if we know the velocity $v(t)$ of a moving object 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. On a short time interval Δt, $v(t)$ is roughly constant, so for a small slice of time, $d_{\text{slice}} = v(t)\Delta t$. As the width of the time interval Δt tends to zero, the exact distance traveled is given by the definite integral

$$d = \int_a^b v(t)\,dt.$$

Finally, if we want to determine the mass of an object of non-constant density, because $M = D \cdot V$ (mass equals density times volume, provided that density is constant), we can consider a small slice of an object on which the density is approximately constant, and 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

$$M = \int_a^b \rho(x)\,dx.$$

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 that varies.

We next turn to the notion of *work*: from physics, a basic principle 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

$$W = F \cdot d = 20 \cdot 4 = 80 \text{ foot-pounds.}$$

If force and distance are measured in English units (pounds and feet), then the units of work are *foot-pounds*. If we work in metric units, where forces are measured in Newtons and distances in meters, the units of work are *Newton-meters*.

Of course, the formula $W = F \cdot d$ only applies when the force is constant over the distance d. In Preview Activity $6.4.1$, we explore one way that we can use a definite integral to compute the total work accomplished when the force exerted varies.

Preview Activity $6.4.1$

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.

1. Find a formula for $B(h)$.
2. 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 $h = 7$.
3. 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?
4. 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?
5. 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?
6. Evaluate the definite integral $\int_0^{50} B(h)\ dh$. 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 a bucket whose weight at height h is given by $B(h) = 12 + 8e^{-0.1h}$ is being lifted in a 50-foot well.

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 done in lifting the bucket. At a height h above the water, the approximate work to move the bucket a small distance Δh is

$$W_{\text{slice}} = B(h)\Delta h = (12 + 8e^{-0.1h})\Delta h.$$

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

$$W = \int_0^{50} B(h)\, dh = \int_0^{50} (12 + 8e^{-0.1h})\, dh.$$

While it 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. Here, it turns out that $W = \int_0^{50} (12 + 8e^{-0.1h})\ dh \approx 679.461$ foot-pounds.

Our work in Preview Activity $6.4.1$ and in the most recent discussion above employs the following important general principle.

Note

For an object being moved in the positive direction along an axis with location x by a force $F(x)$, the total work to move the object from a to b is given by

$$W = \int_a^b F(x)\,dx.$$

Activity 6.4.2

Consider the following situations in which a varying force accomplishes work.

1. 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.)
2. 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?
3. 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 $h = 100$?
4. 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.
    1. Use the given fact that $F(1/3) = 5$ to find the spring constant k.
    2. Find the work done to extend the spring from its natural length to 1 foot beyond its natural length.
    3. 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 6.4.2 we see a *sump crock*[1]. A sump crock provides an outlet for water that may build up beneath the basement floor to prevent flooding the basement.

Image credit to www.warreninspect.com/basement-moisture.

In the crock we see a floating pump. This pump is activated by elevation, so when the water level reaches a particular height, the pump turns on and pumps 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?



Figure 6.4.2. A sump crock.

Example 6.4.1

Suppose that a sump crock has the shape of a frustum of a cone, as pictured in Figure 6.4.3. 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. How much work is required to empty the sump crock if it is initially completely full?

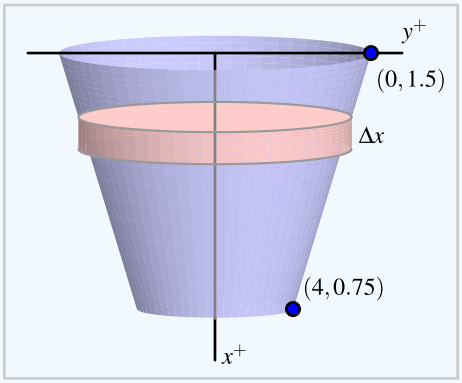


Figure 6.4.3. A sump crock with approximately cylindrical cross-sections.

Answer available at the end of this section.

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 $a = 9.81$ N/kg. That is, in metric units, the weight density of water is 9810 N/m^3.

Activity 6.4.3

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.

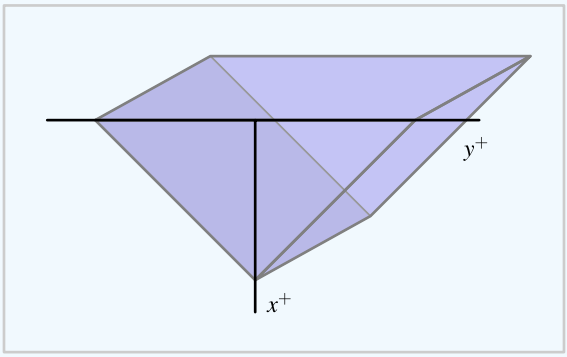


Figure 6.4.4. A trough with triangular ends, as described in Activity 6.4.3, part (c).

1. 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.
2. 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.
3. Consider a trough with triangular ends, as pictured in Figure 6.4.4, 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 building a dam, engineers need to know how much force water will exert against the face of the dam. This force comes from water 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.

We know from experience that the deeper one dives underwater while swimming, the greater the pressure exerted by the water. This is because at a greater depth, there is more water right on top of the swimmer: it is the force that "column" of water exerts that determines the pressure the swimmer experiences. 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. At a depth of d feet, a rectangular column has volume $V = 1 \cdot 1 \cdot d$ ft^3, so the corresponding weight of the water overhead is $62.4d$. This is the amount of force being exerted on a 1 square foot region at a depth d feet underwater, so the pressure exerted by water at depth d is $P = 62.4d$ (lbs/ft^2).

Because pressure is force per unit area, or $P = \frac{F}{A}$, we can compute the total force from a variable pressure by integrating $F = PA$.

Example 6.4.2

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 62.4 pounds per cubic foot. How much force does the water exert against the dam?

Answer available at the end of this section.

Activity 6.4.4

In each of the following problems, determine the total force exerted by water against the surface that is described.

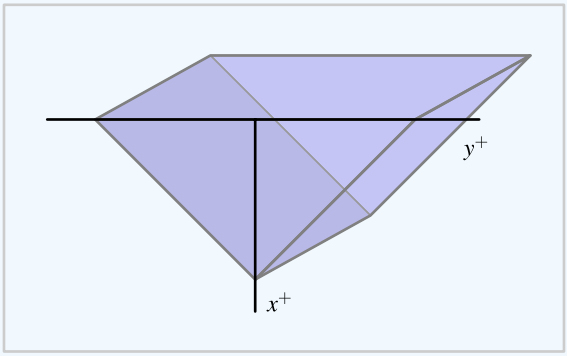


Figure 6.4.6. A trough with triangular ends, as described in Activity 6.4.3, part (c).

1. 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.
2. 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.
3. Consider a trough with triangular ends, as pictured in Figure 6.4.6, 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?

Although there are many different formulas involving work, force, and pressure, the fundamental ideas behind these problems are similar to others we've encountered in applications of the definite integral. We slice the quantity of interest into more manageable pieces and then use a definite integral to add them up.

Summary

- To measure the work done by a varying force in moving an object, we divide the problem into pieces on which we can use the formula $W = F \cdot d$, and then use a definite integral to sum the work done on each piece.
- To find the total force exerted by water against a dam, we use the formula $F = P \cdot A$ 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").

Answers

Example 6.4.1

It is helpful to think of the depth below the surface of the crock as being the independent variable, so we let the positive x-axis point down, and the positive y-axis to the right, as pictured in the figure. Because the pump sits on the surface of the water, 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.

Each slice of water is cylindrical in shape. We see that the radius of each 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 think about the problem in several steps:

1. determining the volume of a typical slice;
2. finding the weight [2] of a typical slice (and thus the force that must be exerted on it);
3. deciding the distance that a typical slice moves;

4. and 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.

We assume that the weight density of water is 62.4 pounds per cubic foot.

Consider a representative cylindrical slice at a depth of x feet below the top of the crock. 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/ft^3, the approximate weight of a representative slice is

$$F_{\text{slice}} = 62.4 \cdot V_{\text{slice}} = 62.4\pi(1.5 - 0.1875x)^2 \Delta x.$$

This is also the approximate force the pump must exert to move the slice.

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, another 9 feet to reach the drain at ground level. Hence, the total distance a representative slice travels is

$$d_{\text{slice}} = x + 9.$$

Finally, the work to move a representative slice is given by

$$W_{\text{slice}} = F_{\text{slice}} \cdot d_{\text{slice}} = 62.4\pi(1.5 - 0.1875x)^2 \Delta x \cdot (x + 9).$$

We sum the work required to move slices throughout the tank (from $x = 0$ to $x = 4$), let $\Delta x \to 0$, and hence

$$W = \int_0^4 62.4\pi(1.5 - 0.1875x)^2 (x + 9)\, dx.$$

When evaluated using appropriate technology, the integral shows that the total work is $W = 3463.2\pi$ foot-pounds.

Example 6.4.2

First, we sketch a picture of the dam, as shown in Figure 6.4.5. Note that, as in problems involving the work to pump out a tank, we let the positive x-axis point down.

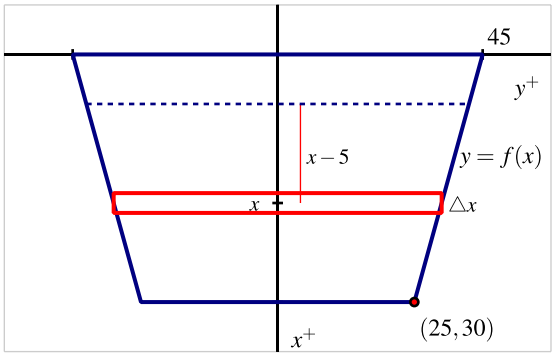


Figure 6.4.5. A trapezoidal dam that is 25 feet tall, 60 feet wide at its base, 90 feet wide at its top, with the water line 5 feet down from the top of its face.

Pressure is constant at a fixed depth, so we consider a slice of water at constant depth on the face, as shown in the figure. The area of this slice is approximately the area of the rectangle pictured. Since the width of that rectangle depends on the variable x, we find a formula for the line that represents one side of the dam. It is straightforward to find that $y = 45 - \frac{3}{5}x$. 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, so the pressure can be considered constant. 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 is given by $P = 62.4d$, we have that at any point on the 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 take slices all the way to the bottom of the dam, where $x = 25$. Hence,

$$F = \int_{x=5}^{x=25} 62.4(x - 5) \cdot 2(45 - \frac{3}{5}x)\, dx.$$

Using technology to evaluate the integral, we find $F = 848640$ pounds.

This page titled 6.4: Physics Applications - Work, Force, and Pressure is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 6.4: Physics Applications - Work, Force, and Pressure by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

6.5: Improper Integrals

Motivating Questions

- What are improper integrals and why are they important?
- What does it mean to say that an improper integral converges or diverges?
- What are some typical improper integrals that we can classify as convergent or divergent?

Another important application of the definite integral measures the likelihood of certain events. For instance, consider a company that manufactures incandescent light bulbs. Based on a large volume of test results, they have determined that the fraction of light bulbs that fail between times $t = a$ and $t = b$ of use (where t is measured in months) is given by

$$\int_a^b 0.3e^{-0.3t}\,dt.$$

For example, the fraction of light bulbs that fail during their third month of use is given by

$$\begin{aligned}\int_2^3 0.3e^{-0.3t}\,dt &= -e^{-0.3t}\Big|_2^3 \\ &= -e^{-0.9} + e^{-0.6} \\ &\approx 0.1422.\end{aligned}$$

Thus about 14.22% of all lightbulbs fail between $t = 2$ and $t = 3$. Clearly we could adjust the limits of integration to measure the fraction of light bulbs that fail during any time period of interest.

Preview Activity 6.5.1

A company with a large customer base has a call center that receives thousands of calls a day. After studying the data that represents how long callers wait for assistance, they find that the function $p(t) = 0.25e^{-0.25t}$ models the time customers wait in the following way: the fraction of customers who wait between $t = a$ and $t = b$ minutes is given by

$$\int_a^b p(t)\,dt.$$

Use this information to answer the following questions.

1. Determine the fraction of callers who wait between 5 and 10 minutes.
2. Determine the fraction of callers who wait between 10 and 20 minutes.
3. Next, let's study the fraction who wait up to a certain number of minutes:
   1. What is the fraction of callers who wait between 0 and 5 minutes?
   2. What is the fraction of callers who wait between 0 and 10 minutes?
   3. Between 0 and 15 minutes? Between 0 and 20?
4. Let $F(b)$ represent the fraction of callers who wait between 0 and b minutes. Find a formula for $F(b)$ that involves a definite integral, and then use the First FTC to find a formula for $F(b)$ that does not involve a definite integral.
5. What is the value of the limit $\lim_{b\to\infty} F(b)$? What is its meaning in the context of the problem?

Improper Integrals Involving Unbounded Intervals

In view of the above examples, we see that we may want to integrate over an interval whose upper limit grows without bound. For example, to find the fraction of light bulbs that fail *eventually*, we wish to find

$$\lim_{b\to\infty}\int_0^b 0.3e^{-0.3t}\,dt,$$

for which we will also use the notation

$$\int_0^\infty 0.3e^{-0.3t}\,dt. \tag{6.5.1}$$

Such an integral can be interpreted as the area of an unbounded region, as pictured at right in Figure 6.5.1.

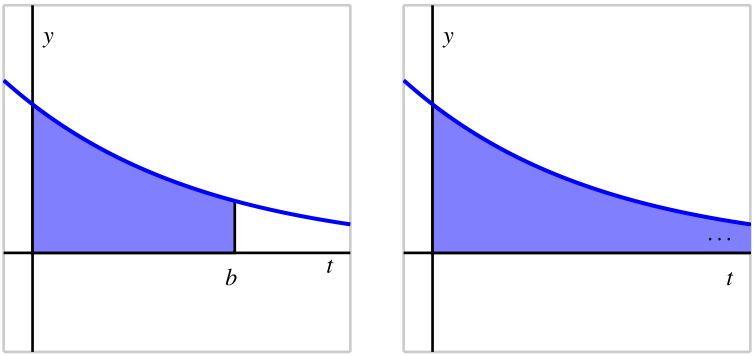


Figure 6.5.1. At left, the area bounded by $p(t) = 0.3e^{-0.3t}$ on the finite interval $[0, b]$; at right, the result of letting $b \to \infty$. By "..." in the righthand figure, we mean that the region extends to the right without bound.

We call an integral for which the interval of integration is unbounded *improper*. For instance, the integrals

$$\int_1^\infty \frac{1}{x^2}\,dx, \quad \int_{-\infty}^0 \frac{1}{1+x^2}\,dx, \quad \text{and} \int_{-\infty}^\infty e^{-x^2}\,dx$$

are all improper because they have limits of integration that involve ∞. To evaluate an improper integral we replace it with a limit of proper integrals. That is,

$$\int_0^\infty f(x)\,dx = \lim_{b\to\infty}\int_0^b f(x)\,dx.$$

We first attempt to evaluate $\int_0^b f(x)\ dx$ using the First FTC, and then evaluate the limit. Is it even possible for the area of an unbounded region to be finite? The following activity explores this issue and others in more detail.

Activity 6.5.2

In this activity we explore the improper integrals $\int_1^\infty \frac{1}{x}\ dx$ and $\int_1^\infty \frac{1}{x^{3/2}}\ dx$.

1. First we investigate $\int_1^\infty \frac{1}{x}\ dx$.
   1. Use the First FTC to determine the exact values of $\int_1^{10} \frac{1}{x}\ dx$, $\int_1^{1000} \frac{1}{x}\ dx$, and $\int_1^{100000} \frac{1}{x}\ dx$. Then, use your computational device to compute a decimal approximation of each result.
   2. Use the First FTC to evaluate the definite integral $\int_1^b \frac{1}{x}\ dx$ (which results in an expression that depends on b).
   3. Now, use your work from (ii.) to evaluate the limit given by
   $$\lim_{b\to\infty}\int_1^b \frac{1}{x}\,dx.$$
2. Next, we investigate $\int_1^\infty \frac{1}{x^{3/2}}\ dx$.
   1. Use the First FTC to determine the exact values of $\int_1^{10} \frac{1}{x^{3/2}}\ dx$, $\int_1^{1000} \frac{1}{x^{3/2}}\ dx$, and $\int_1^{100000} \frac{1}{x^{3/2}}\ dx$. Then, use your calculator to compute a decimal approximation of each result.
   2. Use the First FTC to evaluate the definite integral $\int_1^b \frac{1}{x^{3/2}}\ dx$ (which results in an expression that depends on b).

3. Now, use your work from (ii.) to evaluate the limit given by
$$\lim_{b\to\infty}\int_1^b \frac{1}{x^{3/2}}\,dx.$$

3. Plot the functions $y=\frac{1}{x}$ and $y=\frac{1}{x^{3/2}}$ on the same coordinate axes for the values $x=0\ldots 10$. How would you compare their behavior as x increases without bound? What is similar? What is different?
4. How would you characterize the value of $\int_1^\infty \frac{1}{x}\,dx$? of $\int_1^\infty \frac{1}{x^{3/2}}\,dx$? What does this tell us about the respective areas bounded by these two curves for $x\geq 1$?

Convergence and Divergence

Activity $6.5.2$ suggests that $\lim_{b\to\infty}\int_1^b f(x)\,dx$ is either finite or infinite (or it doesn't exist). With these possibilities in mind, we introduce the following terminology.

Note

If $f(x)$ is nonnegative for $x\geq a$, then we say that the improper integral $\int_a^\infty f(x)\,dx$ ***converges*** provided that
$$\lim_{b\to\infty}\int_a^b f(x)\,dx$$
exists and is finite. Otherwise, we say that $\int_a^\infty f(x)\,dx$ ***diverges***.

We will restrict our interest to improper integrals for which the integrand is nonnegative. Also, we require that $\lim_{x\to\infty} f(x)=0$, for if f does not approach 0 as $x\to\infty$, then it is impossible for $\int_a^\infty f(x)\,dx$ to converge.

Activity $6.5.3$

Determine whether each of the following improper integrals converges or diverges. For each integral that converges, find its exact value.

1. $\displaystyle\int_1^\infty \frac{1}{x^2}\,dx$
2. $\displaystyle\int_0^\infty e^{-x/4}\,dx$
3. $\displaystyle\int_2^\infty \frac{9}{(x+5)^{2/3}}\,dx$
4. $\displaystyle\int_4^\infty \frac{3}{(x+2)^{5/4}}\,dx$
5. $\displaystyle\int_0^\infty xe^{-x/4}\,dx$
6. $\int_1^\infty \frac{1}{x^p}\,dx$, where p is a positive real number

Improper Integrals Involving Unbounded Integrands

An integral is also called improper if the integrand is unbounded on the interval of integration. For example, consider
$$\int_0^1 \frac{1}{\sqrt{x}}\,dx.$$

Because $f(x)=\frac{1}{\sqrt{x}}$ has a vertical asymptote at $x=0$, f is not continuous on $[0,1]$, and the integral represents the area of the unbounded region shown at right in Figure $6.5.2$.

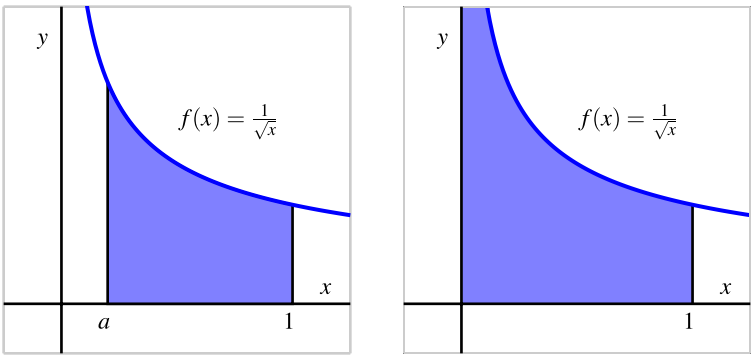


Figure 6.5.2. At left, the area bounded by $f(x) = \frac{1}{\sqrt{x}}$ on the finite interval $[a, 1]$; at right, the result of letting $a \to 0^+$, where we see that the shaded region will extend vertically without bound.

We address the problem of the integrand being unbounded by replacing the improper integral with a limit of proper integrals. For example, to evaluate $\int_0^1 \frac{1}{\sqrt{x}}\ dx$, we replace 0 with a and let a approach 0 from the right. Thus,

$$\int_0^1 \frac{1}{\sqrt{x}}\, dx = \lim_{a\to 0^+} \int_a^1 \frac{1}{\sqrt{x}}\, dx.$$

We evaluate the proper integral $\int_a^1 \frac{1}{\sqrt{x}}\ dx$, and then take the limit. We will say that the improper integral converges if this limit exists, and diverges otherwise. In this example, we observe that

$$\begin{aligned}\int_0^1 \frac{1}{\sqrt{x}}\, dx &= \lim_{a\to 0^+} \int_a^1 \frac{1}{\sqrt{x}}\, dx \\ &= \lim_{a\to 0^+} 2\sqrt{x}\,\Big|_a^1 \\ &= \lim_{a\to 0^+} 2\sqrt{1} - 2\sqrt{a} \\ &= 2,\end{aligned}$$

so the improper integral $\int_0^1 \frac{1}{\sqrt{x}}\ dx$ converges (to the value 2).

We have to be particularly careful with unbounded integrands, for they may arise in ways that may not initially be obvious. Consider, for instance, the integral

$$\int_1^3 \frac{1}{(x-2)^2}\, dx.$$

At first glance we might think that we can simply apply the Fundamental Theorem of Calculus by antidifferentiating $\frac{1}{(x-2)^2}$ to get $-\frac{1}{x-2}$ and then evaluating from 1 to 3. Were we to do so, we would be erroneously applying the FTC because $f(x) = \frac{1}{(x-2)^2}$ fails to be continuous throughout the interval, as seen in Figure 6.5.3.

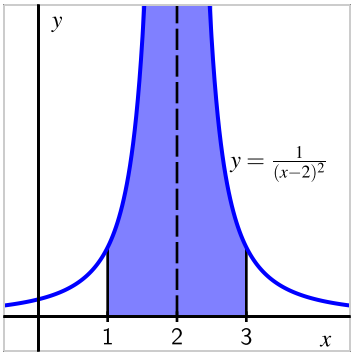


Figure 6.5.3. The function $f(x) = \frac{1}{(x-2)^2}$ on an interval including $x = 2$.

Such an incorrect application of the FTC leads to an impossible result (-2), which would itself suggest that something we did must be wrong. Instead, we must address the vertical asymptote at $x = 2$ by writing

$$\int_1^3 \frac{1}{(x-2)^2}\,dx = \lim_{a\to 2^-} \int_1^a \frac{1}{(x-2)^2}\,dx + \lim_{b\to 2^+} \int_b^3 \frac{1}{(x-2)^2}\,dx.$$

We then evaluate two separate limits of proper integrals. For instance, doing so for the integral with a approaching 2 from the left, we find

$$\begin{aligned}
\int_1^2 \frac{1}{(x-2)^2}\,dx &= \lim_{a\to 2^-} \int_1^a \frac{1}{(x-2)^2}\,dx \\
&= \lim_{a\to 2^-} -\frac{1}{(x-2)}\Big|_1^a \\
&= \lim_{a\to 2^-} -\frac{1}{(a-2)} + \frac{1}{1-2} \\
&= \infty,
\end{aligned}$$

since $\frac{1}{a-2} \to -\infty$ as a approaches 2 from the left. Thus, the improper integral $\int_1^2 \frac{1}{(x-2)^2}\,dx$ diverges; similar work shows that $\int_2^3 \frac{1}{(x-2)^2}\,dx$ also diverges. From either of these two results, we can conclude that that the original integral, $\int_1^3 \frac{1}{(x-2)^2}\,dx$ diverges, too.

Activity 6.5.4

For each of the following definite integrals, decide whether the integral is improper or not. If the integral is proper, evaluate it using the First FTC. If the integral is improper, determine whether or not the integral converges or diverges; if the integral converges, find its exact value.

1. $\int_0^1 \frac{1}{x^{1/3}}\,dx$
2. $\int_0^2 e^{-x}\,dx$
3. $\int_1^4 \frac{1}{\sqrt{4-x}}\,dx$
4. $\int_{-2}^2 \frac{1}{x^2}\,dx$
5. $\int_0^{\pi/2} \tan(x)\,dx$
6. $\int_0^1 \frac{1}{\sqrt{1-x^2}}\,dx$

Summary

- An integral $\int_a^b f(x)\ dx$ can be improper if at least one of a or b is $\pm\infty$, making the interval unbounded, or if f has a vertical asymptote at $x = c$ for some value of c that satisfies $a \leq c \leq b$. One reason that improper integrals are important is that certain probabilities can be represented by integrals that involve infinite limits.
- When we encounter an improper integral, we work to understand it by replacing the improper integral with a limit of proper integrals. For instance, we write

$$\int_a^\infty f(x)\,dx = \lim_{b\to\infty} \int_a^b f(x)\,dx,$$

  and then work to determine whether the limit exists and is finite. For any improper integral, if the resulting limit of proper integrals exists and is finite, we say the improper integral converges. Otherwise, the improper integral diverges.

- An important class of improper integrals is given by

$$\int_1^\infty \frac{1}{x^p}\,dx$$

  where p is a positive real number. We can show that this improper integral converges whenever $p > 1$, and diverges whenever $0 < p \leq 1$. A related class of improper integrals is $\int_0^1 \frac{1}{x^p}\ dx$, which converges for $0 < p < 1$, and diverges for $p \geq 1$.

This page titled 6.5: Improper Integrals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 6.5: Improper Integrals by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

6.6: Using Definite Integrals (Exercises)

6.1: Using Definite Integrals to Find Area and Length

1. Area between two power functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. Area between two trigonometric functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Area between two curves

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Arc length of a curve

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5

Find the exact area of each described region.

1. The finite region between the curves $x = y(y-2)$ and $x = -(y-1)(y-3)$.
2. The region between the sine and cosine functions on the interval $[\frac{\pi}{4}, \frac{3\pi}{4}]$.
3. The finite region between $x = y^2 - y - 2$ and $y = 2x - 1$.
4. The finite region between $y = mx$ and $y = x^2 - 1$, where m is a positive constant.

6

Let $f(x) = 1 - x^2$ and $g(x) = ax^2 - a$, where a is an unknown positive real number. For what value(s) of a is the area between the curves f and g equal to 2?

7

Let $f(x) = 2 - x^2$. Recall that the average value of any continuous function f on an interval $[a, b]$ is given by $\frac{1}{b-a} \int_a^b f(x)\ dx$.

1. Find the average value of $f(x) = 2 - x^2$ on the interval $[0, \sqrt{2}]$. Call this value r.
2. Sketch a graph of $y = f(x)$ and $y = r$. Find their intersection point(s).
3. Show that on the interval $[0, \sqrt{2}]$, the amount of area that lies below $y = f(x)$ and above $y = r$ is equal to the amount of area that lies below $y = r$ and above $y = f(x)$.
4. Will the result of (c) be true for any continuous function and its average value on any interval? Why?

6.2 Using Definite Integrals to Find Volume

1. Solid of revolution from one function about the x-axis

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Solid of revolution from one function about the y-axis

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Solid of revolution from two functions about the x-axis

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Solid of revolution from two functions about a horizontal line

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Solid of revolution from two functions about a different horizontal line

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6. Solid of revolution from two functions about a vertical line

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

7

Consider the curve $f(x) = 3\cos(\frac{x^3}{4})$ and the portion of its graph that lies in the first quadrant between the y-axis and the first positive value of x for which $f(x) = 0$. Let R denote the region bounded by this portion of f, the x-axis, and the y-axis.

1. Set up a definite integral whose value is the exact arc length of f that lies along the upper boundary of R. Use technology appropriately to evaluate the integral you find.
2. Set up a definite integral whose value is the exact area of R. Use technology appropriately to evaluate the integral you find.
3. Suppose that the region R is revolved around the x-axis. Set up a definite integral whose value is the exact volume of the solid of revolution that is generated. Use technology appropriately to evaluate the integral you find.
4. Suppose instead that R is revolved around the y-axis. If possible, set up an integral expression whose value is the exact volume of the solid of revolution and evaluate the integral using appropriate technology. If not possible, explain why.

8

Consider the curves given by $y = \sin(x)$ and $y = \cos(x)$. For each of the following problems, you should include a sketch of the region/solid being considered, as well as a labeled representative slice.

1. Sketch the region R bounded by the y-axis and the curves $y = \sin(x)$ and $y = \cos(x)$ up to the first positive value of x at which they intersect. What is the exact intersection point of the curves?
2. Set up a definite integral whose value is the exact area of R.
3. Set up a definite integral whose value is the exact volume of the solid of revolution generated by revolving R about the x-axis.

4. Set up a definite integral whose value is the exact volume of the solid of revolution generated by revolving R about the y -axis.
5. Set up a definite integral whose value is the exact volume of the solid of revolution generated by revolving R about the line $y = 2$.
6. Set up a definite integral whose value is the exact volume of the solid of revolution generated by revolving R about the line $x = -1$.

9

Consider the finite region R that is bounded by the curves $y = 1 + \frac{1}{2}(x-2)^2$, $y = \frac{1}{2}x^2$, and $x = 0$.

1. Determine a definite integral whose value is the area of the region enclosed by the two curves.
2. Find an expression involving one or more definite integrals whose value is the volume of the solid of revolution generated by revolving the region R about the line $y = -1$.
3. Determine an expression involving one or more definite integrals whose value is the volume of the solid of revolution generated by revolving the region R about the y-axis.
4. Find an expression involving one or more definite integrals whose value is the perimeter of the region R.

6.3 Density, Mass, and Center of Mass

1. Center of mass for a linear density function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Center of mass for a nonlinear density function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Interpreting the density of cars on a road

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Center of mass in a point-mass system

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5

Let a thin rod of length a have density distribution function $\rho(x) = 10e^{-0.1x}$, where x is measured in cm and ρ in grams per centimeter.

1. If the mass of the rod is 30 g, what is the value of a?
2. For the 30g rod, will the center of mass lie at its midpoint, to the left of the midpoint, or to the right of the midpoint? Why?
3. For the 30g rod, find the center of mass, and compare your prediction in (b).
4. At what value of x should the 30g rod be cut in order to form two pieces of equal mass?

6

Consider two thin bars of constant cross-sectional area, each of length 10 cm, with respective mass density functions $\rho(x) = \frac{1}{1+x^2}$ and $p(x) = e^{-0.1x}$.

1. Find the mass of each bar.
2. Find the center of mass of each bar.
3. Now consider a new 10 cm bar whose mass density function is $f(x) = \rho(x) + p(x)$.
    1. Explain how you can easily find the mass of this new bar with little to no additional work.
    2. Similarly, compute $\int_0^{10} xf(x)\ dx$ as simply as possible, in light of earlier computations.
    3. True or false: the center of mass of this new bar is the average of the centers of mass of the two earlier bars. Write at least one sentence to say why your conclusion makes sense.

7

Consider the curve given by $y = f(x) = 2xe^{-1.25x} + (30 - x)e^{-0.25(30-x)}$.

1. Plot this curve in the window $x = 0 \ldots 30$, $y = 0 \ldots 3$ (with constrained scaling so the units on the x and y axis are equal), and use it to generate a solid of revolution about the x-axis. Explain why this curve could generate a reasonable model of a baseball bat.
2. Let x and y be measured in inches. Find the total volume of the baseball bat generated by revolving the given curve about the x-axis. Include units on your answer.
3. Suppose that the baseball bat has constant weight density, and that the weight density is 0.6 ounces per cubic inch. Find the total weight of the bat whose volume you found in (b).
4. Because the baseball bat does not have constant cross-sectional area, we see that the amount of weight concentrated at a location x along the bat is determined by the volume of a slice at location x. Explain why we can think about the function $\rho(x) = 0.6\pi f(x)^2$ (where f is the function given at the start of the problem) as being the weight density function for how the weight of the baseball bat is distributed from $x = 0$ to $x = 30$.
5. Compute the center of mass of the baseball bat.

6.4 Physics Applications: Work, Force, and Pressure

Exercises 6.4.5 Exercises

1. Work to empty a conical tank

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. Work to empty a cylindrical tank

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Work to empty a rectangular pool

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Work to empty a cylindrical tank to differing heights

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5. Force due to hydrostatic pressure

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



6

Consider the curve $f(x) = 3\cos(\frac{x^3}{4})$ and the portion of its graph that lies in the first quadrant between the y-axis and the first positive value of x for which $f(x) = 0$. Let R denote the region bounded by this portion of f, the x-axis, and the y-axis. Assume that x and y are each measured in feet.

1. Picture the coordinate axes rotated 90 degrees clockwise so that the positive x-axis points straight down, and the positive y-axis points to the right. Suppose that R is rotated about the x axis to form a solid of revolution, and we consider this solid as a storage tank. Suppose that the resulting tank is filled to a depth of 1.5 feet with water weighing 62.4 pounds per cubic foot. Find the amount of work required to lower the water in the tank until it is 0.5 feet deep, by pumping the water to the top of the tank.
2. Again picture the coordinate axes rotated 90 degrees clockwise so that the positive x-axis points straight down, and the positive y-axis points to the right. Suppose that R, together with its reflection across the x-axis, forms one end of a storage tank that is 10 feet long. Suppose that the resulting tank is filled completely with water weighing 62.4 pounds per cubic foot. Find a formula for a function that tells the amount of work required to lower the water by h feet.
3. Suppose that the tank described in (b) is completely filled with water. Find the total force due to hydrostatic pressure exerted by the water on one end of the tank.

7

A cylindrical tank, buried on its side, has radius 3 feet and length 10 feet. It is filled completely with water whose weight density is 62.4 lbs/ft^3, and the top of the tank is two feet underground.

1. Set up, but do not evaluate, an integral expression that represents the amount of work required to empty the top half of the water in the tank to a truck whose tank lies 4.5 feet above ground.
2. With the tank now only half-full, set up, but do not evaluate an integral expression that represents the total force due to hydrostatic pressure against one end of the tank.

6.5: Improper Integrals

1. An improper integral on a finite interval

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. An improper integral on an infinite interval

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. An improper integral involving a ratio of exponential functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. A subtle improper integral

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. An improper integral involving a ratio of trigonometric functions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6

Determine, with justification, whether each of the following improper integrals converges or diverges.

1. $\int_e^\infty \frac{\ln(x)}{x}\ dx$
2. $\int_e^\infty \frac{1}{x\ln(x)}\ dx$
3. $\int_e^\infty \frac{1}{x(\ln(x))^2}\ dx$
4. $\int_e^\infty \frac{1}{x(\ln(x))^p}\ dx$, where p is a positive real number
5. $\int_0^1 \frac{\ln(x)}{x}\ dx$
6. $\int_0^1 \ln(x)\ dx$

7

Sometimes we may encounter an improper integral for which we cannot easily evaluate the limit of the corresponding proper integrals. For instance, consider $\int_1^\infty \frac{1}{1+x^3}\ dx$. While it is hard (or perhaps impossible) to find an antiderivative for $\frac{1}{1+x^3}$, we can still determine whether or not the improper integral converges or diverges by comparison to a simpler one. Observe that for all $x > 0$, $1 + x^3 > x^3$, and therefore

$$\frac{1}{1+x^3} < \frac{1}{x^3}.$$

It therefore follows that

$$\int_1^b \frac{1}{1+x^3}\,dx < \int_1^b \frac{1}{x^3}\,dx$$

for every $b > 1$. If we let $b \to \infty$ so as to consider the two improper integrals $\int_1^\infty \frac{1}{1+x^3}\,dx$ and $\int_1^\infty \frac{1}{x^3}\,dx$, we know that the larger of the two improper integrals converges. And thus, since the smaller one lies below a convergent integral, it follows that the smaller one must converge, too. In particular, $\int_1^\infty \frac{1}{1+x^3}\,dx$ must converge, even though we never explicitly evaluated the corresponding limit of proper integrals. We use this idea and similar ones in the exercises that follow.

1. Explain why $x^2 + x + 1 > x^2$ for all $x \geq 1$, and hence show that $\int_1^\infty \frac{1}{x^2+x+1}\,dx$ converges by comparison to $\int_1^\infty \frac{1}{x^2}\,dx$.
2. Observe that for each $x > 1$, $\ln(x) < x$. Explain why

   $$\int_2^b \frac{1}{x}\,dx < \int_2^b \frac{1}{\ln(x)}\,dx$$

   for each $b > 2$. Why must it be true that $\int_2^b \frac{1}{\ln(x)}\,dx$ diverges?
3. Explain why $\sqrt{\frac{x^4+1}{x^4}} > 1$ for all $x > 1$. Then, determine whether or not the improper integral

   $$\int_1^\infty \frac{1}{x} \cdot \sqrt{\frac{x^4+1}{x^4}}\,dx$$

   converges or diverges.

This page titled 6.6: Using Definite Integrals (Exercises) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 6.E: Using Definite Integrals (Exercises) has no license indicated.
- 6.1: Using Definite Integrals to Find Area and Length by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 6.2: Using Definite Integrals to Find Volume by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 6.3: Density, Mass, and Center of Mass by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 6.4: Physics Applications - Work, Force, and Pressure by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 6.5: Improper Integrals by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

7: DIFFERENTIAL EQUATIONS

7.1: An Introduction to Differential Equations

7.2: Qualitative Behavior of Solutions to Differential Equations

7.3: Euler's Method

7.4: Separable Differential Equations

7.5: Modeling with Differential Equations

7.6: Population Growth and the Logistic Equation

7.7: Differential Equations (Exercises)

This page titled 7: Differential Equations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 7: Differential Equations by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

7.1: An Introduction to Differential Equations

Motivating Questions

- What is a differential equation and what kinds of information can it tell us?
- How do differential equations arise in the world around us?
- What do we mean by a solution to a differential equation?

In previous chapters, we have seen that a function's derivative tells us the rate at which the function is changing. The Fundamental Theorem of Calculus helped us determine the total change of a function over an interval from the function's rate of change. For instance, an object's velocity tells us the rate of change of that object's position. By integrating the velocity over a time interval, we can determine how much the position changes over that time interval. If we know where the object is at the beginning of that interval, we have enough information to predict where it will be at the end of the interval.

In this chapter, we introduce the concept of *differential equations*. A differential equation is an equation that provides a description of a function's derivative, which means that it tells us the function's rate of change. Using this information, we would like to learn as much as possible about the function itself. Ideally we would like to have an algebraic description of the function. As we'll see, this may be too much to ask in some situations, but we will still be able to make accurate approximations.

Preview Activity 7.1.1

The position of a moving object is given by the function $s(t)$, where s is measured in feet and t in seconds. We determine that the velocity is $v(t) = 4t + 1$ feet per second.

1. How much does the position change over the time interval $[0, 4]$?
2. Does this give you enough information to determine $s(4)$, the position at time $t = 4$? If so, what is $s(4)$? If not, what additional information would you need to know to determine $s(4)$?
3. Suppose you are told that the object's initial position $s(0) = 7$. Determine $s(2)$, the object's position 2 seconds later.
4. If you are told instead that the object's initial position is $s(0) = 3$, what is $s(2)$?
5. If we only know the velocity $v(t) = 4t + 1$, is it possible that the object's position at all times is $s(t) = 2t^2 + t - 4$? Explain how you know.
6. Are there other possibilities for $s(t)$? If so, what are they?
7. If, in addition to knowing the velocity function is $v(t) = 4t + 1$, we know the initial position $s(0)$, how many possibilities are there for $s(t)$?

What is a differential equation?

A differential equation is an equation that describes the derivative, or derivatives, of a function that is unknown to us. For instance, the equation

$$\frac{dy}{dx} = x \sin x$$

describes the derivative of a function $y(x)$ that is unknown to us.

As many important examples of differential equations involve quantities that change in time, the independent variable in our discussion will frequently be time t. In the preview activity, we considered the differential equation

$$\frac{ds}{dt} = 4t + 1.$$

Knowing the velocity and the starting position of a moving object, we were able to find its position at any later time.

Because differential equations describe the derivative of a function, they give us information about how that function changes. Our goal will be to use this information to predict the value of the function in the future; in this way, differential equations provide us with something like a crystal ball.

Differential equations arise frequently in our every day world. For instance, you may hear a bank advertising:

> *Your money will grow at a 3% annual interest rate with us.*

This innocuous statement is really a differential equation. Let's translate: $A(t)$ will be amount of money you have in your account at time t. The rate at which your money grows is the derivative dA/dt, and we are told that this rate is $0.03A$. This leads to the differential equation

$$\frac{dA}{dt} = 0.03A.$$

This differential equation has a slightly different feel than the previous equation $\frac{ds}{dt} = 4t + 1$. In the earlier example, the rate of change depends only on the independent variable t, and we may find $s(t)$ by integrating the velocity $4t + 1$. In the banking example, however, the rate of change depends on the dependent variable A, so we'll need some new techniques in order to find $A(t)$.

Activity 7.1.2

Express the following statements as differential equations. In each case, you will need to introduce notation to describe the important quantities in the statement so be sure to clearly state what your notation means.

1. The population of a town grows continuously at an annual rate of 1.25%.
2. A radioactive sample loses mass at a rate of 5.6% of its mass every day.
3. You have a bank account that continuously earns 4% interest every year. At the same time, you withdraw money continually from the account at the rate of \$1000 per year.
4. A cup of hot chocolate is sitting in a 70° room. The temperature of the hot chocolate cools continuously by 10% of the difference between the hot chocolate's temperature and the room temperature every minute.
5. A can of cold soda is sitting in a 70° room. The temperature of the soda warms continuously at the rate of 10% of the difference between the soda's temperature and the room's temperature every minute.

Differential equations in the world around us

Differential equations give a natural way to describe phenomena we see in the real world. For instance, physical principles are frequently expressed as a description of how a quantity changes. A good example is Newton's Second Law, which says:

> *The product of an object's mass and acceleration equals the force applied to it.*

For instance, when gravity acts on an object near the earth's surface, it exerts a force equal to mg, the mass of the object times the gravitational constant g. We therefore have

$$ma = mg, \text{ or}$$
$$\frac{dv}{dt} = g,$$

where v is the velocity of the object, and $g = 9.8$ meters per second squared. Notice that this physical principle does not tell us what the object's velocity is, but rather how the object's velocity changes.

Activity 7.1.3

Shown below are two graphs depicting the velocity of falling objects. On the left is the velocity of a skydiver, while on the right is the velocity of a meteorite entering the Earth's atmosphere.

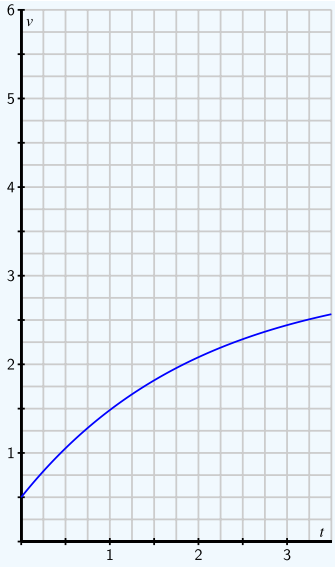


Figure 7.1.1: A skydiver's velocity.



Figure 7.1.2: A meteorite's velocity.

a. Begin with the skydiver's velocity and use the given graph to measure the rate of change dv/dt when the velocity is $v = 0.5, 1.0, 1.5, 2.0,$ and 2.5. Plot your values on the graph below. You will want to think carefully about this: you are plotting the derivative dv/dt as a function of *velocity*.

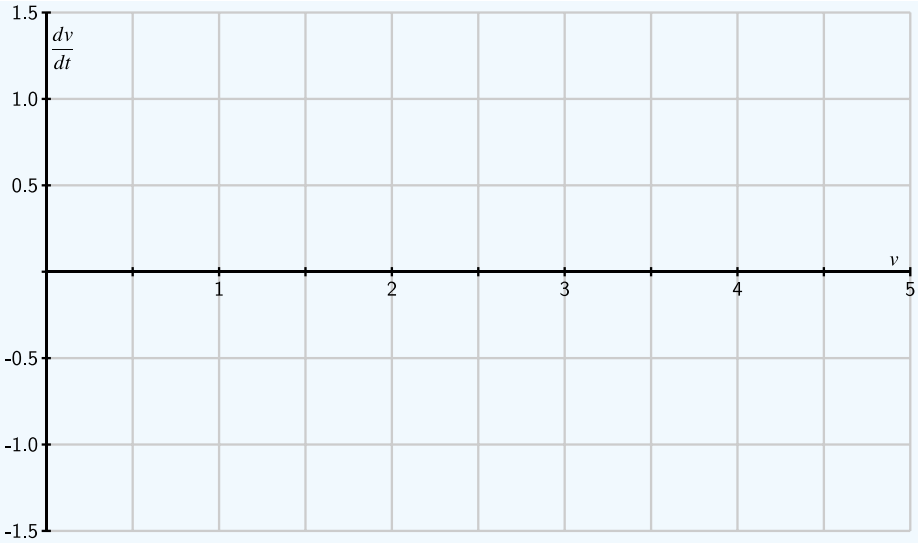


Figure 7.1.3: Plot the rates of change for parts (a) and (b) here.

b. Now do the same thing with the meteorite's velocity: use the given graph to measure the rate of change dv/dt when the velocity is $v = 3.5, 4.0, 4.5,$ and 5.0. Plot your values on the graph above.

c. You should find that all your points lie on a line. Write the equation of this line being careful to use proper notation for the quantities on the horizontal and vertical axes.

d. The relationship you just found is a differential equation. Write a complete sentence that explains its meaning.

e. By looking at the differential equation, determine the values of the velocity for which the velocity increases.

f. By looking at the differential equation, determine the values of the velocity for which the velocity decreases.

g. By looking at the differential equation, determine the values of the velocity for which the velocity remains constant.

The point of this activity is to demonstrate how differential equations model processes in the real world. In this example, two factors influence the velocities: gravity and wind resistance. The differential equation describes how these factors influence the rate of change of the velocities.

Solving a differential equation

A differential equation describes the derivative, or derivatives, of a function that is unknown to us. By a *solution* to a differential equation, we mean simply a function that satisies this description.

For instance, the first differential equation we looked at is

$$\frac{ds}{dt} = 4t + 1,$$

which describes an unknown function $s(t)$. We may check that $s(t) = 2t^2 + t$ is a solution because it satisfies this description. Notice that $s(t) = 2t^2 + t + 4$ is also a solution.

If we have a candidate for a solution, it is straightforward to check whether it is a solution or not. Before we demonstrate, however, let's consider the same issue in a simpler context. Suppose we are given the equation $2x^2 - 2x = 2x + 6$ and asked whether $x = 3$ is a solution. To answer this question, we could rewrite the variable x in the equation with the symbol $\square$:

$$2\square^2 - 2\square = 2\square + 6.$$

To determine whether $x = 3$ is a solution, we can investigate the value of each side of the equation separately when the value 3 is placed in $\square$ and see if indeed the two resulting values are equal. Doing so, we observe that

$$2\square^2 - 2\square = 2 \cdot 3^2 - 2 \cdot 3 = 12,$$

and

$$2\square + 6 = 2 \cdot 3 + 6 = 12.$$

Therefore, $x = 3$ is indeed a solution.

We will do the same thing with differential equations. Consider the differential equation

$$\frac{dv}{dt} = 1.5 - 0.5v, \text{ or}$$
$$\frac{d\square}{dt} = 1.5 - 0.5\square.$$

Let's ask whether $v(t) = 3 - 2e^{-0.5t}$ is a solution [1]. Using this formula for v, observe first that

$$\frac{dv}{dt} = \frac{d\square}{dt} = \frac{d}{dt}[3 - 2e^{-0.5t}] = -2e^{-0.5t} \cdot (-0.5) = e^{-0.5t}$$

and

$$1.5 - 0.5v = 1.5 - 0.5\square = 1.5 - 0.5(3 - 2e^{-0.5t}) = 1.5 - 1.5 + e^{-0.5t} = e^{-0.5t}.$$

At this time, don't worry about why we chose this function; we will learn techniques for finding solutions to differential equations soon enough.

Since $\frac{dv}{dt}$ and $1.5 - 0.5v$ agree for all values of t when $v = 3 - 2e^{-0.5t}$, we have indeed found a solution to the differential equation.

Activity 7.1.4

Consider the differential equation

$$\frac{dv}{dt} = 1.5 - 0.5v.$$

Which of the following functions are solutions of this differential equation?

1. $v(t) = 1.5t - 0.25t^2$.
2. $v(t) = 3 + 2e^{-0.5t}$.
3. $v(t) = 3$.
4. $v(t) = 3 + Ce^{-0.5t}$ where C is any constant.

This activity shows us something interesting. Notice that the differential equation has infinitely many solutions, which are parametrized by the constant C in $v(t) = 3 + Ce^{-0.5t}$. In Figure 7.1.4, we see the graphs of these solutions for a few values of C, as labeled.

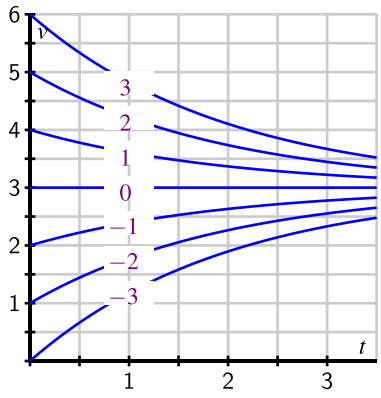


Figure 7.1.4. The family of solutions to the differential equation $\frac{dv}{dt} = 1.5 - 0.5v$.

Notice that the value of C is connected to the initial value of the velocity $v(0)$, since $v(0) = 3 + C$. In other words, while the differential equation describes how the velocity changes as a function of the velocity itself, this is not enough information to determine the velocity uniquely: we also need to know the initial velocity. For this reason, differential equations will typically have infinitely many solutions, one corresponding to each initial value. We have seen this phenomenon before: given the velocity of a moving object $v(t)$, we cannot uniquely determine the object's position function unless we also know its initial position.

If we are given a differential equation and an initial value for the unknown function, we say that we have an *initial value problem*. For instance,

$$\frac{dv}{dt} = 1.5 - 0.5v, \ v(0) = 0.5$$

is an initial value problem. In this problem, we know the value of v at one time and we know how v is changing. Consequently, there should be exactly one function v that satisfies the initial value problem.

This demonstrates the following important general property of initial value problems.

Note

Initial value problems that are “well behaved” have exactly one solution, which exists in some interval around the initial point.

We won't worry about what “well behaved” means—it is a technical condition that will be satisfied by all the differential equations we consider.

To close this section, we note that differential equations may be classified based on certain characteristics they may possess. You may see many different types of differential equations in a later course in differential equations. For now, we would like to introduce a few terms that are used to describe differential equations.

A *first-order* differential equation is one in which only the first derivative of the function occurs. For this reason,

$$\frac{dv}{dt} = 1.5 - 0.5v$$

is a first-order equation while

$$\frac{d^2y}{dt^2} = -10y$$

is a second-order equation.

A differential equation is *autonomous* if the independent variable does not appear in the description of the derivative. For instance,

$$\frac{dv}{dt} = 1.5 - 0.5v$$

is autonomous because the description of the derivative dv/dt does not depend on time. The equation

$$\frac{dy}{dt} = 1.5t - 0.5y,$$

however, is not autonomous.

Summary

- A differential equation is simply an equation that describes the derivative(s) of an unknown function.
- Physical principles, as well as some everyday situations, often describe how a quantity changes, which lead to differential equations.
- A solution to a differential equation is a function whose derivatives satisfy the equation's description. Differential equations typically have infinitely many solutions, parametrized by the initial values.

This page titled 7.1: An Introduction to Differential Equations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 7.1: An Introduction to Differential Equations by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

7.2: Qualitative Behavior of Solutions to Differential Equations

Motivating Questions

- What is a slope field?
- How can we use a slope field to obtain qualitative information about the solutions of a differential equation?
- What are stable and unstable equilibrium solutions of an autonomous differential equation?

In earlier work, we have used the tangent line to the graph of a function f at a point a to approximate the values of f near a. The usefulness of this approximation is that we need to know very little about the function; armed with only the value $f(a)$ and the derivative $f'(a)$, we may find the equation of the tangent line and the approximation

$$f(x) \approx f(a) + f'(a)(x - a).$$

Remember that a first-order differential equation gives us information about the derivative of an unknown function. Since the derivative at a point tells us the slope of the tangent line at this point, a differential equation gives us crucial information about the tangent lines to the graph of a solution. We will use this information about the tangent lines to create a *slope field* for the differential equation, which enables us to sketch solutions to initial value problems. Our aim will be to understand the solutions qualitatively. That is, we would like to understand the basic nature of solutions, such as their long-range behavior, without precisely determining the value of a solution at a particular point.

Preview Activity 7.2.1

Let's consider the initial value problem

$$\frac{dy}{dt} = t - 2, \quad y(0) = 1.$$

a. Use the differential equation to find the slope of the tangent line to the solution $y(t)$ at $t = 0$. Then use the initial value to find the equation of the tangent line at $t = 0$. Sketch this tangent line over the interval $-0.25 \leq t \leq 0.25$ on the axes provided in Figure 7.2.1.

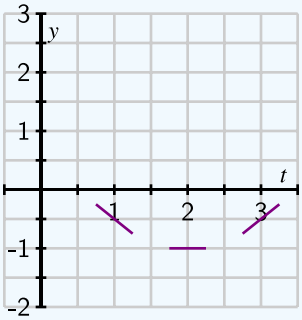


Figure 7.2.1. Grid for plotting partial tangent lines.

b. Also shown in Figure 7.2.1 are the tangent lines to the solution $y(t)$ at the points $t = 1, 2$, and 3 (we will see how to find these later). Use the graph to measure the slope of each tangent line and verify that each agrees with the value specified by the differential equation.

c. Using these tangent lines as a guide, sketch a graph of the solution $y(t)$ over the interval $0 \leq t \leq 3$ so that the lines are tangent to the graph of $y(t)$.

d. Use the Fundamental Theorem of Calculus to find $y(t)$, the solution to this initial value problem.

e. Graph the solution you found in (d) on the axes provided, and compare it to the sketch you made using the tangent lines.

Slope fields

Preview Activity $7.2.1$ shows that we can sketch the solution to an initial value problem if we know an appropriate collection of tangent lines. We can use the differential equation to find the slope of the tangent line at any point of interest, and hence plot such a collection.

Let's continue looking at the differential equation $\frac{dy}{dt} = t - 2$. If $t = 0$, this equation says that $dy/dt = 0 - 2 = -2$. Note that this value holds regardless of the value of y. We will therefore sketch tangent lines for several values of y and $t = 0$ with a slope of -2, as shown in Figure $7.2.2$.

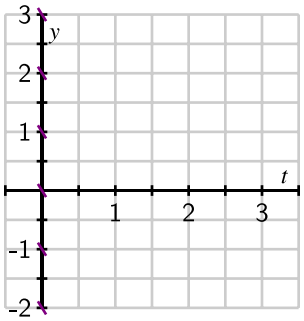


Figure $7.2.2$: Tangent lines at points with $t = 0$.

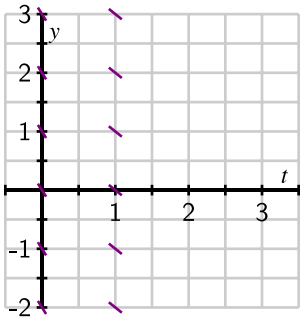


Figure $7.2.3$: Adding tangent lines at points with $t = 1$.

Let's continue in the same way: if $t = 1$, the differential equation tells us that $dy/dt = 1 - 2 = -1$, and this holds regardless of the value of y. We now sketch tangent lines for several values of y and $t = 1$ with a slope of -1 in Figure $7.2.3$.

Similarly, we see that when $t = 2$, $dy/dt = 0$ and when $t = 3$, $dy/dt = 1$. We may therefore add to our growing collection of tangent line plots to achieve Figure $7.2.4$.

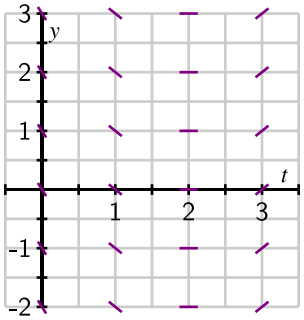


Figure $7.2.4$: Adding tangent lines at points with $t = 2$ and $t = 3$.

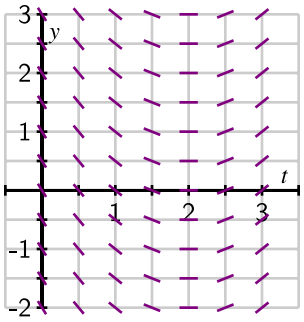


Figure 7.2.5: A completed slope field.

In Figure 7.2.4, we begin to see the solutions to the differential equation emerge. For the sake of even greater clarity, we add more tangent lines to provide the more complete picture shown at right in Figure 7.2.5.

Figure 7.2.5 is called a *slope field* for the differential equation. It allows us to sketch solutions just as we did in the preview activity. We can begin with the initial value $y(0) = 1$ and start sketching the solution by following the tangent line. Whenever the solution passes through a point at which a tangent line is drawn, that line is tangent to the solution. This principle leads us to the sequence of images in Figure 7.2.6.

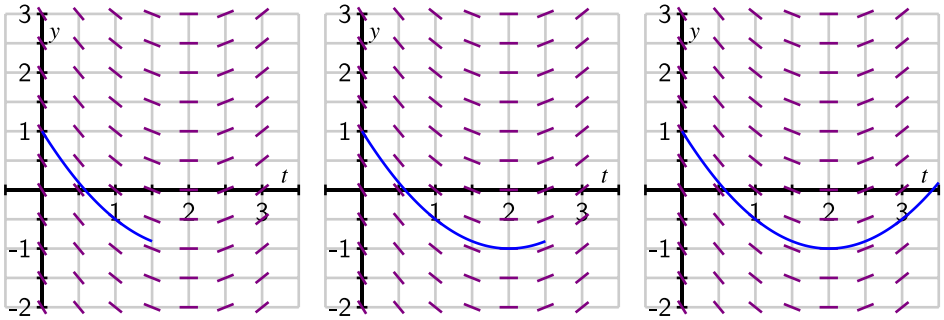


Figure 7.2.6. A sequence of images that show how to sketch the IVP solution that satisfies $y(0) = 1$.

In fact, we can draw solutions for any initial value. Figure 7.2.7 shows solutions for several different initial values for $y(0)$.

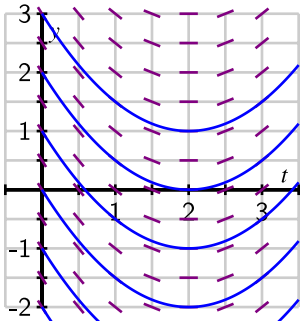


Figure 7.2.7. Different solutions to $\frac{dy}{dt} = t - 2$ that correspond to different initial values.

Just as we did for the equation $\frac{dy}{dt} = t - 2$, we can construct a slope field for any differential equation of interest. The slope field provides us with visual information about how we expect solutions to the differential equation to behave.

Activity 7.2.2

Consider the autonomous differential equation

$$\frac{dy}{dt} = -\frac{1}{2}(y-4).$$

1. Make a plot of $\frac{dy}{dt}$ versus y on the axes provided in Figure 7.2.8. Looking at the graph, for what values of y does y increase and for what values of y does y decrease?

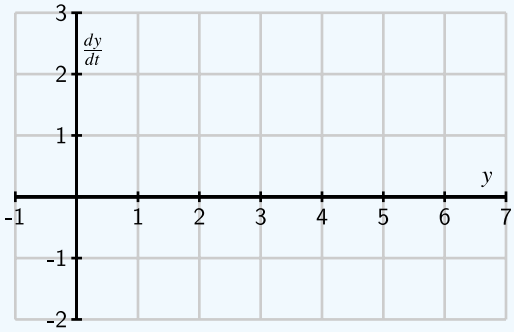


Figure 7.2.8: Axes for plotting $\frac{dy}{dt}$ versus y.

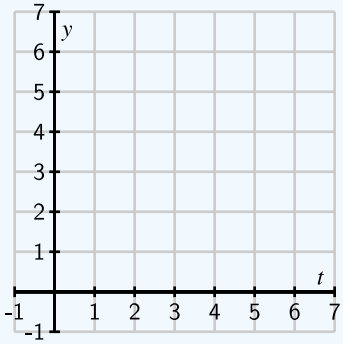


Figure 7.2.9: Axes for plotting the slope field for $\frac{dy}{dt} = -\frac{1}{2}(y-4)$.

2. Next, sketch the slope field for this differential equation on the axes provided in Figure 7.2.9.
3. Use your work in (b) to sketch (on the same axes in Figure 7.2.19.) solutions that satisfy $y(0) = 0$, $y(0) = 2$, $y(0) = 4$ and $y(0) = 6$.
4. Verify that $y(t) = 4 + 2e^{-t/2}$ is a solution to the given differential equation with the initial value $y(0) = 6$. Compare its graph to the one you sketched in (c).
5. What is special about the solution where $y(0) = 4$?

Equilibrium solutions and stability

As our work in Activity 7.2.2 demonstrates, first-order autonomous equations may have solutions that are constant. These are simple to detect by inspecting the differential equation $dy/dt = f(y)$: constant solutions necessarily have a zero derivative, so $dy/dt = 0 = f(y)$.

For example, in Activity 7.2.2, we considered the equation $\frac{dy}{dt} = f(y) = -\frac{1}{2}(y-4)$. Constant solutions are found by setting $f(y) = -\frac{1}{2}(y-4) = 0$, which we immediately see implies that $y = 4$.

Values of y for which $f(y) = 0$ in an autonomous differential equation $\frac{dy}{dt} = f(y)$ are called *equilibrium solutions* of the differential equation.

Activity 7.2.3

Consider the autonomous differential equation

$$\frac{dy}{dt} = -\frac{1}{2}y(y-4).$$

1. Make a plot of $\frac{dy}{dt}$ versus y on the axes provided in Figure 7.2.10. Looking at the graph, for what values of y does y increase and for what values of y does y decrease?

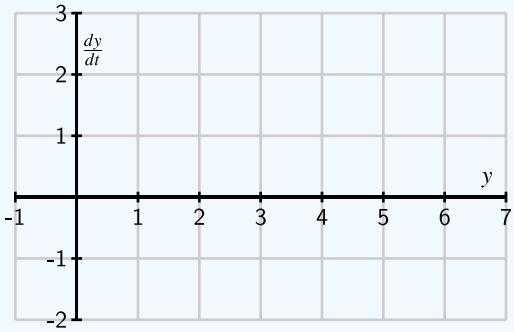


Figure 7.2.10: Axes for plotting dy/dt vs y for $\frac{dy}{dt} = -\frac{1}{2}y(y-4)$.

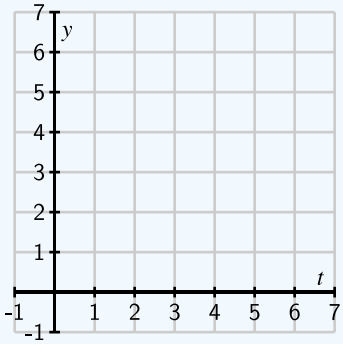


Figure 7.2.11: Axes for plotting the slope field for $\frac{dy}{dt} = -\frac{1}{2}y(y-4)$.

2. Identify any equilibrium solutions of the given differential equation.
3. Now sketch the slope field for the given differential equation on the axes provided in Figure 7.2.11.
4. Sketch the solutions to the given differential equation that correspond to initial values $y(0) = -1, 0, 1, \ldots, 5$.
5. An equilibrium solution $\overline{y}$ is called *stable* if nearby solutions converge to $\overline{y}$. This means that if the initial condition varies slightly from $\overline{y}$, then $\lim_{t\to\infty} y(t) = \overline{y}$. Conversely, an equilibrium solution $\overline{y}$ is called *unstable* if nearby solutions are pushed away from $\overline{y}$. Using your work above, classify the equilibrium solutions you found in (b) as either stable or unstable.
6. Suppose that $y(t)$ describes the population of a species of living organisms and that the initial value $y(0)$ is positive. What can you say about the eventual fate of this population?
7. Now consider a general autonomous differential equation of the form $dy/dt = f(y)$. Remember that an equilibrium solution $\overline{y}$ satisfies $f(\overline{y}) = 0$. If we graph $dy/dt = f(y)$ as a function of y, for which of the differential equations represented in Figure 7.2.12 and Figure 7.2.13 is $\overline{y}$ a stable equilibrium and for which is $\overline{y}$ unstable? Why?

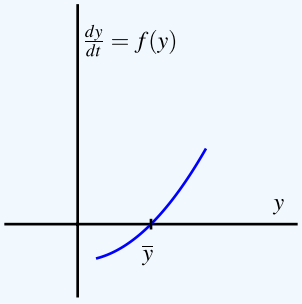


Figure 7.2.12: Plot of $\frac{dy}{dt}$ as a function of y.

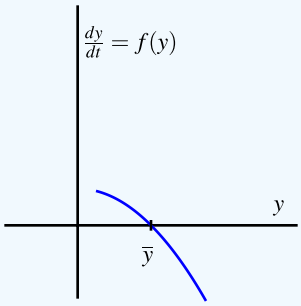


Figure 7.2.13: Plot of $\frac{dy}{dt}$ as a different function of y.

Summary

- A slope field is a plot created by graphing the tangent lines of many different solutions to a differential equation.
- Once we have a slope field, we may sketch the graph of solutions by drawing a curve that is always tangent to the lines in the slope field.
- Autonomous differential equations sometimes have constant solutions that we call equilibrium solutions. These may be classified as stable or unstable, depending on the behavior of nearby solutions.

This page titled 7.2: Qualitative Behavior of Solutions to Differential Equations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 7.2: Qualitative Behavior of Solutions to Differential Equations by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

7.3: Euler's Method

Motivating Questions

- What is Euler's method and how can we use it to approximate the solution to an initial value problem?
- How accurate is Euler's method?

In Section 7.2, we saw how a slope field can be used to sketch solutions to a differential equation. In particular, the slope field is a plot of a large collection of tangent lines to a large number of solutions of the differential equation, and we sketch a single solution by simply following these tangent lines. With a little more thought, we can use this same idea to approximate numerically the solutions of a differential equation.

Preview Activity 7.3.1

Consider the initial value problem

$$\frac{dy}{dt} = \frac{1}{2}(y+1),\ y(0) = 0.$$

a. Use the differential equation to find the slope of the tangent line to the solution $y(t)$ at $t = 0$. Then use the given initial value to find the equation of the tangent line at $t = 0$.

b. Sketch the tangent line on the axes provided in Figure 7.3.1 on the interval $0 \le t \le 2$ and use it to approximate $y(2)$, the value of the solution at $t = 2$.

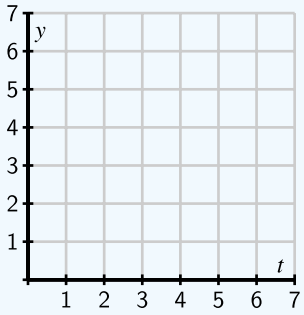


Figure 7.3.1. Grid for plotting the tangent line.

c. Assuming that your approximation for $y(2)$ is the actual value of $y(2)$, use the differential equation to find the slope of the tangent line to $y(t)$ at $t = 2$. Then, write the equation of the tangent line at $t = 2$.

d. Add a sketch of this tangent line on the interval $2 \le t \le 4$ to your plot Figure 7.3.1; use this new tangent line to approximate $y(4)$, the value of the solution at $t = 4$.

e. Repeat the same step to find an approximation for $y(6)$.

Euler's Method

Preview Activity 7.3.1 demonstrates an algorithm known as Euler's[1] Method, which generates a numerical approximation to the solution of an initial value problem. In this algorithm, we will approximate the solution by taking horizontal steps of a fixed size that we denote by Δt.

"Euler" is pronounced "Oy-ler." Among other things, Euler is the mathematician credited with the famous number e; if you incorrectly pronounce his name "You-ler," you fail to appreciate his genius and legacy.

Before explaining the algorithm in detail, let's remember how we compute the slope of a line: the slope is the ratio of the vertical change to the horizontal change, as shown in Figure 7.3.2.

In other words, $m = \frac{\Delta y}{\Delta t}$. Solving for Δy, we see that the vertical change is the product of the slope and the horizontal change, or

$$\Delta y = m\Delta t.$$

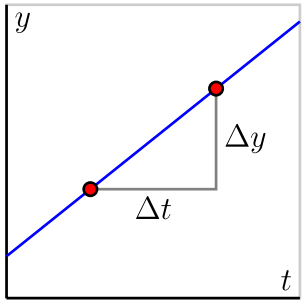


Figure 7.3.2. The role of slope in Euler's Method.

Now, suppose that we would like to solve the initial value problem

$$\frac{dy}{dt} = t - y,\ y(0) = 1.$$

There is an algorithm by which we can find an algebraic formula for the solution to this initial value problem, and we can check that this solution is $y(t) = t - 1 + 2e^{-t}$. But we are instead interested in generating an approximate solution by creating a sequence of points (t_i, y_i), where $y_i \approx y(t_i)$. For this first example, we choose $\Delta t = 0.2$.

Since we know that $y(0) = 1$, we will take the initial point to be $(t_0, y_0) = (0, 1)$ and move horizontally by $\Delta t = 0.2$ to the point (t_1, y_1). Thus, $t_1 = t_0 + \Delta t = 0.2$. Now, the differential equation tells us that the slope of the tangent line at this point is

$$m = \left.\frac{dy}{dt}\right|_{(0,1)} = 0 - 1 = -1,$$

so to move along the tangent line by taking a horizontal step of size $\Delta t = 0.2$, we must also move vertically by

$$\Delta y = m\Delta t = -1 \cdot 0.2 = -0.2.$$

We then have the approximation $y(0.2) \approx y_1 = y_0 + \Delta y = 1 - 0.2 = 0.8$. At this point, we have executed one step of Euler's method, as seen graphically in Figure 7.3.3.

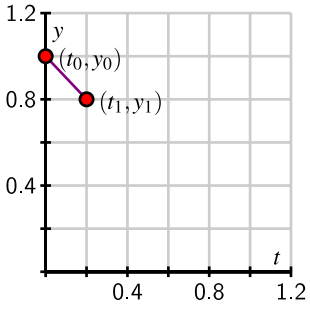


Figure 7.3.3. One step of Euler's method.

Now we repeat this process: at $(t_1, y_1) = (0.2, 0.8)$, the differential equation tells us that the slope is

$$m = \left.\frac{dy}{dt}\right|_{(0.2,0.8)} = 0.2 - 0.8 = -0.6.$$

If we move forward horizontally by Δt to $t_2 = t_1 + \Delta = 0.4$, we must move vertically by

$$\Delta y = -0.6 \cdot 0.2 = -0.12.$$

We consequently arrive at $y_2 = y_1 + \Delta y = 0.8 - 0.12 = 0.68$, which gives $y(0.2) \approx 0.68$. Now we have completed the second step of Euler's method, as shown in Figure 7.3.4.

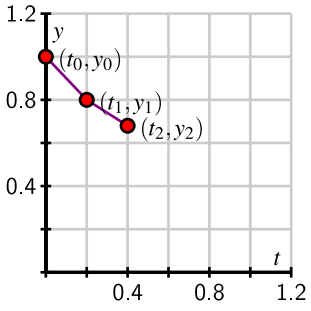


Figure 7.3.4. Two steps of Euler's method.

If we continue in this way, we may generate the points (t_i, y_i) shown in Figure 7.3.5. Because we can find a formula for the actual solution $y(t)$ to this differential equation, we can graph $y(t)$ and compare it to the points generated by Euler's method, as shown in Figure 7.3.6.

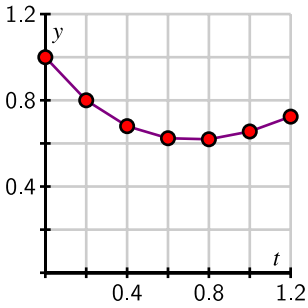


Figure 7.3.5: The points and piecewise linear approximate solution generated by Euler's method.

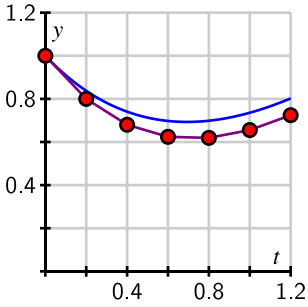


Figure 7.3.6: The approximate solution compared to the exact solution (shown in blue).

Because we need to generate a large number of points (t_i, y_i), it is convenient to organize the implementation of Euler's method in a table as shown. We begin with the given initial data.

t_i	y_i	dy/dt	Δy
0.0000	1.0000		

From here, we compute the slope of the tangent line $m = dy/dt$ using the formula for dy/dt from the differential equation, and then we find Δy, the change in y, using the rule $\Delta y = m\Delta t$.

t_i	y_i	dy/dt	Δy
0.0000	1.0000	-1.0000	-0.2000

Next, we increase t_i by Δt and y_i by Δy to get

t_i	y_i	dy/dt	Δy
0.0000	1.0000	-1.0000	-0.2000
0.2000	0.8000		

We continue the process for however many steps we decide, eventually generating a table like Table 7.3.1.

Table 7.3.1. *Euler's method for 6 steps with* $\Delta t = 0.2$.

t_i	y_i	dy/dt	Δy
0.0000	1.0000	-1.0000	-0.2000
0.2000	0.8000	-0.6000	-0.1200
0.4000	0.6800	-0.2800	-0.0560
0.6000	0.6240	-0.0240	-0.0048
0.8000	0.6192	0.1808	0.0362
1.0000	0.6554	0.3446	0.0689
1.2000	0.7243	0.4757	0.0951

Activity 7.3.2

Consider the initial value problem

$$\frac{dy}{dt} = 2t - 1, \; y(0) = 0$$

1. Use Euler's method with $\Delta t = 0.2$ to approximate the solution at $t_i = 0.2, 0.4, 0.6, 0.8,$ and 1.0. Record your work in the following table, and sketch the points (t_i, y_i) on the axes provided.

   Table 7.3.2. *Table for recording results of Euler's method.*

t_i	y_i	dy/dt	Δy
0.0000	0.0000		
0.2000			
0.4000			
0.6000			
0.8000			
1.0000			

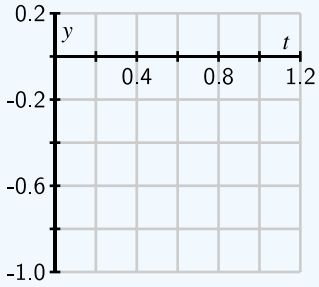


Figure 7.3.7. Grid for plotting points generated by Euler's method.

2. Find the exact solution to the original initial value problem and use this function to find the error in your approximation at each one of the points t_i.
3. Explain why the value y_5 generated by Euler's method for this initial value problem produces the same value as a left Riemann sum for the definite integral $\int_0^1 (2t-1)\,dt$.
4. How would your computations differ if the initial value was $y(0) = 1$? What does this mean about different solutions to this differential equation?

Activity 7.3.3

Consider the differential equation $\frac{dy}{dt} = 6y - y^2$.

1. Sketch the slope field for this differential equation on the axes provided in Figure 7.3.8.

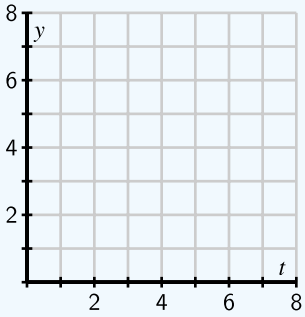


Figure 7.3.8. Grid for plotting the slope field of the given differential equation.

2. Identify any equilibrium solutions and determine whether they are stable or unstable.
3. What is the long-term behavior of the solution that satisfies the initial value $y(0) = 1$?
4. Using the initial value $y(0) = 1$, use Euler's method with $\Delta t = 0.2$ to approximate the solution at $t_i = 0.2, 0.4, 0.6, 0.8$, and 1.0. Record your results in Table 7.3.3 and sketch the corresponding points (t_i, y_i) on the axes provided in Figure 7.3.9. Note the different horizontal scale on the axes in Figure 7.3.9 compared to Figure 7.3.8.

Table 7.3.2. Table for recording results of Euler's method with $\Delta t = 0.2$.

t_i	y_i	dy/dt	Δy
0.0	1.0000		
0.2			
0.4			
0.6			
0.8			

t_i	y_i	dy/dt	Δy
1.0			

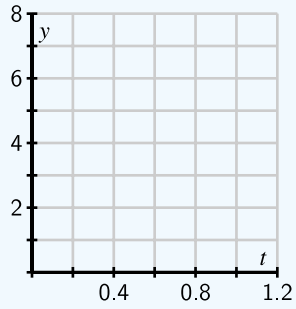


Figure 7.3.9. Axes for plotting the results of Euler's method.

5. What happens if we apply Euler's method to approximate the solution with $y(0) = 6$?

The error in Euler's method

Since we are approximating the solutions to an initial value problem using tangent lines, we should expect that the error in the approximation will be smaller when the step size is smaller. Consider the initial value problem

$$\frac{dy}{dt} = y,\ y(0) = 1,$$

whose solution we can easily find.

The question posed by this initial value problem is "what function do we know that is the same as its own derivative and has value 1 when $t = 0$?" It is not hard to see that the solution is $y(t) = e^t$. We now apply Euler's method to approximate $y(1) = e$ using several values of Δt. These approximations will be denoted by $E_{\Delta t}$, and we'll use them to see how accurate Euler's Method is.

To begin, we apply Euler's method with a step size of $\Delta t = 0.2$. In that case, we find that $y(1) \approx E_{0.2} = 2.4883$. The error is therefore

$$y(1) - E_{0.2} = e - 2.4883 \approx 0.2300.$$

Repeatedly halving Δt gives the following results, expressed in both tabular and graphical form.

Table 7.3.3. Errors that correspond to different Δt values.

Δt	$E_{\Delta t}$	Error
0.200	2.4883	0.2300
0.100	2.5937	0.1245
0.050	2.6533	0.0650
0.025	2.6851	0.0332

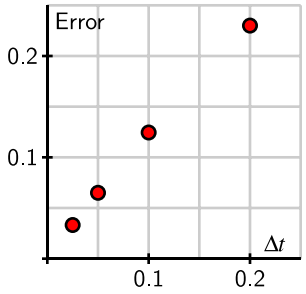


Figure 7.3.10. A plot of the error as a function of Δt.

Notice, both numerically and graphically, that the error is roughly halved when Δt is halved. This example illustrates the following general principle.

If Euler's method is used to approximate the solution to an initial value problem at a point $\bar{t}$, then the error is proportional to Δt. That is,

$$y(\bar{t}) - E_{\Delta t} \approx K\Delta t$$

for some constant of proportionality K.

Summary

- Euler's method is an algorithm for approximating the solution to an initial value problem by following the tangent lines while we take horizontal steps across the t-axis.
- If we wish to approximate $y(\bar{t})$ for some fixed $\bar{t}$ by taking horizontal steps of size Δt, then the error in our approximation is proportional to Δt.

This page titled 7.3: Euler's Method is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 7.3: Euler's Method by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

7.4: Separable Differential Equations

Motivating Questions

- What is a separable differential equation?
- How can we find solutions to a separable differential equation?
- Are some of the differential equations that arise in applications separable?

In Sections 7.2 and 7.3, we have seen several ways to approximate the solution to an initial value problem. Given the frequency with which differential equations arise in the world around us, we would like to have some techniques for finding explicit algebraic solutions of certain initial value problems. In this section, we focus on a particular class of differential equations (called *separable*) and develop a method for finding algebraic formulas for their solutions.

A *separable differential equation* is a differential equation whose algebraic structure allows the variables to be separated in a particular way. For instance, consider the equation

$$\frac{dy}{dt} = ty.$$

We would like to separate the variables t and y so that all occurrences of t appear on the right-hand side, and all occurrences of y appear on the left, multiplied by dy/dt. For this example, we divide both sides by y so that

$$\frac{1}{y}\frac{dy}{dt} = t.$$

Note that when we attempt to separate the variables in a differential equation, we require that one side is a product in which the derivative dy/dt is one factor and the other factor is solely an expression involving y.

Not every differential equation is separable. For example, if we consider the equation

$$\frac{dy}{dt} = t - y,$$

it may seem natural to separate it by writing

$$y + \frac{dy}{dt} = t.$$

As we will see, this will not be helpful, since the left-hand side is not a product of a function of y with $\frac{dy}{dt}$.

Preview Activity 7.4.1

In this preview activity, we explore whether certain differential equations are separable or not, and then revisit some key ideas from earlier work in integral calculus.

1. Which of the following differential equations are separable? If the equation is separable, write the equation in the revised form $g(y)\frac{dy}{dt} = h(t)$.
   1. $\frac{dy}{dt} = -3y.$
   2. $\frac{dy}{dt} = ty - y.$
   3. $\frac{dy}{dt} = t + 1.$
   4. $\frac{dy}{dt} = t^2 - y^2.$
2. Explain why any autonomous differential equation is guaranteed to be separable.
3. Why do we include the term "$+C$" in the expression

$$\int x\ dx = \frac{x^2}{2} + C?$$

4. Suppose we know that a certain function f satisfies the equation
$$\int f'(x)\ dx = \int x\ dx.$$
What can you conclude about f?

Solving separable differential equations

Before we discuss a general approach to solving a separable differential equation, it is instructive to consider an example.

Example 7.4.1

Find all functions y that are solutions to the differential equation
$$\frac{dy}{dt} = \frac{t}{y^2}.$$

Answer available at the end of this section.

The strategy of Example 7.4.1 may be applied to any differential equation of the form $\frac{dy}{dt} = g(y) \cdot h(t)$, and any differential equation of this form is said to be *separable*. We work to solve a separable differential equation by writing
$$\frac{1}{g(y)}\frac{dy}{dt} = h(t),$$
and then integrating both sides with respect to t. After integrating, we try to solve algebraically for y in order to write y as a function of t.

Example 7.4.2

Solve the differential equation
$$\frac{dy}{dt} = 3y.$$

Answer available at the end of this section.

There is one more technical point to make. Notice that $y = 0$ is an equilibrium solution to this differential equation. In solving the equation above, we begin by dividing both sides by y, which is not allowed if $y = 0$. To be perfectly careful, therefore, we should consider the equilibrium solutions separately. In this case, notice that the final form of our solution captures the equilibrium solution by allowing $C = 0$.

Activity 7.4.2

Suppose that the population of a town is growing continuously at an annual rate of 3% per year.

1. Let $P(t)$ be the population of the town in year t. Write a differential equation that describes the annual growth rate.
2. Find the solutions of this differential equation.
3. If you know that the town's population in year 0 is 10,000, find the population $P(t)$.
4. How long does it take for the population to double? This time is called the *doubling time*.
5. Working more generally, find the doubling time if the annual growth rate is k times the population.

Activity 7.4.3

Suppose that a cup of coffee is initially at a temperature of 105° F and is placed in a 75° F room. Newton's law of cooling says that

$$\frac{dT}{dt} = -k(T - 75),$$

where k is a constant of proportionality.

1. Suppose you measure that the coffee is cooling at one degree per minute at the time the coffee is brought into the room. Use the differential equation to determine the value of the constant k.
2. Find all the solutions of this differential equation.
3. What happens to all the solutions as $t \to \infty$? Explain how this agrees with your intuition.
4. What is the temperature of the cup of coffee after 20 minutes?
5. How long does it take for the coffee to cool to 80°?

Activity 7.4.4

Solve each of the following differential equations or initial value problems.

1. $\frac{dy}{dt} - (2 - t)y = 2 - t$
2. $\frac{1}{t}\frac{dy}{dt} = e^{t^2 - 2y}$
3. $y' = 2y + 2$, $y(0) = 2$
4. $y' = 2y^2$, $y(-1) = 2$
5. $\frac{dy}{dt} = \frac{-2ty}{t^2+1}$, $y(0) = 4$

Summary

- A separable differential equation is one that may be rewritten with all occurrences of the dependent variable multiplying the derivative and all occurrences of the independent variable on the other side of the equation.
- We may find the solutions to certain separable differential equations by separating variables, integrating with respect to t, and ultimately solving the resulting algebraic equation for y.
- This technique allows us to solve many important differential equations that arise in the world around us. For instance, questions of growth and decay and Newton's Law of Cooling give rise to separable differential equations. Later, we will learn in Section 7.6 that the important logistic differential equation is also separable.

Answers

Example 7.4.1

We begin by separating the variables and writing

$$y^2 \frac{dy}{dt} = t.$$

Integrating both sides of the equation with respect to the independent variable t shows that

$$\int y^2 \frac{dy}{dt}\, dt = \int t\, dt.$$

Next, we notice that the left-hand side allows us to change the variable of antidifferentiation[1] from t to y. In particular, $dy = \frac{dy}{dt}\, dt$, so we now have

$$\int y^2\, dy = \int t\, dt.$$

This is why we required that the left-hand side be written as a product in which dy/dt is one of the terms.

This equation says that two families of antiderivatives are equal to each other. Therefore, when we find representative antiderivatives of both sides, we know they must differ by an arbitrary constant C. Antidifferentiating and including the integration constant C on the right, we find that

$$\frac{y^3}{3} = \frac{t^2}{2} + C.$$

It is not necessary to include an arbitrary constant on both sides of the equation; we know that $y^3/3$ and $t^2/2$ are in the same family of antiderivatives and must therefore differ by a single constant.

Finally, we solve the last equation above for y as a function of t, which gives

$$y(t) = \sqrt[3]{\frac{3}{2}\,t^2 + 3C}.$$

Of course, the term $3C$ on the right-hand side represents 3 times an unknown constant. It is, therefore, still an unknown constant, which we will rewrite as C. We thus conclude that the funtion

$$y(t) = \sqrt[3]{\frac{3}{2}\,t^2 + C}$$

is a solution to the original differential equation for any value of C.

Notice that because this solution depends on the arbitrary constant C, we have found an infinite family of solutions. This makes sense because we expect to find a unique solution that corresponds to any given initial value.

For example, if we want to solve the initial value problem

$$\frac{dy}{dt} = \frac{t}{y^2},\ y(0) = 2,$$

we know that the solution has the form $y(t) = \sqrt[3]{\frac{3}{2}\,t^2 + C}$ for some constant C. We therefore must find the appropriate value for C that gives the initial value $y(0) = 2$. Hence,

$$2 = y(0) = \sqrt[3]{\frac{3}{2}\,0^2 + C} = \sqrt[3]{C},$$

which shows that $C = 2^3 = 8$. The solution to the initial value problem is then

$$y(t) = \sqrt[3]{\frac{3}{2}\,t^2 + 8}.$$

Example 7.4.2

Following the same strategy as in Example 7.4.1, we have

$$\frac{1}{y}\frac{dy}{dt} = 3.$$

Integrating both sides with respect to t,

$$\int \frac{1}{y}\frac{dy}{dt}\,dt = \int 3\,dt,$$

and thus

$$\int \frac{1}{y}\,dy = \int 3\,dt.$$

Antidifferentiating and including the integration constant, we find that

$$\ln|y| = 3t + C.$$

Finally, we need to solve for y. Here, one point deserves careful attention. By the definition of the natural logarithm function, it follows that

$$|y| = e^{3t+C} = e^{3t}e^{C}.$$

Since C is an unknown constant, e^C is as well, though we do know that it is positive (because e^x is positive for any x). When we remove the absolute value in order to solve for y, however, this constant may be either positive or negative. To account for a possible $+$ or $-$, we denote this updated constant by C to obtain

$$y(t) = Ce^{3t}.$$

This page titled 7.4: Separable Differential Equations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 7.4: Separable Differential Equations by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

7.5: Modeling with Differential Equations

Motivating Questions

- How can we use differential equations to describe phenomena in the world around us?
- How can we use differential equations to better understand these phenomena?

We have seen several ways that differential equations arise in the natural world, from the growth of a population to the temperature of a cup of coffee. In this section, we look more closely at how differential equations give us a natural way to describe various phenoma. As we'll see, the key is to understand the different factors that cause a quantity to change.

Preview Activity 7.5.1

Any time that the rate of change of a quantity is related to the amount of a quantity, a differential equation naturally arises. In the following two problems, we see two such scenarios; for each, we want to develop a differential equation whose solution is the quantity of interest.

1. Suppose you have a bank account in which money grows at an annual rate of 3%.
   1. If you have \$10,000 in the account, at what rate is your money growing?
   2. Suppose that you are also withdrawing money from the account at \$1,000 per year. What is the rate of change in the amount of money in the account? What are the units on this rate of change?
2. Suppose that a water tank holds 100 gallons and that a salty solution, which contains 20 grams of salt in every gallon, enters the tank at 2 gallons per minute.
   1. How much salt enters the tank each minute?
   2. Suppose that initially there are 300 grams of salt in the tank. How much salt is in each gallon at this point in time?
   3. Finally, suppose that evenly mixed solution is pumped out of the tank at the rate of 2 gallons per minute. How much salt leaves the tank each minute?
   4. What is the total rate of change in the amount of salt in the tank?

Developing a differential equation

Preview Activity 7.5.1 demonstrates the kind of thinking we will be doing in this section. In each of the two examples we considered, there is a quantity, such as the amount of money in the bank account or the amount of salt in the tank, that is changing due to several factors. The governing differential equation states that the total rate of change is the difference between the rate of increase and the rate of decrease.

Example 7.5.1

In the Great Lakes region, rivers flowing into the lakes carry a great deal of pollution in the form of small pieces of plastic averaging 1 millimeter in diameter. In order to understand how the amount of plastic in Lake Michigan is changing, construct a model for how this type pollution has built up in the lake.

Answer available at the end of this section.

There are many important lessons to learn from Example 7.5.1. Foremost is how we can develop a differential equation by thinking about the “total rate = rate in - rate out” model. In addition, we note how we can bring together all of our available understanding (plotting $\frac{dP}{dt}$ vs. P, creating a slope field, solving the differential equation) to see how the differential equation describes the behavior of a changing quantity.

We can also explore what happens when certain aspects of the problem change. For instance, let's suppose we are at a time when the plastic pollution entering Lake Michigan has stabilized at $1.5 \cdot 10^5$ cubic meters, and that new legislation is passed to prevent

this type of pollution entering the lake. So, there is no longer any inflow of plastic pollution to the lake. How does the amount of plastic pollution in Lake Michigan now change? For example, how long does it take for the amount of plastic pollution in the lake to halve?

Resetting $t = 0$ at this time, we now have the initial value problem

$$\frac{dP}{dt} = -\frac{1}{100}P,\ P(0) = 1.5 \cdot 10^5.$$

It is a straightforward and familiar exercise to find that the solution to this equation is $P(t) = 1.5 \cdot 10^5 e^{-t/100}$. The time that it takes for half of the pollution to flow out of the lake is given by T where $P(T) = 0.75 \cdot 10^5$. Thus, we must solve the equation

$$0.75 \cdot 10^5 = 1.5 \cdot 10^5 e^{-T/100},$$

or

$$\frac{1}{2} = e^{-T/100}.$$

It follows that

$$T = -100 \ln\left(\frac{1}{2}\right) \approx 69.3\text{years}.$$

In the upcoming activities, we explore some other natural settings in which differential equations model changing quantities.

Activity 7.5.2

Suppose you have a bank account that grows by 5% every year. Let $A(t)$ be the amount of money in the account in year t.

1. What is the rate of change of A with respect to t?
2. Suppose that you are also withdrawing \$10,000 per year. Write a differential equation that expresses the total rate of change of A.
3. Sketch a slope field for this differential equation, find any equilibrium solutions, and identify them as either stable or unstable. Write a sentence or two that describes the significance of the stability of the equilibrium solution.
4. Suppose that you initially deposit \$100,000 into the account. How long does it take for you to deplete the account?
5. What is the smallest amount of money you would need to have in the account to guarantee that you never deplete the money in the account?
6. If your initial deposit is \$300,000, how much could you withdraw every year without depleting the account?

Activity 7.5.3

A dose of morphine is absorbed from the bloodstream of a patient at a rate proportional to the amount in the bloodstream.

1. Write a differential equation for $M(t)$, the amount of morphine in the patient's bloodstream, using k as the constant proportionality.
2. Assuming that the initial dose of morphine is M_0, solve the initial value problem to find $M(t)$. Use the fact that the half-life for the absorption of morphine is two hours to find the constant k.
3. Suppose that a patient is given morphine intravenously at the rate of 3 milligrams per hour. Write a differential equation that combines the intravenous administration of morphine with the body's natural absorption.
4. Find any equilibrium solutions and determine their stability.
5. Assuming that there is initially no morphine in the patient's bloodstream, solve the initial value problem to determine $M(t)$. What happens to $M(t)$ after a very long time?
6. To what rate should a doctor reduce the intravenous rate so that there is eventually 7 milligrams of morphine in the patient's bloodstream?

Summary

- Differential equations arise in a situation when we understand how various factors cause a quantity to change.
- We may use the tools we have developed so far—slope fields, Euler's methods, and our method for solving separable equations—to understand a quantity described by a differential equation.

Answers

Example 7.5.1

First, some basic facts about Lake Michigan.

- The volume of the lake is $5 \cdot 10^{12}$ cubic meters.
- Water flows into the lake at a rate of $5 \cdot 10^{10}$ cubic meters per year. It flows out of the lake at the same rate.
- Each cubic meter flowing into the lake contains roughly $3 \cdot 10^{-8}$ cubic meters of plastic pollution.

Let's denote the amount of pollution in the lake by $P(t)$, where P is measured in cubic meters of plastic and t in years. Our goal is to describe the rate of change of this function; so we want to develop a differential equation describing $P(t)$.

First, we will measure how $P(t)$ increases due to pollution flowing into the lake. We know that $5 \cdot 10^{10}$ cubic meters of water enters the lake every year and each cubic meter of water contains $3 \cdot 10^{-8}$ cubic meters of pollution. Therefore, pollution enters the lake at the rate of

$$\left(5 \cdot 10^{10} \frac{m^3 \text{ water}}{\text{year}}\right)\left(3 \cdot 10^{-8} \frac{m^3 \text{ plastic}}{m^3 \text{ water}}\right) = 1.5 \cdot 10^3 \text{cubic m of plastic per year.}$$

Second, we will measure how $P(t)$ decreases due to pollution flowing out of the lake. If the total amount of pollution is P cubic meters and the volume of Lake Michigan is $5 \cdot 10^{12}$ cubic meters, then the concentration of plastic pollution in Lake Michigan is

$$\frac{P}{5 \cdot 10^{12}} \text{cubic m of plastic per cubic m of water.}$$

Since $5 \cdot 10^{10}$ cubic meters of water flow out each year[1], then the plastic pollution leaves the lake at the rate of

$$\left(\frac{P}{5 \cdot 10^{12}} \frac{m^3 \text{ plastic}}{m^3 \text{ water}}\right)\left(5 \cdot 10^{10} \frac{m^3 \text{ water}}{\text{year}}\right) = \frac{P}{100} \text{cubic m of plastic per year.}$$

and we assume that each cubic meter of water that flows out carries with it the plastic pollution it contains

The total rate of change of P is thus the difference between the rate at which pollution enters the lake and the rate at which pollution leaves the lake; that is,

$$\begin{aligned} \frac{dP}{dt} &= 1.5 \cdot 10^3 - \frac{P}{100} \\ &= \frac{1}{100}(1.5 \cdot 10^5 - P). \end{aligned}$$

We have now found a differential equation that describes the rate at which the amount of pollution is changing. To understand the behavior of $P(t)$, we apply some of the techniques we have recently developed.

Because the differential equation is autonomous, we can sketch dP/dt as a function of P and then construct a slope field, as shown in Figure 7.5.2 and Figure 7.5.3 .

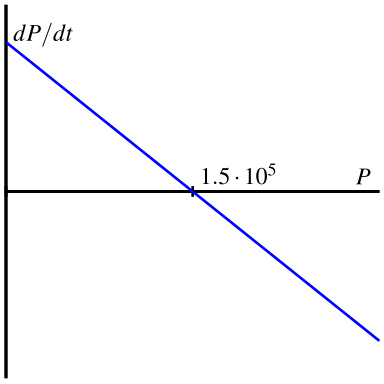


Figure 7.5.2: Plot of $\frac{dP}{dt}$ vs. P for $\frac{dP}{dt} = \frac{1}{100}(1.5 \cdot 10^5 - P)$.

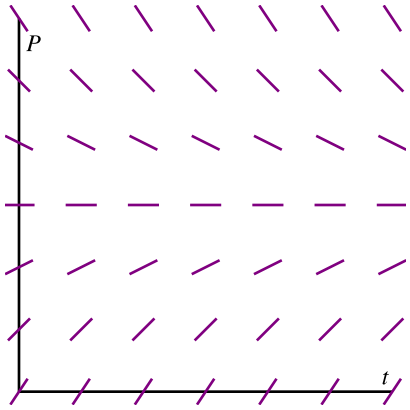


Figure 7.5.3:Plot of the slope field for $\frac{dP}{dt} = \frac{1}{100}(1.5 \cdot 10^5 - P)$.

These plots both show that $P = 1.5 \cdot 10^5$ is a stable equilibrium. Therefore, we should expect that the amount of pollution in Lake Michigan will stabilize near $1.5 \cdot 10^5$ cubic meters of pollution.

Next, assuming that there is initially no pollution in the lake, we will solve the initial value problem

$$\frac{dP}{dt} = \frac{1}{100}(1.5 \cdot 10^5 - P), \;\; P(0) = 0.$$

Separating variables, we find that

$$\frac{1}{1.5 \cdot 10^5 - P}\frac{dP}{dt} = \frac{1}{100}.$$

Integrating with respect to t, we have

$$\int \frac{1}{1.5 \cdot 10^5 - P}\frac{dP}{dt}\, dt = \int \frac{1}{100}\, dt,$$

and thus changing variables on the left and antidifferentiating on both sides, we find that

$$\int \frac{dP}{1.5 \cdot 10^5 - P} = \int \frac{1}{100}\, dt$$

$$-\ln|1.5 \cdot 10^5 - P| = \frac{1}{100}t + C$$

Finally, multiplying both sides by -1 and using the definition of the logarithm, we find that

$$1.5 \cdot 10^5 - P = Ce^{-t/100}. \tag{7.5.1}$$

This is a good time to determine the constant C. Since $P = 0$ when $t = 0$, we have

$$1.5 \cdot 10^5 - 0 = Ce^0 = C,$$

so $C = 1.5 \cdot 10^5$.

Using this value of C in Equation (7.5.1) and solving for P, we arrive at the solution

$$P(t) = 1.5 \cdot 10^5(1 - e^{-t/100}).$$

Superimposing the graph of P on the slope field we saw in Figure 7.5.3, we see, as shown in Figure 7.5.4

We see that, as expected, the amount of plastic pollution stabilizes around $1.5 \cdot 10^5$ cubic meters.

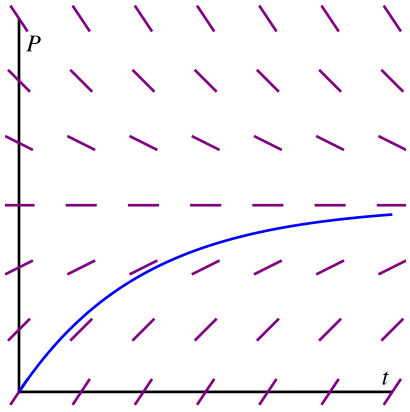


Figure 7.5.4. The solution $P(t)$ and the slope field for the differential equation $\frac{dP}{dt} = \frac{1}{100}(1.5 \cdot 10^5 - P)$.

This page titled 7.5: Modeling with Differential Equations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 7.5: Modeling with Differential Equations by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

7.6: Population Growth and the Logistic Equation

Motivating Questions

- How can we use differential equations to realistically model the growth of a population?
- How can we assess the accuracy of our models?

The growth of the earth's population is one of the pressing issues of our time. Will the population continue to grow? Or will it perhaps level off at some point, and if so, when? In this section, we look at two ways in which we may use differential equations to help us address these questions.

Before we begin, let's consider again two important differential equations that we have seen in earlier work this chapter.

Preview Activity 7.6.1

Recall that one model for population growth states that a population grows at a rate proportional to its size.

1. We begin with the differential equation

$$\frac{dP}{dt} = \frac{1}{2}P.$$

Sketch a slope field below as well as a few typical solutions on the axes provided.

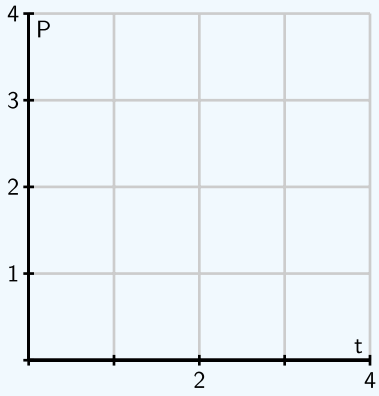


Figure 7.6.1

2. Find all equilibrium solutions of the equation $\frac{dP}{dt} = \frac{1}{2}P$ and classify them as stable or unstable.
3. If $P(0)$ is positive, describe the long-term behavior of the solution to $\frac{dP}{dt} = \frac{1}{2}P$.
4. Let's now consider a modified differential equation given by

$$\frac{dP}{dt} = \frac{1}{2}P(3 - P).$$

As before, sketch a slope field as well as a few typical solutions on the following axes provided.

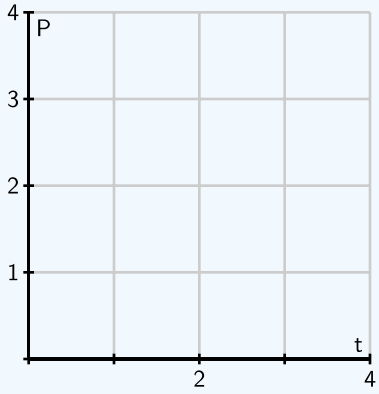


Figure 7.6.2

5. Find any equilibrium solutions and classify them as stable or unstable.
6. If $P(0)$ is positive, describe the long-term behavior of the solution.

7.6.1 The earth's population

We will now begin studying the earth's population. To get started, in Table 7.6.1 are some data for the earth's population in recent years that we will use in our investigations.

Table 7.6.1. Some recent population data for planet Earth.

Year	1998	1999	2000	2001	2002	2005	2006	2007	2008	2009	2010
Pop (billions)	5.932	6.008	6.084	6.159	6.234	6.456	6.531	6.606	6.681	6.756	6.831

Activity 7.6.2

Our first model will be based on the following assumption:

> *The rate of change of the population is proportional to the population.*

On the face of it, this seems pretty reasonable. When there is a relatively small number of people, there will be fewer births and deaths so the rate of change will be small. When there is a larger number of people, there will be more births and deaths so we expect a larger rate of change.

If $P(t)$ is the population t years after the year 2000, we may express this assumption as

$$\frac{dP}{dt} = kP$$

where k is a constant of proportionality.

1. Use the data in the table to estimate the derivative $P'(0)$ using a central difference. Assume that $t = 0$ corresponds to the year 2000.
2. What is the population $P(0)$?
3. Use your results from (a) and (b) to estimate the constant of proportionality k in the differential equation.
4. Now that we know the value of k, we have the initial value problem

$$\frac{dP}{dt} = kP, \ P(0) = 6.084.$$

Find the solution to this initial value problem.

5. What does your solution predict for the population in the year 2010? Is this close to the actual population given in the table?
6. When does your solution predict that the population will reach 12 billion?
7. What does your solution predict for the population in the year 2500?
8. Do you think this is a reasonable model for the earth's population? Why or why not? Explain your thinking using a couple of complete sentences.

Our work in Activity 7.6.2 shows that that the exponential model is fairly accurate for years relatively close to 2000. However, if we go too far into the future, the model predicts increasingly large rates of change, which causes the population to grow arbitrarily large. This does not make much sense since it is unrealistic to expect that the earth would be able to support such a large population.

The constant k in the differential equation has an important interpretation. Let's rewrite the differential equation $\frac{dP}{dt} = kP$ by solving for k, so that we have

$$k = \frac{dP/dt}{P}.$$

We see that k is the ratio of the rate of change to the population; in other words, it is the contribution to the rate of change from a single person. We call this the *per capita growth rate*.

In the exponential model we introduced in Activity 7.6.2, the per capita growth rate is constant. This means that when the population is large, the per capita growth rate is the same as when the population is small. It is natural to think that the per capita growth rate should decrease when the population becomes large, since there will not be enough resources to support so many people. We expect it would be a more realistic model to assume that the per capita growth rate depends on the population P.

In the previous activity, we computed the per capita growth rate in a single year by computing k, the quotient of $\frac{dP}{dt}$ and P (which we did for $t = 0$). If we return to the data in Table 7.6.1 and compute the per capita growth rate over a range of years, we generate the data shown in Figure 7.6.3, which shows how the per capita growth rate is a function of the population, P.

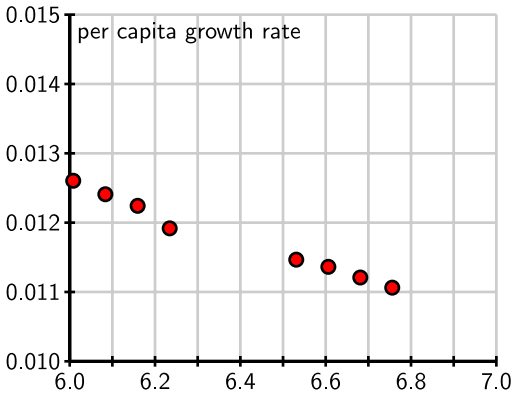


Figure 7.6.3. A plot of per capita growth rate vs. population P.

From the data, we see that the per capita growth rate appears to decrease as the population increases. In fact, the points seem to lie very close to a line, which is shown at two different scales in Figure 7.6.4.

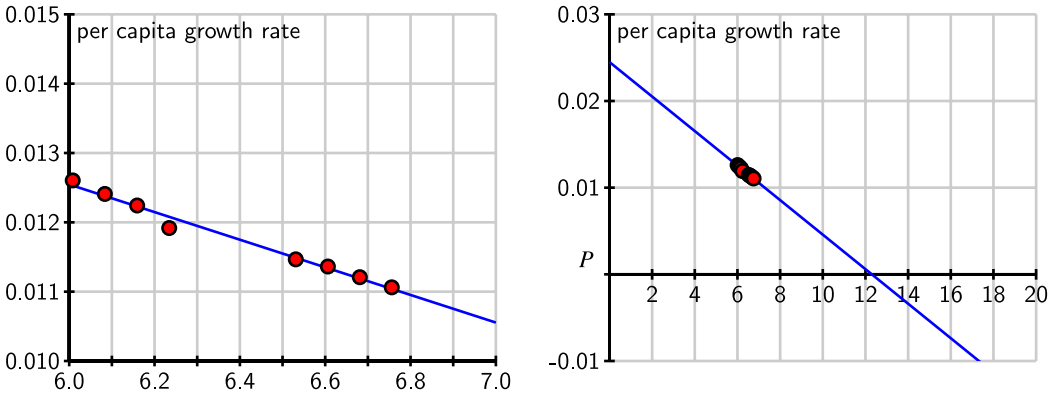


Figure 7.6.4. The line that approximates per capita growth as a function of population, P.

Looking at this line carefully, we can find its equation to be

$$\frac{dP/dt}{P} = 0.025 - 0.002P.$$

If we multiply both sides by P, we arrive at the differential equation

$$\frac{dP}{dt} = P(0.025 - 0.002P).$$

Graphing the dependence of dP/dt on the population P, we see that this differential equation demonstrates a quadratic relationship between $\frac{dP}{dt}$ and P, as shown in Figure 7.6.5.

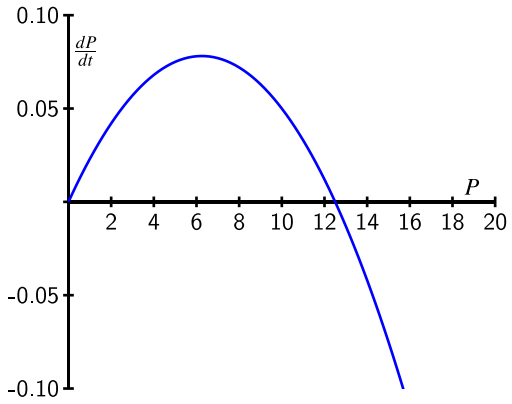


Figure 7.6.5. A plot of $\frac{dP}{dt}$ vs. P for the differential equation $\frac{dP}{dt} = P(0.025 - 0.002P)$.

The equation $\frac{dP}{dt} = P(0.025 - 0.002P)$ is an example of the *logistic equation*, and is the second model for population growth that we will consider. We expect that it will be more realistic, because the per capita growth rate is a decreasing function of the population.

Indeed, the graph in Figure 7.6.5 shows that there are two equilibrium solutions, $P = 0$, which is unstable, and $P = 12.5$, which is a stable equilibrium. The graph shows that any solution with $P(0) > 0$ will eventually stabilize around 12.5. Thus, our model predicts the world's population will eventually stabilize around 12.5 billion.

A prediction for the long-term behavior of the population is a valuable conclusion to draw from our differential equation. We would, however, also like to answer some quantitative questions. For instance, how long will it take to reach a population of 10 billion? To answer this question, we need to find an explicit solution of the equation.

Solving the logistic differential equation

Since we would like to apply the logistic model in more general situations, we state the logistic equation in its more general form,

$$\frac{dP}{dt} = kP(N-P). \tag{7.6.1}$$

The equilibrium solutions here are $P = 0$ and $1 - \frac{P}{N} = 0$, which shows that $P = N$. The equilibrium at $P = N$ is called the *carrying capacity* of the population for it represents the stable population that can be sustained by the environment.

We now solve the logistic equation (7.6.1). The equation is separable, so we separate the variables

$$\frac{1}{P(N-P)}\frac{dP}{dt} = k,$$

and integrate to find that

$$\int \frac{1}{P(N-P)}\, dP = \int k\, dt.$$

To find the antiderivative on the left, we use the partial fraction decomposition

$$\frac{1}{P(N-P)} = \frac{1}{N}\left[\frac{1}{P} + \frac{1}{N-P}\right].$$

Now we are ready to integrate, with

$$\int \frac{1}{N}\left[\frac{1}{P} + \frac{1}{N-P}\right]\, dP = \int k\, dt.$$

On the left, observe that N is constant, so we can remove a factor of $\frac{1}{N}$ and antidifferentiate to find that

$$\frac{1}{N}(\ln|P| - \ln|N-P|) = kt + C.$$

Multiplying both sides of this last equation by N and using a rule of logarithms, we next find that

$$\ln\left|\frac{P}{N-P}\right| = kNt + C.$$

From the definition of the logarithm, replacing e^C with C, and letting C absorb the absolute value signs, we now know that

$$\frac{P}{N-P} = Ce^{kNt}.$$

At this point, all that remains is to determine C and solve algebraically for P.

If the initial population is $P(0) = P_0$, then it follows that $C = \frac{P_0}{N-P_0}$, so

$$\frac{P}{N-P} = \frac{P_0}{N-P_0}e^{kNt}.$$

We will solve this equation for P by multiplying both sides by $(N-P)(N-P_0)$ to obtain

$$\begin{aligned} P(N-P_0) &= P_0(N-P)e^{kNt} \\ &= P_0Ne^{kNt} - P_0Pe^{kNt}. \end{aligned}$$

Solving for P_0Ne^{kNt}, expanding, and factoring, it follows that

$$\begin{aligned} P_0Ne^{kNt} &= P(N-P_0) + P_0Pe^{kNt} \\ &= P(N - P_0 + P_0e^{kNt}). \end{aligned}$$

Dividing to solve for P, we see that

$$P = \frac{P_0Ne^{kNt}}{N - P_0 + P_0e^{kNt}}.$$

Finally, we choose to multiply the numerator and denominator by $\frac{1}{P_0}e^{-kNt}$ to obtain

$$P(t) = \frac{N}{\left(\frac{N-P_0}{P_0}\right)e^{-kNt}+1}.$$

While that was a lot of algebra, notice the result: we have found an explicit solution to the logistic equation.

Solution to the Logistic Equation

The solution to the initial value problem

$$\frac{dP}{dt} = kP(N-P),\ P(0) = P_0,$$

is

$$P(t) = \frac{N}{\left(\frac{N-P_0}{P_0}\right)e^{-kNt}+1}. \tag{7.6.1}$$

For the logistic equation describing the earth's population that we worked with earlier in this section, we have

$$k = 0.002, N = 12.5, \text{ and } P_0 = 6.084.$$

This gives the solution

$$P(t) = \frac{12.5}{1.0546e^{-0.025t}+1},$$

whose graph is shown in Figure 7.6.6.

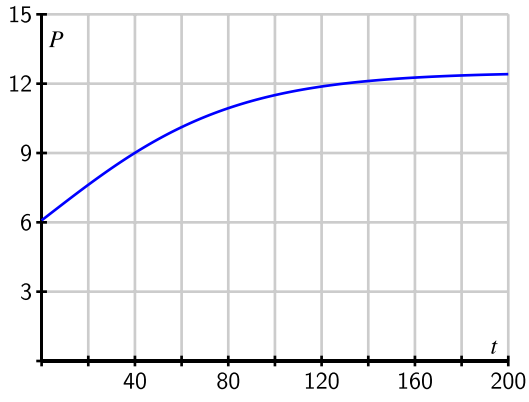


Figure 7.6.6. The solution to the logistic equation modeling the earth's population.

The graph shows the population leveling off at 12.5 billion, as we expected, and that the population will be around 10 billion in the year 2050. These results, which we have found using a relatively simple mathematical model, agree fairly well with predictions made using a much more sophisticated model developed by the United Nations.

The logistic equation is good for modeling any situation in which limited growth is possible. For instance, it could model the spread of a flu virus through a population contained on a cruise ship, the rate at which a rumor spreads within a small town, or the behavior of an animal population on an island. Through our work in this section, we have completely solved the logistic equation, regardless of the values of the constants N, k, and P_0. Anytime we encounter a logistic equation, we can apply the formula we found in Equation 7.6.1.

Activity 7.6.3

Consider the logistic equation

$$\frac{dP}{dt} = kP(N - P)$$

with the graph of $\frac{dP}{dt}$ vs. P shown in Figure 7.6.7.

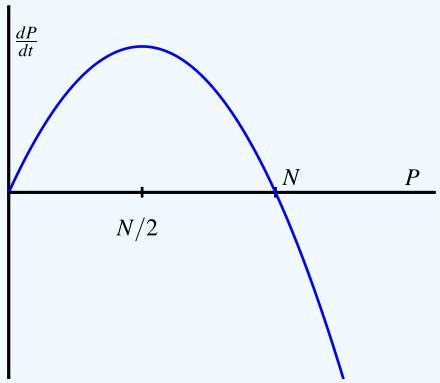


Figure 7.6.7. Plot of $\frac{dP}{dt}$ vs. P.

1. At what value of P is the rate of change greatest?
2. Consider the model for the earth's population that we created. At what value of P is the rate of change greatest? How does that compare to the population in recent years?
3. According to the model we developed, what will the population be in the year 2100?
4. According to the model we developed, when will the population reach 9 billion?
5. Now consider the general solution to the general logistic initial value problem that we found, given by

$$P(t) = \frac{N}{\left(\frac{N-P_0}{P_0}\right)e^{-kNt} + 1}.$$

Verify algebraically that $P(0) = P_0$ and that $\lim_{t\to\infty} P(t) = N$.

Summary

- If we assume that the rate of growth of a population is proportional to the population, we are led to a model in which the population grows without bound and at a rate that grows without bound.
- By assuming that the per capita growth rate decreases as the population grows, we are led to the logistic model of population growth, which predicts that the population will eventually stabilize at the carrying capacity.

This page titled 7.6: Population Growth and the Logistic Equation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 7.6: Population Growth and the Logistic Equation by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

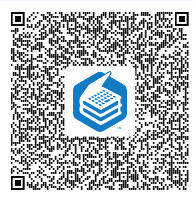
7.7: Differential Equations (Exercises)

7.1: An Introduction to Differential Equations

1. Matching solutions with equations

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

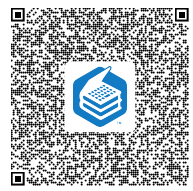
View this Interactive Content online



2. Finding constant to complete solution

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Choosing solution of $dy/dt = k(1 - Ay)$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

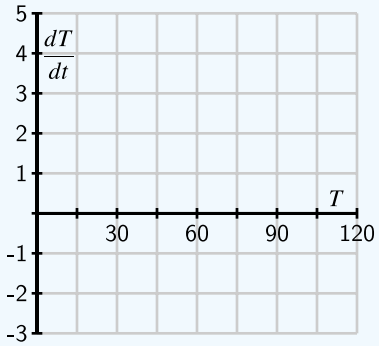


4.

Suppose that $T(t)$ represents the temperature of a cup of coffee set out in a room, where T is expressed in degrees Fahrenheit and t in minutes. A physical principle known as Newton's Law of Cooling tells us that

$$\frac{dT}{dt} = -\frac{1}{15}T + 5.$$

1. Supposes that $T(0) = 105$. What does the differential equation give us for the value of $\frac{dT}{dt}|_{T=105}$? Explain in a complete sentence the meaning of these two facts.
2. Is T increasing or decreasing at $t = 0$?
3. What is the approximate temperature at $t = 1$?
4. On the graph below, make a plot of dT/dt as a function of T.



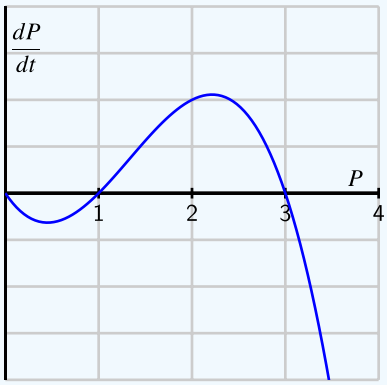
5. For which values of T does T increase? For which values of T does T decrease?
6. What do you think is the temperature of the room? Explain your thinking.
7. Verify that $T(t) = 75 + 30e^{-t/15}$ is the solution to the differential equation with initial value $T(0) = 105$. What happens to this solution after a long time?

5.

Suppose that the population of a particular species is described by the function $P(t)$, where P is expressed in millions. Suppose further that the population's rate of change is governed by the differential equation

$$\frac{dP}{dt} = f(P)$$

where $f(P)$ is the function graphed below.



1. For which values of the population P does the population increase?
2. For which values of the population P does the population decrease?
3. If $P(0) = 3$, how will the population change in time?
4. If the initial population satisfies $0 < P(0) < 1$, what will happen to the population after a very long time?
5. If the initial population satisfies $1 < P(0) < 3$, what will happen to the population after a very long time?
6. If the initial population satisfies $3 < P(0)$, what will happen to the population after a very long time?
7. This model for a population's growth is sometimes called “growth with a threshold.” Explain why this is an appropriate name.

6

In this problem, we test further what it means for a function to be a solution to a given differential equation.

1. Consider the differential equation

$$\frac{dy}{dt} = y - t.$$

Determine whether the following functions are solutions to the given differential equation.

1. $y(t) = t + 1 + 2e^t$
2. $y(t) = t + 1$
3. $y(t) = t + 2$

2. When you weigh bananas in a scale at the grocery store, the height h of the bananas is described by the differential equation

$$\frac{d^2h}{dt^2} = -kh$$

where k is the *spring constant*, a constant that depends on the properties of the spring in the scale. After you put the bananas in the scale, you (cleverly) observe that the height of the bananas is given by $h(t) = 4\sin(3t)$. What is the value of the spring constant?

7.2: Qualitative Behavior of Solutions to DE's

Exercises 7.2.4 Exercises

1. Graphing equilibrium solutions

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. Sketching solution curves

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Matching equations with direction fields

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Describing equilibrium solutions

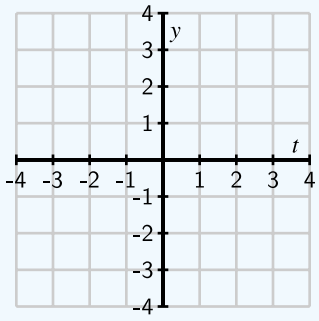
This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5

Consider the differential equation

$$\frac{dy}{dt} = t - y.$$



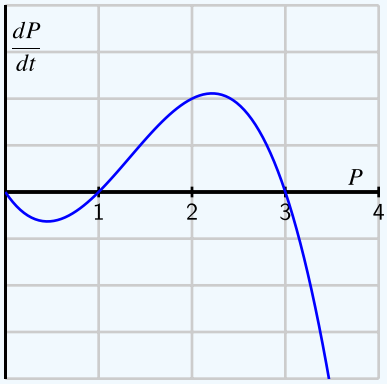
1. Sketch a slope field on the axes at right.
2. Sketch the solutions whose initial values are $y(0) = -4, -3, \ldots, 4$.
3. What do your sketches suggest is the solution whose initial value is $y(0) = -1$? Verify that this is indeed the solution to this initial value problem.
4. By considering the differential equation and the graphs you have sketched, what is the relationship between t and y at a point where a solution has a local minimum?

6

Consider the situation from problem 2 of Section 7.1: Suppose that the population of a particular species is described by the function $P(t)$, where P is expressed in millions. Suppose further that the population's rate of change is governed by the differential equation

$$\frac{dP}{dt} = f(P)$$

where $f(P)$ is the function graphed below.



1. Sketch a slope field for this differential equation. You do not have enough information to determine the actual slopes, but you should have enough information to determine where slopes are positive, negative, zero, large, or small, and hence determine the qualitative behavior of solutions.
2. Sketch some solutions to this differential equation when the initial population $P(0) > 0$.
3. Identify any equilibrium solutions to the differential equation and classify them as stable or unstable.
4. If $P(0) > 1$, what is the eventual fate of the species? if $P(0) < 1$?
5. Remember that we referred to this model for population growth as "growth with a threshold." Explain why this characterization makes sense by considering solutions whose inital value is close to 1.

7

The population of a species of fish in a lake is $P(t)$ where P is measured in thousands of fish and t is measured in months. The growth of the population is described by the differential equation

$$\frac{dP}{dt} = f(P) = P(6 - P).$$

1. Sketch a graph of $f(P) = P(6 - P)$ and use it to determine the equilibrium solutions and whether they are stable or unstable. Write a complete sentence that describes the long-term behavior of the fish population.
2. Suppose now that the owners of the lake allow fishers to remove 1000 fish from the lake every month (remember that $P(t)$ is measured in *thousands* of fish). Modify the differential equation to take this into account. Sketch the new graph of dP/dt versus P. Determine the new equilibrium solutions and decide whether they are stable or unstable.
3. Given the situation in part (b), give a description of the long-term behavior of the fish population.
4. Suppose that fishermen remove h thousand fish per month. How is the differential equation modified?
5. What is the largest number of fish that can be removed per month without eliminating the fish population? If fish are removed at this maximum rate, what is the eventual population of fish?

8

Let $y(t)$ be the number of thousands of mice that live on a farm; assume time t is measured in years.[1]

1. The population of the mice grows at a yearly rate that is twenty times the number of mice. Express this as a differential equation.
2. At some point, the farmer brings C cats to the farm. The number of mice that the cats can eat in a year is

   $$M(y) = C\frac{y}{2 + y}$$

   thousand mice per year. Explain how this modifies the differential equation that you found in part a).
3. Sketch a graph of the function $M(y)$ for a single cat $C = 1$ and explain its features by looking, for instance, at the behavior of $M(y)$ when y is small and when y is large.

4. Suppose that $C = 1$. Find the equilibrium solutions and determine whether they are stable or unstable. Use this to explain the long-term behavior of the mice population depending on the initial population of the mice.
5. Suppose that $C = 60$. Find the equilibrium solutions and determine whether they are stable or unstable. Use this to explain the long-term behavior of the mice population depending on the initial population of the mice.
6. What is the smallest number of cats you would need to keep the mice population from growing arbitrarily large?

This problem is based on an ecological analysis presented in a research paper by C.S. Hollings: The Components of Predation as Revealed by a Study of Small Mammal Predation of the European Pine Sawfly, *Canadian Entomology 91*: 283-320.

7.3: Euler's Method

1. A few steps of Euler's method

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Using Euler's method for a solution of $y' = 4y$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Using Euler's method with different time steps

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4

Newton's Law of Cooling says that the rate at which an object, such as a cup of coffee, cools is proportional to the difference in the object's temperature and room temperature. If $T(t)$ is the object's temperature and T_r is room temperature, this law is expressed at

$$\frac{dT}{dt} = -k(T - T_r),$$

where k is a constant of proportionality. In this problem, temperature is measured in degrees Fahrenheit and time in minutes.

1. Two calculus students, Alice and Bob, enter a 70° classroom at the same time. Each has a cup of coffee that is 100°. The differential equation for Alice has a constant of proportionality $k = 0.5$, while the constant of proportionality for Bob is $k = 0.1$. What is the initial rate of change for Alice's coffee? What is the initial rate of change for Bob's coffee?

2. What feature of Alice's and Bob's cups of coffee could explain this difference?
3. As the heating unit turns on and off in the room, the temperature in the room is

$$T_r = 70 + 10 \sin t.$$

Implement Euler's method with a step size of $\Delta t = 0.1$ to approximate the temperature of Alice's coffee over the time interval $0 \leq t \leq 50$. This will most easily be performed using a spreadsheet such as *Excel*. Graph the temperature of her coffee and room temperature over this interval.

4. In the same way, implement Euler's method to approximate the temperature of Bob's coffee over the same time interval. Graph the temperature of his coffee and room temperature over the interval.
5. Explain the similarities and differences that you see in the behavior of Alice's and Bob's cups of coffee.

5

We have seen that the error in approximating the solution to an initial value problem is proportional to Δt. That is, if $E_{\Delta t}$ is the Euler's method approximation to the solution to an initial value problem at $\bar{t}$, then

$$y(\bar{t}) - E_{\Delta t} \approx K \Delta t$$

for some constant of proportionality K.

In this problem, we will see how to use this fact to improve our estimates, using an idea called *accelerated convergence*.

1. We will create a new approximation by assuming the error is *exactly* proportional to Δt, according to the formula

$$y(\bar{t}) - E_{\Delta t} = K \Delta t.$$

Using our earlier results from the initial value problem $dy/dt = y$ and $y(0) = 1$ with $\Delta t = 0.2$ and $\Delta t = 0.1$, we have

$$y(1) - 2.4883 = 0.2K$$
$$y(1) - 2.5937 = 0.1K.$$

This is a system of two linear equations in the unknowns $y(1)$ and K. Solve this system to find a new approximation for $y(1)$. (You may remember that the exact value is $y(1) = e = 2.71828 \ldots$.)

2. Use the other data, $E_{0.05} = 2.6533$ and $E_{0.025} = 2.6851$ to do similar work as in (a) to obtain another approximation. Which gives the better approximation? Why do you think this is?
3. Let's now study the initial value problem

$$\frac{dy}{dt} = t - y, \; y(0) = 0.$$

Approximate $y(0.3)$ by applying Euler's method to find approximations $E_{0.1}$ and $E_{0.05}$. Now use the idea of accelerated convergence to obtain a better approximation. (For the sake of comparison, you want to note that the actual value is $y(0.3) = 0.0408$.)

6

In this problem, we'll modify Euler's method to obtain better approximations to solutions of initial value problems. This method is called the *Improved Euler's method*.

In Euler's method, we walk across an interval of width Δt using the slope obtained from the differential equation at the left endpoint of the interval. Of course, the slope of the solution will most likely change over this interval. We can improve our approximation by trying to incorporate the change in the slope over the interval.

Let's again consider the initial value problem $dy/dt = y$ and $y(0) = 1$, which we will approximate using steps of width $\Delta t = 0.2$. Our first interval is therefore $0 \le t \le 0.2$. At $t = 0$, the differential equation tells us that the slope is 1, and the approximation we obtain from Euler's method is that $y(0.2) \approx y_1 = 1 + 1(0.2) = 1.2$.

This gives us some idea for how the slope has changed over the interval $0 \le t \le 0.2$. We know the slope at $t = 0$ is 1, while the slope at $t = 0.2$ is 1.2, trusting in the Euler's method approximation. We will therefore refine our estimate of the initial slope to be the average of these two slopes; that is, we will estimate the slope to be $(1 + 1.2)/2 = 1.1$. This gives the new approximation $y(1) = y_1 = 1 + 1.1(0.2) = 1.22$.

The first few steps look like what is found in Table 7.3.15.

Table 7.3.15. The first several steps of the improved Euler's method

t_i	y_i	Slope at (t_{i+1}, y_{i+1})	Average slope
0.0	1.0000	1.2000	1.1000
0.2	1.2200	1.4640	1.3420
0.4	1.4884	1.7861	1.6372
⋮	⋮	⋮	⋮

1. Continue with this method to obtain an approximation for $y(1) = e$.
2. Repeat this method with $\Delta t = 0.1$ to obtain a better approximation for $y(1)$.
3. We saw that the error in Euler's method is proportional to Δt. Using your results from parts (a) and (b), what power of Δt appears to be proportional to the error in the Improved Euler's Method?

7.4: Separable Differential Equations

1. Initial value problem for $dy/dx = x^8 y$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Interactive Content

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Initial value problem for $dy/dt = y^2(8+t)$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Initial value problem for $du/dt = e^{6u+10t}$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Initial value problem for $dy/dx = 170yx^{16}$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6

The mass of a radioactive sample decays at a rate that is proportional to its mass.

1. Express this fact as a differential equation for the mass $M(t)$ using k for the constant of proportionality.
2. If the initial mass is M_0, find an expression for the mass $M(t)$.
3. The *half-life* of the sample is the amount of time required for half of the mass to decay. Knowing that the half-life of Carbon-14 is 5730 years, find the value of k for a sample of Carbon-14.
4. How long does it take for a sample of Carbon-14 to be reduced to one-quarter its original mass?
5. Carbon-14 naturally occurs in our environment; any living organism takes in Carbon-14 when it eats and breathes. Upon dying, however, the organism no longer takes in Carbon-14. Suppose that you find remnants of a pre-historic firepit. By analyzing the charred wood in the pit, you determine that the amount of Carbon-14 is only 30% of the amount in living trees. Estimate the age of the firepit. [2]

This approach is the basic idea behind radiocarbon dating.

7

Consider the initial value problem

$$\frac{dy}{dt} = -\frac{t}{y},\ y(0) = 8$$

1. Find the solution of the initial value problem and sketch its graph.
2. For what values of t is the solution defined?
3. What is the value of y at the last time that the solution is defined?

4. By looking at the differential equation, explain why we should not expect to find solutions with the value of y you noted in (c).

8

Suppose that a cylindrical water tank with a hole in the bottom is filled with water. The water, of course, will leak out and the height of the water will decrease. Let $h(t)$ denote the height of the water. A physical principle called *Torricelli's Law* implies that the height decreases at a rate proportional to the square root of the height.

1. Express this fact using k as the constant of proportionality.
2. Suppose you have two tanks, one with $k = -1$ and another with $k = -10$. What physical differences would you expect to find?
3. Suppose you have a tank for which the height decreases at 20 inches per minute when the water is filled to a depth of 100 inches. Find the value of k.
4. Solve the initial value problem for the tank in part (c), and graph the solution you determine.
5. How long does it take for the water to run out of the tank?
6. Is the solution that you found valid for all time t? If so, explain how you know this. If not, explain why not.

9

The *Gompertz equation* is a model that is used to describe the growth of certain populations. Suppose that $P(t)$ is the population of some organism and that

$$\frac{dP}{dt} = -P \ln\left(\frac{P}{3}\right) = -P(\ln P - \ln 3).$$

1. Sketch a slope field for $P(t)$ over the range $0 \leq P \leq 6$.
2. Identify any equilibrium solutions and determine whether they are stable or unstable.
3. Find the population $P(t)$ assuming that $P(0) = 1$ and sketch its graph. What happens to $P(t)$ after a very long time?
4. Find the population $P(t)$ assuming that $P(0) = 6$ and sketch its graph. What happens to $P(t)$ after a very long time?
5. Verify that the long-term behavior of your solutions agrees with what you predicted by looking at the slope field.

7.5: Modeling with Differential Equations

1. Mixing problem

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Mixing problem

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Population growth problem

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Radioactive decay problem

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Investment problem

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6

Congratulations, you just won the lottery! In one option presented to you, you will be paid one million dollars a year for the next 25 years. You can deposit this money in an account that will earn 5% each year.

1. Set up a differential equation that describes the rate of change in the amount of money in the account. Two factors cause the amount to grow—first, you are depositing one millon dollars per year and second, you are earning 5% interest.
2. If there is no amount of money in the account when you open it, how much money will you have in the account after 25 years?
3. The second option presented to you is to take a lump sum of 10 million dollars, which you will deposit into a similar account. How much money will you have in that account after 25 years?
4. Do you prefer the first or second option? Explain your thinking.
5. At what time does the amount of money in the account under the first option overtake the amount of money in the account under the second option?

7

When a skydiver jumps from a plane, gravity causes her downward velocity to increase at the rate of $g \approx 9.8$ meters per second squared. At the same time, wind resistance causes her velocity to decrease at a rate proportional to the velocity.

1. Using k to represent the constant of proportionality, write a differential equation that describes the rate of change of the skydiver's velocity.
2. Find any equilibrium solutions and decide whether they are stable or unstable. Your result should depend on k.
3. Suppose that the initial velocity is zero. Find the velocity $v(t)$.

4. A typical terminal velocity for a skydiver falling face down is 54 meters per second. What is the value of k for this skydiver?
5. How long does it take to reach 50% of the terminal velocity?

8

During the first few years of life, the rate at which a baby gains weight is proportional to the reciprocal of its weight.

1. Express this fact as a differential equation.
2. Suppose that a baby weighs 8 pounds at birth and 9 pounds one month later. How much will he weigh at one year?
3. Do you think this is a realistic model for a long time?

9

Suppose that you have a water tank that holds 100 gallons of water. A briny solution, which contains 20 grams of salt per gallon, enters the tank at the rate of 3 gallons per minute.

At the same time, the solution is well mixed, and water is pumped out of the tank at the rate of 3 gallons per minute.

1. Since 3 gallons enters the tank every minute and 3 gallons leaves every minute, what can you conclude about the volume of water in the tank.
2. How many grams of salt enters the tank every minute?
3. Suppose that $S(t)$ denotes the number of grams of salt in the tank in minute t. How many grams are there in each gallon in minute t?
4. Since water leaves the tank at 3 gallons per minute, how many grams of salt leave the tank each minute?
5. Write a differential equation that expresses the total rate of change of S.
6. Identify any equilibrium solutions and determine whether they are stable or unstable.
7. Suppose that there is initially no salt in the tank. Find the amount of salt $S(t)$ in minute t.
8. What happens to $S(t)$ after a very long time? Explain how you could have predicted this only knowing how much salt there is in each gallon of the briny solution that enters the tank.

7.6: Population Growth and the Logistic Equation

Exercises 7.6.4 Exercises

1. Analyzing a logistic equation

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Analyzing a logistic model

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Finding a logistic function for an infection model

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Analyzing a population growth model

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5

The logistic equation may be used to model how a rumor spreads through a group of people. Suppose that $p(t)$ is the fraction of people that have heard the rumor on day t. The equation

$$\frac{dp}{dt} = 0.2p(1-p)$$

describes how p changes. Suppose initially that one-tenth of the people have heard the rumor; that is, $p(0) = 0.1$.

1. What happens to $p(t)$ after a very long time?
2. Determine a formula for the function $p(t)$.
3. At what time is p changing most rapidly?
4. How long does it take before 80% of the people have heard the rumor?

6

Suppose that $b(t)$ measures the number of bacteria living in a colony in a Petri dish, where b is measured in thousands and t is measured in days. One day, you measure that there are 6,000 bacteria and the per capita growth rate is 3. A few days later, you measure that there are 9,000 bacteria and the per capita growth rate is 2.

1. Assume that the per capita growth rate $\frac{db/dt}{b}$ is a linear function of b. Use the measurements to find this function and write a logistic equation to describe $\frac{db}{dt}$.
2. What is the carrying capacity for the bacteria?
3. At what population is the number of bacteria increasing most rapidly?

4. If there are initially 1,000 bacteria, how long will it take to reach 80% of the carrying capacity?

7

Suppose that the population of a species of fish is controlled by the logistic equation

$$\frac{dP}{dt} = 0.1P(10 - P),$$

where P is measured in thousands of fish and t is measured in years.

1. What is the carrying capacity of this population?
2. Suppose that a long time has passed and that the fish population is stable at the carrying capacity. At this time, humans begin harvesting 20% of the fish every year. Modify the differential equation by adding a term to incorporate the harvesting of fish.
3. What is the new carrying capacity?
4. What will the fish population be one year after the harvesting begins?
5. How long will it take for the population to be within 10% of the carrying capacity?

This page titled 7.7: Differential Equations (Exercises) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 7.E: Differential Equations (Exercises) has no license indicated.
- 7.1: An Introduction to Differential Equations by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 7.2: Qualitative Behavior of Solutions to Differential Equations by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 7.3: Euler's Method by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 7.4: Separable Differential Equations by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 7.5: Modeling with Differential Equations by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 7.6: Population Growth and the Logistic Equation by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

8: SEQUENCES AND SERIES

8.1: Sequences
8.2: Geometric Series
8.3: Series of Real Numbers
8.4: Alternating Series
8.5: Taylor Polynomials and Taylor Series
8.6: Power Series
8.7: Sequences and Series (Exercises)

This page titled 8: Sequences and Series is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 8: Sequences and Series by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

8.1: Sequences

Motivating Questions

- What is a sequence?
- What does it mean for a sequence to converge?
- What does it mean for a sequence to diverge?

We encounter sequences every day. Your monthly utility payments, the annual interest you earn on investments, the amount you spend on groceries each week; all are examples of sequences. Other sequences with which you may be familiar include the Fibonacci sequence $1, 1, 2, 3, 5, 8, \ldots$, where each term is the sum of the two preceding terms, and the triangular numbers $1, 3, 6, 10, 15, 21, 28, 36, 45, 55, \ldots$, the number of vertices in the triangles shown in Figure 8.1.1.

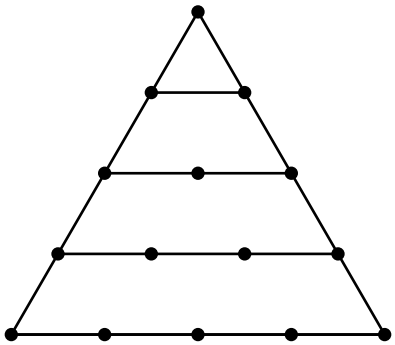


Figure 8.1.1. Triangular numbers

Sequences of integers are of such interest to mathematicians and others that they have a journal [1] devoted to them and an on-line encyclopedia [2] that catalogs a huge number of integer sequences and their connections. Sequences are also used in digital recordings and images.

The Journal of Integer Sequences at http://www.cs.uwaterloo.ca/journals/JIS/
The On-Line Encyclopedia of Integer Sequences at http://oeis.org/

Our studies in calculus have dealt with continuous functions. Sequences model *discrete* instead of continuous information. We will study ways to represent and work with discrete information in this chapter as we investigate *sequences* and *series*, and ultimately see key connections between the discrete and continuous.

Preview Activity 8.1.1

Suppose you receive $\$5000$ through an inheritance. You decide to invest this money into a fund that pays 8% annually, compounded monthly. That means that each month your investment earns $\frac{0.08}{12} \cdot P$ additional dollars, where P is your principal balance at the start of the month. So in the first month your investment earns

$$5000\left(\frac{0.08}{12}\right)$$

or $\$33.33$. If you reinvest this money, you will then have $\$5033.33$ in your account at the end of the first month. From this point on, assume that you reinvest all of the interest you earn.

1. How much interest will you earn in the second month? How much money will you have in your account at the end of the second month?

2. Complete Table 8.1.2 to determine the interest earned and total amount of money in this investment each month for one year.

Table 8.1.1. *Interest*

Month	Interest earned	Total amount of money in the account
0	\$0.00	\$5000.00
1	\$33.33	\$5033.33
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		

3. As we will see later, the amount of money P_n in the account after month n is given by

$$P_n = 5000\left(1 + \frac{0.08}{12}\right)^n.$$

Use this formula to check your calculations in Table 8.1.2. Then find the amount of money in the account after 5 years.

4. How many years will it be before the account has doubled in value to \$10000?

Sequences

As Preview Activity 8.1.1 illustrates, many discrete phenomena can be represented as lists of numbers (like the amount of money in an account over a period of months). We call any such list a *sequence*. A sequence is nothing more than list of terms in some order. We often list the entries of the sequence with subscripts,

$$s_1, s_2, \ldots, s_n \ldots,$$

where the subscript denotes the position of the entry in the sequence.

Definition 8.1.1

A sequence is a list of terms $s_1, s_2, s_3, \ldots$ in a specified order.

We can think of a sequence as a function f whose domain is the set of positive integers where $f(n) = s_n$ for each positive integer n. This alternative view will be be useful in many situations.

We often denote the sequence

$$s_1, s_2, s_3, \ldots$$

by $\{s_n\}$. The value s_n (alternatively $s(n)$) is called the nth *term* in the sequence. If the terms are all 0 after some fixed value of n, we say the sequence is finite. Otherwise the sequence is infinite. With infinite sequences, we are often interested in their end behavior and the idea of *convergent* sequences.

Activity 8.1.2

1. Let s_n be the nth term in the sequence $1, 2, 3, \ldots$. Find a formula for s_n and use appropriate technological tools to draw a graph of entries in this sequence by plotting points of the form (n, s_n) for some values of n. Most graphing calculators can plot sequences; directions follow for the TI-84.
   - In the MODEmenu, highlight SEQin the FUNCline and press ENTER.
   - In the Y=menu, you will now see lines to enter sequences. Enter a value for nMin (where the sequence starts), a function for u(n) (the nth term in the sequence), and the value of u(nMin).
   - Set your window coordinates (this involves choosing limits for n as well as the window coordinates XMin, XMax, YMin, and YMax.
   - The GRAPHkey will draw a plot of your sequence.

   Using your knowledge of limits of continuous functions as $x \to \infty$, decide if this sequence $\{s_n\}$ has a limit as $n \to \infty$. Explain your reasoning.

2. Let s_n be the nth term in the sequence $1, \frac{1}{2}, \frac{1}{3}, \ldots$. Find a formula for s_n. Draw a graph of some points in this sequence. Using your knowledge of limits of continuous functions as $x \to \infty$, decide if this sequence $\{s_n\}$ has a limit as $n \to \infty$. Explain your reasoning.
3. Let s_n be the nth term in the sequence $2, \frac{3}{2}, \frac{4}{3}, \frac{5}{4}, \ldots$. Find a formula for s_n. Using your knowledge of limits of continuous functions as $x \to \infty$, decide if this sequence $\{s_n\}$ has a limit as $n \to \infty$. Explain your reasoning.

Next we formalize the ideas from Activity 8.1.2.

Activity 8.1.3

1. Recall our earlier work with limits involving infinity in Section 2.8. State clearly what it means for a continuous function f to have a limit L as $x \to \infty$.
2. Given that an infinite sequence of real numbers is a function from the integers to the real numbers, apply the idea from part (a) to explain what you think it means for a sequence $\{s_n\}$ to have a limit as $n \to \infty$.
3. Based on your response to the part (b), decide if the sequence $\left\{\frac{1+n}{2+n}\right\}$ has a limit as $n \to \infty$. If so, what is the limit? If not, why not?

In Activities 8.1.2 and 8.1.3 we investigated a sequence $\{s_n\}$ that has a limit as n goes to infinity. More formally, we make the following definition.

Definition 8.1.2

A sequence $\{s_n\}$ ***converges*** or is a ***convergent sequence*** provided that there is a number L such that we can make s_n as close to L as we want by taking n sufficiently large. In this situation, we call L the ***limit*** of the convergent sequence and write

$$\lim_{n\to\infty} s_n = L.$$

If the sequence $\{s_n\}$ does not converge, we say that the sequence $\{s_n\}$ ***diverges***.

The idea of sequence having a limit as $n \to \infty$ is the same as the idea of a continuous function having a limit as $x \to \infty$. The only difference is that sequences are discrete instead of continuous.

Activity 8.1.4

Use graphical and/or algebraic methods to determine whether each of the following sequences converges or diverges.

1. $\left\{\dfrac{1+2n}{3n-2}\right\}$
2. $\left\{\dfrac{5+3^n}{10+2^n}\right\}$
3. $\left\{\frac{10^n}{n!}\right\}$ (where ! is the *factorial* symbol and $n! = n(n-1)(n-2)\cdots(2)(1)$ for any positive integer n (as convention we define $0!$ to be 1)).

Summary

- A sequence is a list of objects in a specified order. We will typically work with sequences of real numbers. We can think of a sequence as a function from the positive integers to the set of real numbers.
- A sequence $\{s_n\}$ of real numbers converges to a number L if we can make every value of s_k for $k \geq n$ as close as we want to L by choosing n sufficiently large.
- A sequence diverges if it does not converge.

This page titled 8.1: Sequences is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.1: Sequences by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

8.2: Geometric Series

Motivating Questions

- What is a geometric series?
- What is a partial sum of a geometric series? What is a simplified form of the nth partial sum of a geometric series?
- Under what conditions does a geometric series converge? What is the sum of a convergent geometric series?

Many important sequences are generated by addition. In Preview Activity 8.2.1, we see an example of a sequence that is connected to a sum.

Preview Activity 8.2.1

Warfarin is an anticoagulant that prevents blood clotting; often it is prescribed to stroke victims in order to help ensure blood flow. The level of warfarin has to reach a certain concentration in the blood in order to be effective.

Suppose warfarin is taken by a particular patient in a 5 mg dose each day. The drug is absorbed by the body and some is excreted from the system between doses. Assume that at the end of a 24 hour period, 8% of the drug remains in the body. Let $Q(n)$ be the amount (in mg) of warfarin in the body before the $(n+1)$st dose of the drug is administered.

1. Explain why $Q(1) = 5 \times 0.08$ mg.
2. Explain why $Q(2) = (5 + Q(1)) \times 0.08$ mg. Then show that
$$Q(2) = (5 \times 0.08)\,(1 + 0.08)\,\text{mg}.$$
3. Explain why $Q(3) = (5 + Q(2)) \times 0.08$ mg. Then show that
$$Q(3) = (5 \times 0.08)\left(1 + 0.08 + 0.08^2\right)\text{mg}.$$
4. Explain why $Q(4) = (5 + Q(3)) \times 0.08$ mg. Then show that
$$Q(4) = (5 \times 0.08)\left(1 + 0.08 + 0.08^2 + 0.08^3\right)\text{mg}.$$
5. There is a pattern that you should see emerging. Use this pattern to find a formula for $Q(n)$, where n is an arbitrary positive integer.
6. Complete Table 8.2.1 with values of $Q(n)$ for the provided n-values (reporting $Q(n)$ to 10 decimal places). What appears to be happening to the sequence $Q(n)$ as n increases?

Table 8.2.1. Values of $Q(n)$ for selected values of n

n	1	2	3	4	5	6	7	8	9	10
$Q(n)$	0.40									

Geometric Series

In Preview Activity 8.2.1 we encountered the sum

$$(5 \times 0.08)\left(1 + 0.08 + 0.08^2 + 0.08^3 + \cdots + 0.08^{n-1}\right)$$

for the long-term level of Warfarin in the patient's system. This sum has the form

$$a + ar + ar^2 + \cdots + ar^{n-1} \tag{8.2.1}$$

where $a = 5 \times 0.08$ and $r = 0.08$. Such a sum is called a *finite geometric series* with ratio r.

Activity 8.2.2

Let a and r be real numbers (with $r \neq 1$) and let

$$S_n = a + ar + ar^2 + \cdots + ar^{n-1}.$$

In this activity we will find a shortcut formula for S_n that does not involve a sum of n terms.

1. Multiply S_n by r. What does the resulting sum look like?
2. Subtract rS_n from S_n and explain why
$$S_n - rS_n = a - ar^n. \tag{8.2.2}$$
3. Solve equation (8.2.2) for S_n to find a simple formula for S_n that does not involve adding n terms.

The sum of the terms of a sequence is called a *series*. We summarize the result of Activity 8.2.2 in the following way.

Note

A finite geometric series S_n is a sum of the form

$$S_n = a + ar + ar^2 + \cdots + ar^{n-1}, \tag{8.2.3}$$

where a and r are real numbers such that $r \neq 1$. The finite geometric series S_n can be written more simply as

$$S_n = a + ar + ar^2 + \cdots + ar^{n-1} = \frac{a(1-r^n)}{1-r}. \tag{8.2.4}$$

We now apply Equation (8.2.4) to the example involving Warfarin from Preview Activity 8.2.1. Recall that

$$Q(n) = (5 \times 0.08)\left(1 + 0.08 + 0.08^2 + 0.08^3 + \cdots + 0.08^{n-1}\right) \text{mg},$$

so $Q(n)$ is a geometric series with $a = 5 \times 0.08 = 0.4$ and $r = 0.08$. Thus,

$$Q(n) = 0.4\left(\frac{1-0.08^n}{1-0.08}\right) = \frac{1}{2.3}\left(1-0.08^n\right).$$

Notice that as n goes to infinity, the value of 0.08^n goes to 0. So,

$$\lim_{n\to\infty} Q(n) = \lim_{n\to\infty} \frac{1}{2.3}\left(1-0.08^n\right) = \frac{1}{2.3} \approx 0.435.$$

Therefore, the long-term level of Warfarin in the blood under these conditions is $\frac{1}{2.3}$, which is approximately 0.435 mg.

To determine the long-term effect of Warfarin, we considered a finite geometric series of n terms, and then considered what happened as n was allowed to grow without bound. In this sense, we were actually interested in an infinite geometric series (the result of letting n go to infinity in the finite sum).

Definition 8.2.1

An infinite geometric series is an infinite sum of the form

$$a + ar + ar^2 + \cdots = \sum_{n=0}^{\infty} ar^n. \tag{8.2.5}$$

The value of r in the geometric series (8.2.5) is called the *common ratio* of the series because the ratio of the $(n+1)$st term, ar^n, to the nth term, ar^{n-1}, is always r:

$$\frac{ar^n}{ar^{n-1}} = r.$$

Geometric series are common in mathematics and arise naturally in many different situations. As a familiar example, suppose we want to write the number with repeating decimal expansion

$$N = 0.121\overline{212}$$

as a rational number. Observe that

$$\begin{aligned} N &= 0.12 + 0.0012 + 0.000012 + \cdots \\ &= \left(\frac{12}{100}\right) + \left(\frac{12}{100}\right)\left(\frac{1}{100}\right) + \left(\frac{12}{100}\right)\left(\frac{1}{100}\right)^2 + \cdots. \end{aligned}$$

This is an infinite geometric series with $a = \frac{12}{100}$ and $r = \frac{1}{100}$.

By using the formula for the value of a finite geometric sum, we can also develop a formula for the value of an infinite geometric series. We explore this idea in the following activity.

Activity 8.2.3

Let $r \neq 1$ and a be real numbers and let

$$S = a + ar + ar^2 + \cdots ar^{n-1} + \cdots$$

be an infinite geometric series. For each positive integer n, let

$$S_n = a + ar + ar^2 + \cdots + ar^{n-1}.$$

Recall that

$$S_n = a\frac{1 - r^n}{1 - r}.$$

1. What should we allow n to approach in order to have S_n approach S?
2. What is the value of $\lim_{n\to\infty} r^n$ for $|r| > 1$? for $|r| < 1$? Explain.
3. If $|r| < 1$, use the formula for S_n and your observations in (a) and (b) to explain why S is finite and find a resulting formula for S.

We can now find the value of the geometric series

$$N = \left(\frac{12}{100}\right) + \left(\frac{12}{100}\right)\left(\frac{1}{100}\right) + \left(\frac{12}{100}\right)\left(\frac{1}{100}\right)^2 + \cdots.$$

Using $a = \frac{12}{100}$ and $r = \frac{1}{100}$, we see that

$$N = \frac{12}{100}\left(\frac{1}{1 - \frac{1}{100}}\right) = \frac{12}{100}\left(\frac{100}{99}\right) = \frac{4}{33}.$$

The sum of a finite number of terms of an infinite geometric series is often called a *partial sum* of the series. Thus,

$$S_n = a + ar + ar^2 + \cdots + ar^{n-1} = \sum_{k=0}^{n-1} ar^k.$$

is called the nth partial sum of the series $\sum_{k=0}^{\infty} ar^k$. We summarize our recent work with geometric series as follows.

Note

- An infinite geometric series is an infinite sum of the form

$$a + ar + ar^2 + \cdots = \sum_{n=0}^{\infty} ar^n, \tag{8.2.6}$$

  where a and r are real numbers such that $r \neq 0$.

- The nth partial sum S_n of an infinite geometric series is

$$S_n = a + ar + ar^2 + \cdots + ar^{n-1}.$$

- If $|r| < 1$, then using the fact that $S_n = a\frac{1-r^n}{1-r}$, it follows that the sum S of the infinite geometric series (8.2.6) is

$$S = \lim_{n\to\infty} S_n = \lim_{n\to\infty} a\frac{1-r^n}{1-r} = \frac{a}{1-r}$$

Activity 8.2.4

The formulas we have derived for an infinite geometric series and its partial sum have assumed we begin indexing the sums at $n = 0$. If instead we have a sum that does not begin at $n = 0$, we can factor out common terms and use the established formulas. This process is illustrated in the examples in this activity.

1. Consider the sum

$$\sum_{k=1}^{\infty} (2)\left(\frac{1}{3}\right)^k = (2)\left(\frac{1}{3}\right) + (2)\left(\frac{1}{3}\right)^2 + (2)\left(\frac{1}{3}\right)^3 + \cdots.$$

   Remove the common factor of $(2)\left(\frac{1}{3}\right)$ from each term and hence find the sum of the series.

2. Next let a and r be real numbers with $-1 < r < 1$. Consider the sum

$$\sum_{k=3}^{\infty} ar^k = ar^3 + ar^4 + ar^5 + \cdots.$$

   Remove the common factor of ar^3 from each term and find the sum of the series.

3. Finally, we consider the most general case. Let a and r be real numbers with $-1 < r < 1$, let n be a positive integer, and consider the sum

$$\sum_{k=n}^{\infty} ar^k = ar^n + ar^{n+1} + ar^{n+2} + \cdots.$$

   Remove the common factor of ar^n from each term to find the sum of the series.

Summary

- An infinite geometric series is an infinite sum of the form

$$\sum_{k=0}^{\infty} ar^k$$

  where a and r are real numbers and $r \neq 0$.

- The nth partial sum of the geometric series $\sum_{k=0}^{\infty} ar^k$ is

$$S_n = \sum_{k=0}^{n-1} ar^k.$$

A formula for the nth partial sum of a geometric series is

$$S_n = a\frac{1-r^n}{1-r}.$$

If $|r| < 1$, the infinite geometric series $\sum_{k=0}^{\infty} ar^k$ has the finite sum $\frac{a}{1-r}$.

This page titled 8.2: Geometric Series is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.2: Geometric Series by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

8.3: Series of Real Numbers

Motivating Questions

- What is an infinite series?
- What is the nth partial sum of an infinite series?
- How do we add up an infinite number of numbers? In other words, what does it mean for an infinite series of real numbers to converge?
- What does it mean for an infinite series of real numbers to diverge?

In Section 8.2, we encountered infinite geometric series. For example, by writing

$$N = 0.1212121212\cdots = \frac{12}{100} + \frac{12}{100}\cdot\frac{1}{100} + \frac{12}{100}\cdot\frac{1}{100^2} + \cdots$$

as a geometric series, we found a way to write N as a single fraction: $N = \frac{4}{33}$. In this section, we explore other types of infinite sums. In Preview Activity 8.3.1 we see how one such sum is related to the famous number e.

Preview Activity 8.3.1

Have you ever wondered how your calculator can produce a numeric approximation for complicated numbers like e, π or $\ln(2)$? After all, the only operations a calculator can really perform are addition, subtraction, multiplication, and division, the operations that make up polynomials. This activity provides the first steps in understanding how this process works. Throughout the activity, let $f(x) = e^x$.

1. Find the tangent line to f at $x = 0$ and use this linearization to approximate e. That is, find a formula $L(x)$ for the tangent line, and compute $L(1)$, since $L(1) \approx f(1) = e$.
2. The linearization of e^x does not provide a good approximation to e since 1 is not very close to 0. To obtain a better approximation, we alter our approach a bit. Instead of using a straight line to approximate e, we put an appropriate bend in our estimating function to make it better fit the graph of e^x for x close to 0. With the linearization, we had both $f(x)$ and $f'(x)$ share the same value as the linearization at $x = 0$. We will now use a quadratic approximation $P_2(x)$ to $f(x) = e^x$ centered at $x = 0$ which has the property that $P_2(0) = f(0)$, $P_2'(0) = f'(0)$, and $P_2''(0) = f''(0)$.
    1. Let $P_2(x) = 1 + x + \frac{x^2}{2}$. Show that $P_2(0) = f(0)$, $P_2'(0) = f'(0)$, and $P_2''(0) = f''(0)$. Then, use $P_2(x)$ to approximate e by observing that $P_2(1) \approx f(1)$.
    2. We can continue approximating e with polynomials of larger degree whose higher derivatives agree with those of f at 0. This turns out to make the polynomials fit the graph of f better for more values of x around 0. For example, let $P_3(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$. Show that $P_3(0) = f(0)$, $P_3'(0) = f'(0)$, $P_3''(0) = f''(0)$, and $P_3'''(0) = f'''(0)$. Use $P_3(x)$ to approximate e in a way similar to how you did so with $P_2(x)$ above.

Infinite Series

Preview Activity 8.3.1 shows that an approximation to e using a linear polynomial is 2, an approximation to e using a quadratic polynomial is 2.5, and an approximation using a cubic polynomial is 2.6667. If we continue this process we can obtain approximations from quartic (degree 4) and quintic (degree 5) polynomials, giving us the following approximations to e:

linear	$1 + 1$	2
quadratic	$1 + 1 + \frac{1}{2}$	2.5
cubic	$1 + 1 + \frac{1}{2} + \frac{1}{6}$	$2.\overline{6}$
quartic	$1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24}$	$2.708\overline{3}$

quintic	$1+1+\frac{1}{2}+\frac{1}{6}+\frac{1}{24}+\frac{1}{120}$	$2.71\overline{6}$

We see an interesting pattern here. The number e is being approximated by the sum

$$1+1+\frac{1}{2}+\frac{1}{6}+\frac{1}{24}+\frac{1}{120}+\cdots+\frac{1}{n!} \tag{8.3.1}$$

for increasing values of n.

We can use summation notation as a shorthand [1] for writing the sum in Equation (8.3.1) to get

$$e \approx 1+1+\frac{1}{2}+\frac{1}{6}+\frac{1}{24}+\frac{1}{120}+\cdots+\frac{1}{n!}=\sum_{k=0}^{n}\frac{1}{k!}. \tag{8.3.2}$$

Note that $0!$ appears in Equation (8.3.2). By definition, $0!=1$.

We can calculate this sum using as large an n as we want, and the larger n is the more accurate the approximation (8.3.2) is. This suggests that we can write the number e as the infinite sum

$$e=\sum_{k=0}^{\infty}\frac{1}{k!}. \tag{8.3.3}$$

This sum is an example of an *infinite series*. Note that the series (8.3.3) is the sum of the terms of the infinite sequence $\left\{\frac{1}{n!}\right\}$. In general, we use the following notation and terminology.

Definition 8.3.1

An ***infinite series*** of real numbers is the sum of the entries in an infinite sequence of real numbers. In other words, an infinite series is sum of the form

$$a_1+a_2+\cdots+a_n+\cdots=\sum_{k=1}^{\infty}a_k,$$

where $a_1, a_2, \ldots,$ are real numbers.

We use summation notation to identify a series. If the series adds the entries of a sequence $\{a_n\}_{n\geq 1}$, we will write the series as

$$\sum_{k\geq 1}a_k$$

or

$$\sum_{k=1}^{\infty}a_k.$$

Each of these notations is simply shorthand for the infinite sum $a_1+a_2+\cdots+a_n+\cdots$.

Is it even possible to sum an infinite list of numbers? That doing so is possible in some situations shouldn't come as a surprise. We have already examined the special case of geometric series in the previous section. Moreover, the definite integral is defineed as the limit of a sequence of finite sums, which provides insight into how we'll think of infinite sums generally. As we investigate other infinite sums, there are two key questions we seek to answer: (1) is the sum finite? and (2) if yes, what is its value?

Activity 8.3.2

Consider the series

$$\sum_{k=1}^{\infty} \frac{1}{k^2}.$$

While it is physically impossible to add an infinite collection of numbers, we can, of course, add any finite collection of them. In what follows, we investigate how understanding how to find the nth partial sum (that is, the sum of the first n terms) enables us to make sense of the infinite sum.

1. Sum the first two numbers in this series. That is, find a numeric value for
$$\sum_{k=1}^{2} \frac{1}{k^2}$$
2. Next, add the first three numbers in the series.
3. Continue adding terms in this series to complete the list below. Carry each sum to at least 8 decimal places.
   - $\sum_{k=1}^{1} \frac{1}{k^2} = 1$
   - $\sum_{k=1}^{2} \frac{1}{k^2} =$
   - $\sum_{k=1}^{3} \frac{1}{k^2} =$
   - $\sum_{k=1}^{4} \frac{1}{k^2} =$
   - $\sum_{k=1}^{5} \frac{1}{k^2} =$
   - $\sum_{k=1}^{6} \frac{1}{k^2} =$
   - $\sum_{k=1}^{7} \frac{1}{k^2} =$
   - $\sum_{k=1}^{8} \frac{1}{k^2} =$
   - $\sum_{k=1}^{9} \frac{1}{k^2} =$
   - $\sum_{k=1}^{10} \frac{1}{k^2} =$
4. The sums in the table in part c form a sequence whose nth term is $S_n = \sum_{k=1}^{n} \frac{1}{k^2}$. Based on your calculations in the table, do you think the sequence $\{S_n\}$ converges or diverges? Explain. How do you think this sequence $\{S_n\}$ is related to the series $\sum_{k=1}^{\infty} \frac{1}{k^2}$?

The example in Activity 8.3.2 illustrates how we define the sum of an infinite series. We construct a new sequence of numbers (called the *sequence of partial sums*) where the nth term in the sequence consists of the sum of the first n terms of the series. If this sequence converges, the corresponding infinite series is said to converge, and we say that we can find the sum of the series. More formally, we have the following definition.

Definition 8.3.2

The ***nth partial sum*** of the series $\sum_{k=1}^{\infty} a_k$ is the finite sum $S_n = \sum_{k=1}^{n} a_k$.

In other words, the nth partial sum S_n of a series is the sum of the first n terms in the series,

$$S_n = a_1 + a_2 + \cdots + a_n.$$

We investigate the behavior of the series by examining the sequence

$$S_1, S_2, \ldots, S_n, \ldots$$

of its partial sums. If the sequence of partial sums converges to some finite number, we say that the corresponding series *converges*. Otherwise, we say the series *diverges*. From our work in Activity 8.3.2, the series

$$\sum_{k=1}^{\infty} \frac{1}{k^2}$$

appears to converge to some number near 1.54977. We formalize the concept of convergence and divergence of an infinite series in the following definition.

Definition 8.3.3

The infinite series

$$\sum_{k=1}^{\infty} a_k$$

converges (or is ***convergent***) if the sequence $\{S_n\}$ of partial sums converges, where

$$S_n = \sum_{k=1}^{n} a_k.$$

If $\lim_{n\to\infty} S_n = S$, then we call S the sum of the series $\sum_{k=1}^{\infty} a_k$. That is,

$$\sum_{k=1}^{\infty} a_k = \lim_{n\to\infty} S_n = S.$$

If the sequence of partial sums does not converge, then the series

$$\sum_{k=1}^{\infty} a_k$$

diverges (or is ***divergent***).

The early terms in a series do not influence whether or not the series converges or diverges. Rather, the convergence or divergence of a series

$$\sum_{k=1}^{\infty} a_k$$

is determined by what happens to the terms a_k for very large values of k. To see why, suppose that m is some constant larger than 1. Then

$$\sum_{k=1}^{\infty} a_k = (a_1 + a_2 + \cdots + a_m) + \sum_{k=m+1}^{\infty} a_k.$$

Since $a_1 + a_2 + \cdots + a_m$ is a finite number, the series $\sum_{k=1}^{\infty} a_k$ will converge if and only if the series $\sum_{k=m+1}^{\infty} a_k$ converges. Because the starting index of the series doesn't affect whether the series converges or diverges, we will often just write

$$\sum a_k$$

when we are interested in questions of convergence/divergence as opposed to the exact sum of a series.

In Section 8.2 we encountered the special family of infinite geometric series. Recall that a geometric series has the form $\sum_{k=0}^{\infty} ar^k$, where a and r are real numbers (and $r \neq 1$). We found that the nth partial sum S_n of a geometric series is given by the convenient formula

$$S_n = \frac{1 - r^n}{1 - r},$$

and thus a geometric series converges if $|r| < 1$. Geometric series diverge for all other values of r.

It is generally a difficult question to determine if a given nongeometric series converges or diverges. There are several tests we can use that we will consider in the following sections.

The Divergence Test

The first question we ask about any infinite series is usually "Does the series converge or diverge?" There is a straightforward way to check that certain series diverge, and we explore this test in the next activity.

Activity 8.3.3

If the *series* $\sum a_k$ converges, then an important result necessarily follows regarding the *sequence* $\{a_n\}$. This activity explores this result.

Assume that the series $\sum_{k=1}^{\infty} a_k$ converges and has sum equal to L.

1. What is the nth partial sum S_n of the series $\sum_{k=1}^{\infty} a_k$?
2. What is the $(n-1)$st partial sum S_{n-1} of the series $\sum_{k=1}^{\infty} a_k$?
3. What is the difference between the nth partial sum and the $(n-1)$st partial sum of the series $\sum_{k=1}^{\infty} a_k$?
4. Since we are assuming that $\sum_{k=1}^{\infty} a_k = L$, what does that tell us about $\lim_{n\to\infty} S_n$? Why? What does that tell us about $\lim_{n\to\infty} S_{n-1}$? Why?
5. Combine the results of the previous two parts of this activity to determine $\lim_{n\to\infty} a_n = \lim_{n\to\infty}(S_n - S_{n-1})$.

The result of Activity 8.3.3 is the following important conditional statement:

> If the series $\sum_{k=1}^{\infty} a_k$ converges, then the sequence $\{a_k\}$ of kth terms converges to 0.

It is logically equivalent to say that if the sequence $\{a_k\}$ does not converge to 0, then the series $\sum_{k=1}^{\infty} a_k$ cannot converge. This statement is called the Divergence Test.

The Divergence Test

If $\lim_{k\to\infty} a_k \neq 0$, then the series $\sum a_k$ diverges.

Activity 8.3.4

Determine if the Divergence Test applies to the following series. If the test does not apply, explain why. If the test does apply, what does it tell us about the series?

1. $\sum \dfrac{k}{k+1}$
2. $\sum (-1)^k$
3. $\sum \dfrac{1}{k}$

Note well: be very careful with the Divergence Test. This test only tells us what happens to a series if the terms of the corresponding sequence do not converge to 0. If the sequence of the terms of the series does converge to 0, the Divergence Test does not apply: indeed, as we will soon see, a series whose terms go to zero may either converge or diverge.

The Integral Test

The Divergence Test settles the questions of divergence or convergence of series $\sum a_k$ in which $\lim_{k\to\infty} a_k \neq 0$. Determining the convergence or divergence of series $\sum a_k$ in which $\lim_{k\to\infty} a_k = 0$ turns out to be more complicated. Often, we have to investigate

the sequence of partial sums or apply some other technique.

Next, we consider the ***harmonic series*** [2]

$$\sum_{k=1}^{\infty} \frac{1}{k}.$$

The first 9 partial sums of this series are shown following.

$$\sum_{k=1}^{1} \frac{1}{k} = 1 \qquad \sum_{k=1}^{4} \frac{1}{k} = 2.083333333 \qquad \sum_{k=1}^{7} \frac{1}{k} = 2.592857143$$

$$\sum_{k=1}^{2} \frac{1}{k} = 1.5 \qquad \sum_{k=1}^{5} \frac{1}{k} = 2.283333333 \qquad \sum_{k=1}^{8} \frac{1}{k} = 2.717857143$$

$$\sum_{k=1}^{3} \frac{1}{k} = 1.833333333 \qquad \sum_{k=1}^{6} \frac{1}{k} = 2.450000000 \qquad \sum_{k=1}^{9} \frac{1}{k} = 2.828968254$$

This series is called harmonic because each term in the series after the first is the harmonic mean of the term before it and the term after it. The harmonic mean of two numbers a and b is $\frac{2ab}{a+b}$. See "What's Harmonic about the Harmonic Series", by David E. Kullman (in the College Mathematics Journal, Vol. 32, No. 3 (May, 2001), 201-203) for an interesting discussion of the harmonic mean.

This information alone doesn't seem to be enough to tell us if the series $\sum_{k=1}^{\infty} \frac{1}{k}$ converges or diverges. The partial sums could eventually level off to some fixed number or continue to grow without bound. Even if we look at larger partial sums, such as $\sum_{n=1}^{1000} \frac{1}{k} \approx 7.485470861$, the result isn't particularly convincing one way or another. The Integral Test is one way to determine whether or not the harmonic series converges, and we explore this test further in the next activity.

Activity 8.3.5

Consider the harmonic series $\sum_{k=1}^{\infty} \frac{1}{k}$. Recall that the harmonic series will converge provided that its sequence of partial sums converges. The nth partial sum S_n of the series $\sum_{k=1}^{\infty} \frac{1}{k}$ is

$$\begin{aligned} S_n &= \sum_{k=1}^{n} \frac{1}{k} \\ &= 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \\ &= 1(1) + (1)\left(\frac{1}{2}\right) + (1)\left(\frac{1}{3}\right) + \cdots + (1)\left(\frac{1}{n}\right). \end{aligned}$$

Through this last expression for S_n, we can visualize this partial sum as a sum of areas of rectangles with heights $\frac{1}{m}$ and bases of length 1, as shown in Figure 8.3.1, which uses the 9th partial sum.

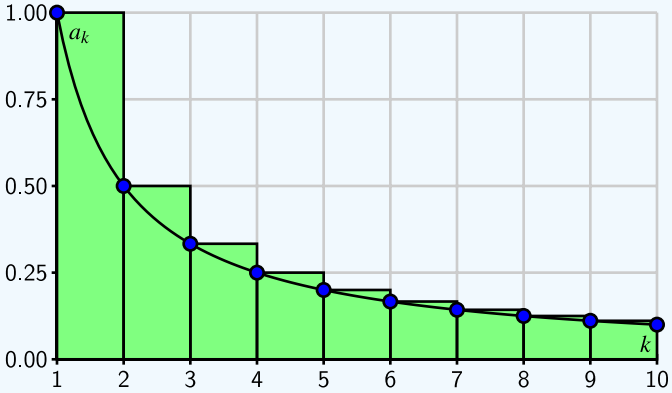


Figure 8.3.1. A picture of the 9th partial sum of the harmonic series as a sum of areas of rectangles.

The graph of the continuous function f defined by $f(x) = \frac{1}{x}$ is overlaid on this plot.

1. Explain how this picture represents a particular Riemann sum.
2. What is the definite integral that corresponds to the Riemann sum you considered in (a)?
3. Which is larger, the definite integral in (b), or the corresponding partial sum S_9 of the series? Why?
4. If instead of considering the 9th partial sum, we consider the nth partial sum, and we let n go to infinity, we can then compare the series $\sum_{k=1}^{\infty} \frac{1}{k}$ to the improper integral $\int_1^{\infty} \frac{1}{x}\ dx$. Which of these quantities is larger? Why?
5. Does the improper integral $\int_1^{\infty} \frac{1}{x}\ dx$ converge or diverge? What does that result, together with your work in (d), tell us about the series $\sum_{k=1}^{\infty} \frac{1}{k}$?

The ideas from Activity 8.3.5 hold more generally. Suppose that f is a continuous decreasing function and that $a_k = f(k)$ for each value of k. Consider the corresponding series $\sum_{k=1}^{\infty} a_k$. The partial sum

$$S_n = \sum_{k=1}^{n} a_k$$

can always be viewed as a left hand Riemann sum of $f(x)$, using rectangles with bases of width 1 and heights given by the values a_k. A representative picture is shown at left in Figure 8.3.2. Since f is a decreasing function, we have that

$$S_n > \int_1^n f(x)\ dx.$$

Taking the limit as n goes to infinity shows that

$$\sum_{k=1}^{\infty} a_k > \int_1^{\infty} f(x)\ dx.$$

Therefore, if the improper integral $\int_1^{\infty} f(x)\ dx$ diverges, so does the series $\sum_{k=1}^{\infty} a_k$.

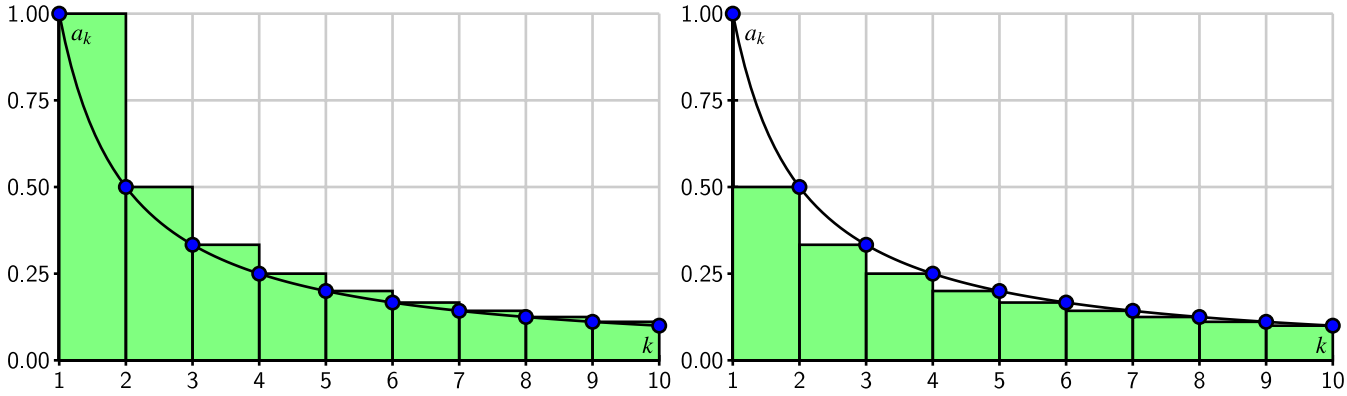


Figure 8.3.2. Comparing an improper integral to a series

What's more, if we look at the right hand Riemann sums of f on $[1, n]$ as shown at right in Figure 8.3.5, we see that

$$\int_1^{\infty} f(x)\ dx > \sum_{k=2}^{\infty} a_k.$$

So if $\int_1^{\infty} f(x)\ dx$ converges, then so does $\sum_{k=2}^{\infty} a_k$, which also means that the series $\sum_{k=1}^{\infty} a_k$ also converges. Our preceding discussion has demonstrated the truth of the Integral Test.

The Integral Test

Let f be a real valued function and assume f is decreasing and positive for all x larger than some number c. Let $a_k = f(k)$ for each positive integer k.

1. If the improper integral $\int_c^\infty f(x)\ dx$ converges, then the series $\sum_{k=1}^\infty a_k$ converges.
2. If the improper integral $\int_c^\infty f(x)\ dx$ diverges, then the series $\sum_{k=1}^\infty a_k$ diverges.

The Integral Test compares a given infinite series to a natural, corresponding improper integral and says that the infinite series and corresponding improper integral both have the same convergence status. In the next activity, we apply the Integral Test to determine the convergence or divergence of a class of important series.

Activity 8.3.6

The series $\sum \frac{1}{k^p}$ are special series called p-series. We have already seen that the p-series with $p = 1$ (the harmonic series) diverges. We investigate the behavior of other p-series in this activity.

1. Evaluate the improper integral $\int_1^\infty \frac{1}{x^2}\ dx$. Does the series $\sum_{k=1}^\infty \frac{1}{k^2}$ converge or diverge? Explain.
2. Evaluate the improper integral $\int_1^\infty \frac{1}{x^p}\ dx$ where $p > 1$. For which values of p can we conclude that the series $\sum_{k=1}^\infty \frac{1}{k^p}$ converges?
3. Evaluate the improper integral $\int_1^\infty \frac{1}{x^p}\ dx$ where $p < 1$. What does this tell us about the corresponding p-series $\sum_{k=1}^\infty \frac{1}{k^p}$?
4. Summarize your work in this activity by completing the following statement.

   The p-series $\sum_{k=1}^\infty \frac{1}{k^p}$ converges if and only if .

The Limit Comparison Test

The Integral Test allows us to determine the convergence of an entire family of series: the p-series. However, we have seen that it is often difficult to integrate functions, so the Integral Test is not one that we can use all of the time. In fact, even for a relatively simple series such as $\sum \frac{k^2+1}{k^4+2k+2}$, the Integral Test is not an option. In what follows we develop a test that applies to series of rational functions by comparing their behavior to the behavior of p-series.

Activity 8.3.7

Consider the series $\sum \frac{k+1}{k^3+2}$. Since the convergence or divergence of a series only depends on the behavior of the series for large values of k, we might examine the terms of this series more closely as k gets large.

1. By computing the value of $\frac{k+1}{k^3+2}$ for $k = 100$ and $k = 1000$, explain why the terms $\frac{k+1}{k^3+2}$ are essentially $\frac{k}{k^3}$ when k is large.
2. Let's formalize our observations in (a) a bit more. Let $a_k = \frac{k+1}{k^3+2}$ and $b_k = \frac{k}{k^3}$. Calculate

   $$\lim_{k\to\infty} \frac{a_k}{b_k}.$$

   What does the value of the limit tell you about a_k and b_k for large values of k? Compare your response from part (a).

3. Does the series $\sum \frac{k}{k^3}$ converge or diverge? Why? What do you think that tells us about the convergence or divergence of the series $\sum \frac{k+1}{k^3+2}$? Explain.

Activity 8.3.7 illustrates how we can compare one series with positive terms to another whose convergence status we know. Suppose we have two series $\sum a_k$ and $\sum b_k$ with positive terms and we know the convergence status of the series $\sum a_k$. Recall that the convergence or divergence of a series depends only on the terms of the series for large values of k, so if we know that a_k and b_k are proportional for large k, then the two series $\sum a_k$ and $\sum b_k$ should behave the same way. In other words, if there is a positive finite constant c such that

$$\lim_{k\to\infty} \frac{b_k}{a_k} = c,$$

then $b_k \approx ca_k$ for large values of k. So

$$\sum b_k \approx \sum ca_k = c\sum a_k.$$

Since multiplying by a nonzero constant does not affect the convergence or divergence of a series, it follows that the series $\sum a_k$ and $\sum b_k$ either both converge or both diverge. The formal statement of this fact is called the Limit Comparison Test.

The Limit Comparison Test

Let $\sum a_k$ and $\sum b_k$ be series with positive terms. If

$$\lim_{k\to\infty} \frac{b_k}{a_k} = c$$

for some positive (finite) constant c, then $\sum a_k$ and $\sum b_k$ either both converge or both diverge.

The Limit Comparison Test shows that if we have a series $\sum \frac{p(k)}{q(k)}$ of rational functions where $p(k)$ is a polynomial of degree m and $q(k)$ a polynomial of degree l, then the series $\sum \frac{p(k)}{q(k)}$ will behave like the series $\sum \frac{k^m}{k^l}$. So this test allows us to determine the convergence or divergence of series whose terms are rational functions.

Activity 8.3.8

Use the Limit Comparison Test to determine the convergence or divergence of the series

$$\sum \frac{3k^2+1}{5k^4+2k+2}.$$

by comparing it to the series $\sum \frac{1}{k^2}$.

The Ratio Test

The Limit Comparison Test works well if we can find a series with known behavior to compare. But such series are not always easy to find. In this section we will examine a test that allows us to examine the behavior of a series by comparing it to a geometric series, without knowing in advance which geometric series we need.

Activity 8.3.9

Consider the series defined by

$$\sum_{k=1}^{\infty} \frac{2^k}{3^k-k}. \tag{8.3.4}$$

This series is not a geometric series, but this activity will illustrate how we might compare this series to a geometric one. Recall that a series $\sum a_k$ is geometric if the ratio $\frac{a_{k+1}}{a_k}$ is always the same. For the series in (8.3.4), note that $a_k = \frac{2^k}{3^k-k}$.

1. To see if $\sum \frac{2^k}{3^k-k}$ is comparable to a geometric series, we analyze the ratios of successive terms in the series. Complete Table 8.3.6, listing your calculations to at least 8 decimal places.

   Table 8.3.1. Ratios of successive terms in the series $\sum \frac{2^k}{3^k-k}$

k	5	10	20	21	22	23	24	25

$\frac{a_{k+1}}{a_k}$								

2. Based on your calculations in Table 8.3.1, what can we say about the ratio $\frac{a_{k+1}}{a_k}$ if k is large?
3. Do you agree or disagree with the statement: "the series $\sum \frac{2^k}{3^k-k}$ is approximately geometric when k is large"? If not, why not? If so, do you think the series $\sum \frac{2^k}{3^k-k}$ converges or diverges? Explain.

We can generalize the argument in Activity 8.3.9 in the following way. Consider the series $\sum a_k$. If

$$\frac{a_{k+1}}{a_k} \approx r$$

for large values of k, then $a_{k+1} \approx ra_k$ for large k and the series $\sum a_k$ is approximately the geometric series $\sum ar^k$ for large k. Since the geometric series with ratio r converges only for $-1 < r < 1$, we see that the series $\sum a_k$ will converge if

$$\lim_{k\to\infty} \frac{a_{k+1}}{a_k} = r$$

for a value of r such that $|r| < 1$. This result is known as the Ratio Test.

The Ratio Test

Let $\sum a_k$ be an infinite series. Suppose

$$\lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|} = r.$$

1. If $0 \le r < 1$, then the series $\sum a_k$ converges.
2. If $1 < r$, then the series $\sum a_k$ diverges.
3. If $r = 1$, then the test is inconclusive.

Note well: The Ratio Test looks at the limit of the ratio of consecutive terms of a given series; in so doing, the test is asking, "is this series approximately geometric?" If so, the test uses the limit of the ratio of consecutive terms to determine if the given series converges.

We have now encountered several tests for determining convergence or divergence of series.

- The Divergence Test can be used to show that a series diverges, but never to prove that a series converges.
- We used the Integral Test to determine the convergence status of an entire class of series, the p-series.
- The Limit Comparison Test works well for series that involve rational functions and which can therefore by compared to p -series.
- Finally, the Ratio Test allows us to compare our series to a geometric series; it is particularly useful for series that involve n th powers and factorials.
- Two other tests, the Direct Comparison Test and the Root Test, are discussed in the exercises.

One of the challenges of determining whether a series converges or diverges is finding which test answers that question.

Activity 8.3.10

Determine whether each of the following series converges or diverges. Explicitly state which test you use.

1. $\sum \frac{k}{2^k}$
2. $\sum \frac{k^3+2}{k^2+1}$
3. $\sum \frac{10^k}{k!}$

4. $\sum \dfrac{k^3 - 2k^2 + 1}{k^6 + 4}$

Summary

- An infinite series is a sum of the elements in an infinite sequence. In other words, an infinite series is a sum of the form

$$a_1 + a_2 + \cdots + a_n + \cdots = \sum_{k=1}^{\infty} a_k$$

  where a_k is a real number for each positive integer k.

- The nth partial sum S_n of the series $\sum_{k=1}^{\infty} a_k$ is the sum of the first n terms of the series. That is,

$$S_n = a_1 + a_2 + \cdots + a_n = \sum_{k=1}^{n} a_k.$$

- The sequence $\{S_n\}$ of partial sums of a series $\sum_{k=1}^{\infty} a_k$ tells us about the convergence or divergence of the series. In particular
  - The series $\sum_{k=1}^{\infty} a_k$ converges if the sequence $\{S_n\}$ of partial sums converges. In this case we say that the series is the limit of the sequence of partial sums and write

$$\sum_{k=1}^{\infty} a_k = \lim_{n\to\infty} S_n.$$

  - The series $\sum_{k=1}^{\infty} a_k$ diverges if the sequence $\{S_n\}$ of partial sums diverges.

This page titled 8.3: Series of Real Numbers is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.3: Series of Real Numbers by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

8.4: Alternating Series

Motivating Questions

- What is an alternating series?
- Under what conditions does an alternating series converge? Why?
- How well does the nth partial sum of a convergent alternating series approximate the actual sum of the series? Why?

So far, we've considered series with exclusively nonnegative terms. Next, we consider series that have some negative terms. For instance, the geometric series

$$2 - \frac{4}{3} + \frac{8}{9} - \cdots + 2\left(-\frac{2}{3}\right)^n + \cdots,$$

has $a = 2$ and $r = -\frac{2}{3}$, so that every other term alternates in sign. This series converges to

$$S = \frac{a}{1-r} = \frac{2}{1-\left(-\frac{2}{3}\right)} = \frac{6}{5}.$$

In Preview Activity $8.4.1$ and our following discussion, we investigate the behavior of similar series where consecutive terms have opposite signs.

Preview Activity $8.4.1$

Preview Activity 8.3.1 showed how we can approximate the number e with linear, quadratic, and other polynomial approximations. We use a similar approach in this activity to obtain linear and quadratic approximations to $\ln(2)$. Along the way, we encounter a type of series that is different than most of the ones we have seen so far. Throughout this activity, let $f(x) = \ln(1+x)$.

1. Find the tangent line to f at $x = 0$ and use this linearization to approximate $\ln(2)$. That is, find $L(x)$, the tangent line approximation to $f(x)$, and use the fact that $L(1) \approx f(1)$ to estimate $\ln(2)$.
2. The linearization of $\ln(1+x)$ does not provide a very good approximation to $\ln(2)$ since 1 is not that close to 0. To obtain a better approximation, we alter our approach; instead of using a straight line to approximate $\ln(2)$, we use a quadratic function to account for the concavity of $\ln(1+x)$ for x close to 0. With the linearization, both the function's value and slope agree with the linearization's value and slope at $x = 0$. We will now make a quadratic approximation $P_2(x)$ to $f(x) = \ln(1+x)$ centered at $x = 0$ with the property that $P_2(0) = f(0)$, $P_2'(0) = f'(0)$, and $P_2''(0) = f''(0)$.
    1. Let $P_2(x) = x - \frac{x^2}{2}$. Show that $P_2(0) = f(0)$, $P_2'(0) = f'(0)$, and $P_2''(0) = f''(0)$. Use $P_2(x)$ to approximate $\ln(2)$ by using the fact that $P_2(1) \approx f(1)$.
    2. We can continue approximating $\ln(2)$ with polynomials of larger degree whose derivatives agree with those of f at 0. This makes the polynomials fit the graph of f better for more values of x around 0. For example, let $P_3(x) = x - \frac{x^2}{2} + \frac{x^3}{3}$. Show that $P_3(0) = f(0)$, $P_3'(0) = f'(0)$, $P_3''(0) = f''(0)$, and $P_3'''(0) = f'''(0)$. Taking a similar approach to preceding questions, use $P_3(x)$ to approximate $\ln(2)$.
    3. If we used a degree 4 or degree 5 polynomial to approximate $\ln(1+x)$, what approximations of $\ln(2)$ do you think would result? Use the preceding questions to conjecture a pattern that holds, and state the degree 4 and degree 5 approximation.

The Alternating Series Test

Preview Activity $8.4.1$ gives us several approximations to $\ln(2)$. The linear approximation is 1, and the quadratic approximation is $1 - \frac{1}{2} = \frac{1}{2}$. If we continue this process, cubic, quartic (degree 4), quintic (degree 5), and higher degree polynomials give us the approximations to $\ln(2)$ in Table $8.4.1$.

Table 8.4.1.

linear	1	1
quadratic	$1-\frac{1}{2}$	0.5
cubic	$1-\frac{1}{2}+\frac{1}{3}$	$0.8\overline{3}$
quartic	$1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}$	$0.58\overline{3}$
quintic	$1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}$	$0.78\overline{3}$

The pattern here shows that $\ln(2)$ can be approximated by the partial sums of the infinite series

$$\sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k} \tag{8.4.1}$$

where the alternating signs are indicated by the factor $(-1)^{k+1}$. We call such a series an *alternating series*.

Using computational technology, we find that the sum of the first 100 terms in this series is 0.6881721793. As a comparison, $\ln(2) \approx 0.6931471806$. This shows that even though the series (8.4.1) converges to $\ln(2)$, it must do so quite slowly, since the sum of the first 100 terms isn't particularly close to $\ln(2)$. We will investigate the issue of how quickly an alternating series converges later in this section.

Definition 8.4.1

An alternating series is a series of the form

$$\sum_{k=0}^{\infty}(-1)^k a_k,$$

where $a_k > 0$ for each k.

We have some flexibility in how we write an alternating series; for example, the series

$$\sum_{k=1}^{\infty}(-1)^{k+1} a_k,$$

whose index starts at $k=1$, is also alternating. As we will soon see, there are several very nice results that hold for alternating series, while alternating series can also demonstrate some unusual behaivior.

It is important to remember that most of the series tests we have seen in previous sections apply only to series with nonnegative terms. Alternating series require a different test.

Activity 8.4.2

Remember that, by definition, a series converges if and only if its corresponding sequence of partial sums converges.

1. Calculate the first few partial sums (to 10 decimal places) of the alternating series
$$\sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k}.$$
Label each partial sum with the notation $S_n = \sum_{k=1}^{n}(-1)^{k+1}\frac{1}{k}$ for an appropriate choice of n.
2. Plot the sequence of partial sums from part (a). What do you notice about this sequence?

Activity 8.4.2 illustrates the general behavior of any convergent alternating series. We see that the partial sums of the alternating

harmonic series oscillate around a fixed number that turns out to be the sum of the series.

Recall that if $\lim_{k\to\infty} a_k \neq 0$, then the series $\sum a_k$ diverges by the Divergence Test. From this point forward, we will thus only consider alternating series

$$\sum_{k=1}^{\infty} (-1)^{k+1} a_k$$

in which the sequence a_k consists of positive numbers that decrease to 0. The nth partial sum S_n is

$$S_n = \sum_{k=1}^{n} (-1)^{k+1} a_k.$$

Notice that

- $S_2 = a_1 - a_2$, and since $a_1 > a_2$ we have $0 < S_2 < S_1$.
- $S_3 = S_2 + a_3$ and so $S_2 < S_3$. But $a_3 < a_2$, so $S_3 < S_1$. Thus, $0 < S_2 < S_3 < S_1$.
- $S_4 = S_3 - a_4$ and so $S_4 < S_3$. But $a_4 < a_3$, so $S_2 < S_4$. Thus, $0 < S_2 < S_4 < S_3 < S_1$.
- $S_5 = S_4 + a_5$ and so $S_4 < S_5$. But $a_5 < a_4$, so $S_5 < S_3$. Thus, $0 < S_2 < S_4 < S_5 < S_3 < S_1$.

This pattern continues as illustrated in Figure $8.4.1$ (with n odd) so that each partial sum lies between the previous two partial sums.

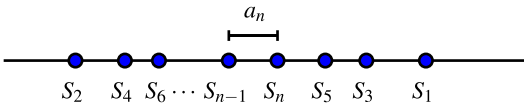


Figure $8.4.1$. Partial sums of an alternating series

Note further that the absolute value of the difference between the $(n-1)$st partial sum S_{n-1} and the nth partial sum S_n is

$$|S_n - S_{n-1}| = a_n.$$

Because the sequence $\{a_n\}$ converges to 0, the distance between successive partial sums becomes as close to zero as we'd like, and thus the sequence of partial sums converges (even though we don't know the exact value to which it converges).

The preceding discussion has demonstrated the truth of the Alternating Series Test.

The Alternating Series Test

Given an alternating series $\sum (-1)^k a_k$, if the sequence $\{a_k\}$ of positive terms decreases to 0 as $k \to \infty$, then the alternating series converges.

Note that if the limit of the sequence $\{a_k\}$ is not 0, then the alternating series diverges.

Activity $8.4.3$

Which series converge and which diverge? Justify your answers.

1. $\displaystyle\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2+2}$
2. $\displaystyle\sum_{k=1}^{\infty} \frac{(-1)^{k+1} 2k}{k+5}$

3. $\displaystyle\sum_{k=2}^{\infty} \frac{(-1)^k}{\ln(k)}$

Estimating Alternating Sums

If the series converges, the argument for the Alternating Series Test also provides us with a method to determine how close the n th partial sum S_n is to the actual sum of the series. To see how this works, let S be the sum of a convergent alternating series, so

$$S = \sum_{k=1}^{\infty} (-1)^k a_k.$$

Recall that the sequence of partial sums oscillates around the sum S so that

$$|S - S_n| < |S_{n+1} - S_n| = a_{n+1}.$$

Therefore, the value of the term a_{n+1} provides an error estimate for how well the partial sum S_n approximates the actual sum S. We summarize this fact in the statement of the Alternating Series Estimation Theorem.

Alternating Series Estimation Theorem

If the alternating series $\sum_{k=1}^{\infty} (-1)^{k+1} a_k$ has positive terms a_k that decrease to zero as $k \to \infty$, and $S_n = \sum_{k=1}^{n} (-1)^{k+1} a_k$ is the nth partial sum of the alternating series, then

$$\left| \sum_{k=1}^{\infty} (-1)^{k+1} a_k - S_n \right| \leq a_{n+1}.$$

Example 8.4.1

Determine how well the 100th partial sum S_{100} of

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}$$

approximates the sum of the series.

Answer available at the end of this section.

Activity 8.4.4

Determine the number of terms it takes to approximate the sum of the convergent alternating series

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^4}$$

to within 0.0001.

Absolute and Conditional Convergence

A series such as

$$1 - \frac{1}{4} - \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \frac{1}{36} - \frac{1}{49} - \frac{1}{64} - \frac{1}{81} - \frac{1}{100} + \cdots \tag{8.4.2}$$

whose terms are neither all nonnegative nor alternating is different from any series that we have considered so far. The behavior of such a series can be rather complicated, but there is an important connection between a series with some negative terms and series with all positive terms.

Activity 8.4.5

1. Explain why the series
$$1-\frac{1}{4}-\frac{1}{9}+\frac{1}{16}+\frac{1}{25}+\frac{1}{36}-\frac{1}{49}-\frac{1}{64}-\frac{1}{81}-\frac{1}{100}+\cdots$$
must have a sum that is less than the series
$$\sum_{k=1}^{\infty}\frac{1}{k^2}.$$
2. Explain why the series
$$1-\frac{1}{4}-\frac{1}{9}+\frac{1}{16}+\frac{1}{25}+\frac{1}{36}-\frac{1}{49}-\frac{1}{64}-\frac{1}{81}-\frac{1}{100}+\cdots$$
must have a sum that is greater than the series
$$\sum_{k=1}^{\infty}-\frac{1}{k^2}.$$
3. Given that the terms in the series
$$1-\frac{1}{4}-\frac{1}{9}+\frac{1}{16}+\frac{1}{25}+\frac{1}{36}-\frac{1}{49}-\frac{1}{64}-\frac{1}{81}-\frac{1}{100}+\cdots$$
converge to 0, what do you think the previous two results tell us about the convergence status of this series?

As the example in Activity 8.4.5 suggests, if a series $\sum a_k$ has some negative terms but $\sum |a_k|$ converges, then the original series, $\sum a_k$, must also converge. That is, if $\sum |a_k|$ converges, then so must $\sum a_k$.

As we just observed, this is the case for the series (8.4.2), because the corresponding series of the absolute values of its terms is the convergent p-series $\sum \frac{1}{k^2}$. But there are series, such as the alternating harmonic series $\sum (-1)^{k+1}\frac{1}{k}$, that converge while the corresponding series of absolute values, $\sum \frac{1}{k}$, diverges. We distinguish between these behaviors by introducing the following language.

Definition 8.4.6

Consider a series $\sum a_k$.

1. The series $\sum a_k$ ***converges absolutely*** (or is ***absolutely convergent***) provided that $\sum |a_k|$ converges.
2. The series $\sum a_k$ ***converges conditionally*** (or is ***conditionally convergent***) provided that $\sum |a_k|$ diverges and $\sum a_k$ converges.

In this terminology, the series (8.4.2) converges absolutely while the alternating harmonic series is conditionally convergent.

Activity 8.4.6

1. Consider the series $\sum (-1)^k \frac{\ln(k)}{k}$.
    1. Does this series converge? Explain.
    2. Does this series converge absolutely? Explain what test you use to determine your answer.
2. Consider the series $\sum (-1)^k \frac{\ln(k)}{k^2}$.
    1. Does this series converge? Explain.

2. Does this series converge absolutely? Hint: Use the fact that $\ln(k) < \sqrt{k}$ for large values of k and then compare to an appropriate p-series.

Conditionally convergent series turn out to be very interesting. If the sequence $\{a_n\}$ decreases to 0, but the series $\sum a_k$ diverges, the conditionally convergent series $\sum (-1)^k a_k$ is right on the borderline of being a divergent series. As a result, any conditionally convergent series converges very slowly. Furthermore, some very strange things can happen with conditionally convergent series, as illustrated in some of the exercises.

Summary of Tests for Convergence of Series

We have discussed several tests for convergence/divergence of series in our sections and in exercises. We close this section of the text with a summary of all the tests we have encountered, followed by an activity that challenges you to decide which convergence test to apply to several different series.

Geometric Series

The geometric series $\sum ar^k$ with ratio r converges for $-1 < r < 1$ and diverges for $|r| \geq 1$.

The sum of the convergent geometric series $\sum_{k=0}^{\infty} ar^k$ is $\frac{a}{1-r}$.

Divergence Test

If the sequence a_n does not converge to 0, then the series $\sum a_k$ diverges.

This is the first test to apply because the conclusion is simple. However, if $\lim_{n\to\infty} a_n = 0$, no conclusion can be drawn.

Integral Test

Let f be a positive, decreasing function on an interval $[c, \infty)$ and let $a_k = f(k)$ for each positive integer $k \geq c$.

- If $\int_c^{\infty} f(t)\ dt$ converges, then $\sum a_k$ converges.
- If $\int_c^{\infty} f(t)\ dt$ diverges, then $\sum a_k$ diverges.

Use this test when $f(x)$ is easy to integrate.

Direct Comp. Test

(see Ex 4 in Section 8.3)

Let $0 \leq a_k \leq b_k$ for each positive integer k.

- If $\sum b_k$ converges, then $\sum a_k$ converges.
- If $\sum a_k$ diverges, then $\sum b_k$ diverges.

Use this test when you have a series with known behavior that you can compare to — this test can be difficult to apply.

Limit Comp. Test

Let a_n and b_n be sequences of positive terms. If

$$\lim_{k\to\infty} \frac{a_k}{b_k} = L$$

for some positive finite number L, then the two series $\sum a_k$ and $\sum b_k$ either both converge or both diverge.

Easier to apply in general than the comparison test, but you must have a series with known behavior to compare.

Useful to apply to series of rational functions.

Ratio Test

Let $a_k \neq 0$ for each k and suppose

$$\lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|} = r.$$

- If $r < 1$, then the series $\sum a_k$ converges absolutely.
- If $r > 1$, then the series $\sum a_k$ diverges.
- If $r = 1$, then test is inconclusive.

This test is useful when a series involves factorials and powers.

Root Test

(see Exercise 2 in Section 8.3)

Let $a_k \geq 0$ for each k and suppose

$$\lim_{k\to\infty} \sqrt[k]{a_k} = r.$$

- If $r < 1$, then the series $\sum a_k$ converges.
- If $r > 1$, then the series $\sum a_k$ diverges.
- If $r = 1$, then test is inconclusive.

In general, the Ratio Test can usually be used in place of the Root Test. However, the Root Test can be quick to use when a_k involves kth powers.

Alt. Series Test

If a_n is a positive, decreasing sequence so that $\lim_{n\to\infty} a_n = 0$, then the alternating series $\sum (-1)^{k+1} a_k$ converges.

This test applies only to alternating series — we assume that the terms a_n are all positive and that the sequence $\{a_n\}$ is decreasing.

Alt. Series Est.

Let $S_n = \sum_{k=1}^{n} (-1)^{k+1} a_k$ be the nth partial sum of the alternating series $\sum_{k=1}^{\infty} (-1)^{k+1} a_k$. Assume $a_n > 0$ for each positive integer n, the sequence a_n decreases to 0 and $\lim_{n\to\infty} S_n = S$. Then it follows that $|S - S_n| < a_{n+1}$.

This bound can be used to determine the accuracy of the partial sum S_n as an approximation of the sum of a convergent alternating series.

Activity 8.4.7

For (a)-(j), use appropriate tests to determine the convergence or divergence of the following series. Throughout, if a series is a convergent geometric series, find its sum.

1. $\displaystyle\sum_{k=3}^{\infty} \frac{2}{\sqrt{k-2}}$
2. $\displaystyle\sum_{k=1}^{\infty} \frac{k}{1+2k}$
3. $\displaystyle\sum_{k=0}^{\infty} \frac{2k^2+1}{k^3+k+1}$

4. $\sum_{k=0}^{\infty} \frac{100^k}{k!}$
5. $\sum_{k=1}^{\infty} \frac{2^k}{5^k}$
6. $\sum_{k=1}^{\infty} \frac{k^3-1}{k^5+1}$
7. $\sum_{k=2}^{\infty} \frac{3^{k-1}}{7^k}$
8. $\sum_{k=2}^{\infty} \frac{1}{k^k}$
9. $\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{\sqrt{k+1}}$
10. $\sum_{k=2}^{\infty} \frac{1}{k\ln(k)}$
11. Determine a value of n so that the nth partial sum S_n of the alternating series $\sum_{n=2}^{\infty} \frac{(-1)^n}{\ln(n)}$ approximates the sum to within 0.001.

Summary

- An alternating series is a series whose terms alternate in sign. It has the form
$$\sum (-1)^k a_k$$
where a_k is a positive real number for each k.
- The sequence of partial sums of a convergent alternating series oscillates around the sum of the series if the sequence of nth terms converges to 0. That is why the Alternating Series Test shows that the alternating series $\sum_{k=1}^{\infty} (-1)^k a_k$ converges whenever the sequence $\{a_n\}$ of nth terms decreases to 0.
- The difference between the $n-1$st partial sum S_{n-1} and the nth partial sum S_n of a convergent alternating series $\sum_{k=1}^{\infty} (-1)^k a_k$ is $|S_n - S_{n-1}| = a_n$. Since the partial sums oscillate around the sum S of the series, it follows that
$$|S - S_n| < a_n.$$
So the nth partial sum of a convergent alternating series $\sum_{k=1}^{\infty} (-1)^k a_k$ approximates the actual sum of the series to within a_n.

Answers

Example $8.4.1$

If we let S be the sum of the series $\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}$, then we know that
$$|S_{100} - S| < a_{101}.$$
Now
$$a_{101} = \frac{1}{101} \approx 0.0099,$$
so the 100th partial sum is within 0.0099 of the sum of the series. We have discussed the fact (and will later verify) that
$$S = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} = \ln(2),$$
and so $S \approx 0.693147$ while
$$S_{100} = \sum_{k=1}^{100} \frac{(-1)^{k+1}}{k} \approx 0.6881721793.$$
We see that the actual difference between S and S_{100} is approximately 0.0049750013, which is indeed less than 0.0099.

This page titled 8.4: Alternating Series is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.4: Alternating Series by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

8.5: Taylor Polynomials and Taylor Series

Motivating Questions

What is a Taylor polynomial? For what purposes are Taylor polynomials used?
What is a Taylor series?
How do we determine the accuracy when we use a Taylor polynomial to approximate a function?

So far, each infinite series we have discussed has been a series of real numbers, such as

$$1+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^k}+\cdots=\sum_{k=0}^{\infty}\frac{1}{2^k} \tag{8.5.1}$$

In the remainder of this chapter, we will include series that involve a variable. For instance, if in the geometric series in Equation (8.5.1) we replace the ratio $r=\frac{1}{2}$ with the variable x, we have the infinite (still geometric) series

$$1+x+x^2+\cdots+x^k+\cdots=\sum_{k=0}^{\infty}x^k \tag{8.5.2}$$

Here we see something very interesting: because a geometric series converges whenever its ratio r satisfies $|r|<1$, and the sum of a convergent geometric series is $\frac{a}{1-r}$, we can say that for $|x|<1$

$$1+x+x^2+\cdots+x^k+\cdots=\frac{1}{1-x}. \tag{8.5.3}$$

Equation (8.5.3) states that the non-polynomial function $\frac{1}{1-x}$ on the right is equal to the infinite polynomial expresssion on the left. Because the terms on the left get very small as k gets large, we can truncate the series and say, for example, that

$$1+x+x^2+x^3\approx\frac{1}{1-x}$$

Preview Activity 8.5.1

Preview Activity 8.3.1 showed how we can approximate the number e using linear, quadratic, and other polynomial functions; we then used similar ideas in Preview Activity 8.4.1 to approximate $\ln(2)$. In this activity, we review and extend the process to find the "best" quadratic approximation to the exponential function e^x around the origin. Let $f(x)=e^x$ throughout this activity.

a. Find a formula for $P_1(x)$, the linearization of $f(x)$ at $x=0$. (We label this linearization P_1 because it is a first degree polynomial approximation.) Recall that $P_1(x)$ is a good approximation to $f(x)$ for values of x close to 0 . Plot f and P_1 near $x=0$ to illustrate this fact.
b. Since $f(x)=e^x$ is not linear, the linear approximation eventually is not a very good one. To obtain better approximations, we want to develop a different approximation that "bends" to make it more closely fit the graph of f near $x=0$. To do so, we add a quadratic term to $P_1(x)$. In other words, we let

$$P_2(x)=P_1(x)+c_2x^2$$

for some real number c_2. We need to determine the value of c_2 that makes the graph of $P_2(x)$ best fit the graph of $f(x)$ near $x=0$.

Remember that $P_1(x)$ was a good linear approximation to $f(x)$ near 0 ; this is because $P_1(0)=f(0)$ and $P_1'(0)=f'(0)$. It is therefore reasonable to seek a value of c_2 so that

$$P_2(0) = f(0), \quad P_2'(0) = f'(0), \quad \text{and } P_2''(0) = f''(0).$$

Remember, we are letting $P_2(x) = P_1(x) + c_2x^2$.

i. Calculate $P_2(0)$ to show that $P_2(0) = f(0)$.
ii. Calculate $P_2'(0)$ to show that $P_2'(0) = f'(0)$.
iii. Calculate $P_2''(x)$. Then find a value for c_2 so that $P_2''(0) = f''(0)$.
iv. Explain why the condition $P_2''(0) = f''(0)$ will put an appropriate "bend" in the graph of P_2 to make P_2 fit the graph of f around $x = 0$.

Taylor Polynomials

Preview Activity $8.5.1$ illustrates the first steps in the process of approximating functions with polynomials. Using this process we can approximate trigonometric, exponential, logarithmic, and other nonpolynomial functions as closely as we like (for certain values of x) with polynomials. This is extraordinarily useful in that it allows us to calculate values of these functions to whatever precision we like using only the operations of addition, subtraction, multiplication, and division, which can be easily programmed in a computer.

We next extend the approach in Preview Activity 8.5.1 to arbitrary functions at arbitrary points. Let f be a function that has as many derivatives as we need at a point $x = a$. Recall that $P_1(x)$ is the tangent line to f at $(a, f(a))$ and is given by the formula

$$P_1(x) = f(a) + f'(a)(x - a).$$

$P_1(x)$ is the linear approximation to f near a that has the same slope and function value as f at the point $x = a$.

We next want to find a quadratic approximation

$$P_2(x) = P_1(x) + c_2(x - a)^2$$

so that $P_2(x)$ more closely models $f(x)$ near $x = a$. Consider the following calculations of the values and derivatives of $P_2(x)$:

$P_2(x) = P_1(x) + c_2(x - a)^2 \quad P_2(a) = P_1(a) = f(a)$
$P_2'(x) = P_1'(x) + 2c_2(x - a) \quad P_2'(a) = P_1'(a) = f'(a)$
$P_2''(x) = 2c_2 \quad P_2''(a) = 2c_2.$

To make $P_2(x)$ fit $f(x)$ better than $P_1(x)$, we want $P_2(x)$ and $f(x)$ to have the same concavity at $x = a$, in addition to having the same slope and function value. That is, we want to have

$$P_2''(a) = f''(a).$$

This implies that

$$2c_2 = f''(a)$$

and thus

$$c_2 = \frac{f''(a)}{2}.$$

Therefore, the quadratic approximation $P_2(x)$ to f centered at $x = a$ is

$$P_2(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2.$$

This approach extends naturally to polynomials of higher degree. We define polynomials

$$P_3(x) = P_2(x) + c_3(x - a)^3,$$
$$P_4(x) = P_3(x) + c_4(x - a)^4,$$
$$P_5(x) = P_4(x) + c_5(x - a)^5,$$

and in general

$$P_n(x) = P_{n-1}(x) + c_n(x-a)^n.$$

The defining property of these polynomials is that for each n, $P_n(x)$ and all its first n derivatives must agree with those of f at $x = a$. In other words we require that

$$P_n^{(k)}(a) = f^{(k)}(a)$$

for all k from 0 to n.
To see the conditions under which this happens, suppose

$$P_n(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + \cdots + c_n(x-a)^n.$$

Then

$$\begin{aligned} P_n^{(0)}(a) &= c_0 \\ P_n^{(1)}(a) &= c_1 \\ P_n^{(2)}(a) &= 2c_2 \\ P_n^{(3)}(a) &= (2)(3)c_3 \\ P_n^{(4)}(a) &= (2)(3)(4)c_4 \\ P_n^{(5)}(a) &= (2)(3)(4)(5)c_5 \end{aligned}$$

and, in general,

$$P_n^{(k)}(a) = (2)(3)(4)\cdots(k-1)(k)c_k = k!c_k.$$

So having $P_n^{(k)}(a) = f^{(k)}(a)$ means that $k!c_k = f^{(k)}(a)$ and therefore

$$c_k = \frac{f^{(k)}(a)}{k!}$$

for each value of k. Using this expression for c_k, we have found the formula for the polynomial approximation of f that we seek. Such a polynomial is called a Taylor polynomial.

Definition: 8.5.1: Taylor Polynomials

The nth order Taylor polynomial of f centered at $x = a$ is given by

$$\begin{aligned} P_n(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k. \end{aligned}$$

This degree n polynomial approximates $f(x)$ near $x = a$ and has the property that $P_n^{(k)}(a) = f^{(k)}(a)$ for $k = 0, 1, \ldots, n$.

Example 8.5.1

Example 8.5.1. Determine the third order Taylor polynomial for $f(x) = e^x$, as well as the general nth order Taylor polynomial for f centered at $x = 0$.

Solution

We know that $f'(x) = e^x$ and so $f''(x) = e^x$ and $f'''(x) = e^x$. Thus,

$$f(0) = f'(0) = f''(0) = f'''(0) = 1.$$

So the third order Taylor polynomial of $f(x) = e^x$ centered at $x = 0$ is

$$\begin{aligned} P_3(x) &= f(0) + f'(0)(x-0) + \frac{f''(0)}{2!}(x-0)^2 + \frac{f'''(0)}{3!}(x-0)^3 \\ &= 1 + x + \frac{x^2}{2} + \frac{x^3}{6}. \end{aligned}$$

In general, for the exponential function f we have $f^{(k)}(x) = e^x$ for every positive integer k. Thus, the k th term in the nth order Taylor polynomial for $f(x)$ centered at $x = 0$ is

$$\frac{f^{(k)}(0)}{k!}(x-0)^k = \frac{1}{k!}x^k.$$

Therefore, the nth order Taylor polynomial for $f(x) = e^x$ centered at $x = 0$ is

$$P_n(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{1}{n!}x^n = \sum_{k=0}^{n} \frac{x^k}{k!}.$$

Activity 8.5.2

We have just seen that the nth order Taylor polynomial centered at $a = 0$ for the exponential function e^x is

$$\sum_{k=0}^{n} \frac{x^k}{k!}.$$

In this activity, we determine small order Taylor polynomials for several other familiar functions, and look for general patterns.

a. Let $f(x) = \frac{1}{1-x}$.

i. Calculate the first four derivatives of $f(x)$ at $x = 0$. Then find the fourth order Taylor polynomial $P_4(x)$ for $\frac{1}{1-x}$ centered at 0 .
ii. Based on your results from part (i), determine a general formula for $f^{(k)}(0)$

b. Let $f(x) = \cos(x)$.

i. Calculate the first four derivatives of $f(x)$ at $x = 0$. Then find the fourth order Taylor polynomial $P_4(x)$ for $\cos(x)$ centered at 0 .
ii. Based on your results from part (i), find a general formula for $f^{(k)}(0)$. (Think about how k being even or odd affects the value of the k th derivative.)

c. Let $f(x) = \sin(x)$.

i. Calculate the first four derivatives of $f(x)$ at $x = 0$. Then find the fourth order Taylor polynomial $P_4(x)$ for $\sin(x)$ centered at 0 .
ii. Based on your results from part (i), find a general formula for $f^{(k)}(0)$. (Think about how k being even or odd affects the value of the k th derivative.)

It is possible that an nth order Taylor polynomial is not a polynomial of degree n; that is, the order of the approximation can be different from the degree of the polynomial. For example, in Activity 8.5.3 we found that the second order Taylor polynomial $P_2(x)$ centered at 0 for $\sin(x)$ is $P_2(x) = x$. In this case, the second order Taylor polynomial is a degree 1 polynomial.

Taylor Series

In Activity 8.5.2 we saw that the fourth order Taylor polynomial $P_4(x)$ for $\sin(x)$ centered at 0 is

$$P_4(x) = x - \frac{x^3}{3!}.$$

The pattern we found for the derivatives $f^{(k)}(0)$ describe the higher-order Taylor polynomials, e.g.,

$$P_5(x) = x - \tfrac{x^3}{3!} + \tfrac{x^5}{5!},$$
$$P_7(x) = x - \tfrac{x^3}{3!} + \tfrac{x^5}{5!} - \tfrac{x^7}{7!},$$
$$P_9(x) = x - \tfrac{x^3}{3!} + \tfrac{x^5}{5!} - \tfrac{x^7}{7!} + \tfrac{x^9}{9!},$$

and so on. It is instructive to consider the graphical behavior of these functions; Figure 8.5.1 shows the graphs of a few of the Taylor polynomials centered at 0 for the sine function.

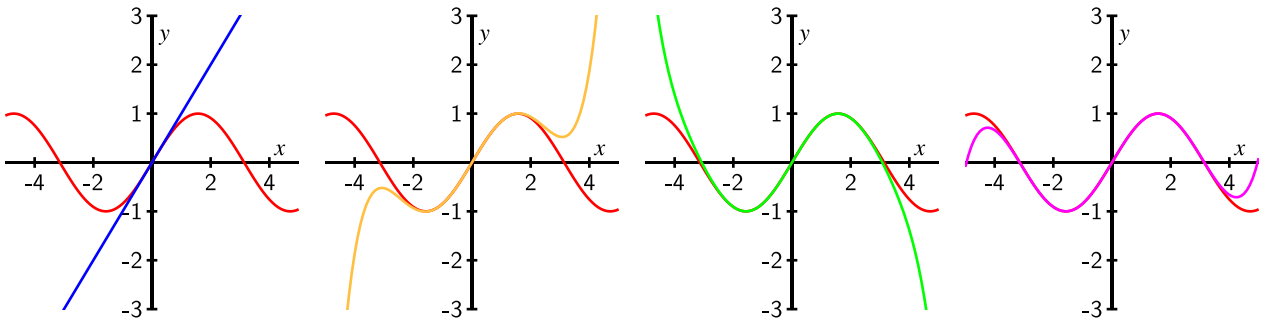


Figure 8.5.1: The order $1, 5, 7$, and 9 Taylor polynomials centered at $x = 0$ for $f(x) = \sin(x)$.

Notice that $P_1(x)$ is close to the sine function only for values of x that are close to 0 , but as we increase the degree of the Taylor polynomial the Taylor polynomials provide a better fit to the graph of the sine function over larger intervals. This illustrates the general behavior of Taylor polynomials: for any sufficiently well-behaved function, the sequence $\{P_n(x)\}$ of Taylor polynomials converges to the function f on larger and larger intervals (though those intervals may not necessarily increase without bound). If the Taylor polynomials ultimately converge to f on its entire domain, we write

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

Definition: 8.5.2

Let f be a function all of whose derivatives exist at $x = a$. The Taylor series for f centered at $x = a$ is the series $T_f(x)$ defined by

$$T_f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k. \tag{8.5.4}$$

In the special case where $a = 0$ in Definition 8.5.3, the Taylor series is also called the Maclaurin series for f. From Example 8.5.1 we know the nth order Taylor polynomial centered at 0 for the exponential function e^x; thus, the Maclaurin series for e^x is

$$\sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Activity 8.5.3

In Activity \({8{.}5{.}\eatSpaces{2} \endSpaces}\ we determined small order Taylor polynomials for a few familiar functions, and also found general patterns in the derivatives evaluated at 0 . Use that information to write the Taylor series centered at 0 for the following functions.

a. $f(x) = \frac{1}{1-x}$
b. $f(x) = \cos(x)$ (You will need to carefully consider how to indicate that many of the coefficients are 0 . Think about a general way to represent an even integer.)
c. $f(x) = \sin(x)$ (You will need to carefully consider how to indicate that many of the coefficients are 0 . Think about a general way to represent an odd integer.)
d. $f(x) = \frac{1}{1+x}$

Activity 8.5.4

a. Plot the graphs of several of the Taylor polynomials centered at 0 (of order at least 5) for e^x and convince yourself that these Taylor polynomials converge to e^x for every value of x.
b. Draw the graphs of several of the Taylor polynomials centered at 0 (of order at least 6) for $\cos(x)$ and convince yourself that these Taylor polynomials converge to $\cos(x)$ for every value of x. Write the Taylor series centered at 0 for $\cos(x)$.
c. Draw the graphs of several of the Taylor polynomials centered at 0 for $\frac{1}{1-x}$. Based on your graphs, for what values of x do these Taylor polynomials appear to converge to $\frac{1}{1-x}$? How is this situation different from what we observe with e^x and $\cos(x)$? In addition, write the Taylor series centered at 0 for $\frac{1}{1-x}$.

The Maclaurin series for $e^x, \sin(x), \cos(x)$, and $\frac{1}{1-x}$ will be used frequently, so we should be certain to know and recognize them well.

The Interval of Convergence of a Taylor Series

In the previous section (in Figure 8.5.1 and Activity 8.5.4) we observed that the Taylor polynomials centered at 0 for $e^x, \cos(x)$, and $\sin(x)$ converged to these functions for all values of x in their domain, but that the Taylor polynomials centered at 0 for $\frac{1}{1-x}$ converge to $\frac{1}{1-x}$ on the interval $(-1, 1)$ and diverge for all other values of x. So the Taylor series for a function $f(x)$ does not need to converge for all values of x in the domain of f.
Our observations suggest two natural questions: can we determine the values of x for which a given Taylor series converges? And does the Taylor series for a function f actually converge to $f(x)$?

Example 8.5.1

Graphical evidence suggests that the Taylor series centered at 0 for e^x converges for all values of x. To verify this, use the Ratio Test to determine all values of x for which the Taylor series

$$\sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{8.5.5}$$

converges absolutely.

Solution

Recall that the Ratio Test applies only to series of nonnegative terms. In this example, the variable x may have negative values. But we are interested in absolute convergence, so we apply the Ratio Test to the series

$$\sum_{k=0}^{\infty} \left| \frac{x^k}{k!} \right| = \sum_{k=0}^{\infty} \frac{|x|^k}{k!}.$$

Now, observe that

$$\begin{aligned}\lim_{k\to\infty}\frac{a_{k+1}}{a_k} &= \lim_{k\to\infty}\frac{\frac{|x|^{k+1}}{(k+1)!}}{\frac{|x|^k}{k}}\\ &= \lim_{k\to\infty}\frac{|x|^{k+1}k!}{|x|^k(k+1)!}\\ &= \lim_{k\to\infty}\frac{|x|}{k+1}\\ &= 0\end{aligned}$$

for any value of x. So the Taylor series $(8.5.4)$ converges absolutely for every value of x, and thus converges for every value of x.

One question still remains: while the Taylor series for e^x converges for all x, what we have done does not tell us that this Taylor series actually converges to e^x for each x. We'll return to this question when we consider the error in a Taylor approximation near the end of this section.

We can apply the main idea from Example 8.5.4 in general. To determine the values of x for which a Taylor series

$$\sum_{k=0}^{\infty} c_k(x-a)^k$$

centered at $x=a$ will converge, we apply the Ratio Test with $a_k = \left|c_k(x-a)^k\right|$. The series converges if $\lim_{k\to\infty}\frac{a_{k+1}}{a_k} < 1$.

Observe that

$$\frac{a_{k+1}}{a_k} = |x-a|\frac{|c_{k+1}|}{|c_k|}$$

so when we apply the Ratio Test, we get

$$\lim_{k\to\infty}\frac{a_{k+1}}{a_k} = \lim_{k\to\infty}|x-a|\frac{|c_{k+1}|}{|c_k|}.$$

Note suppose that

$$\lim_{k\to\infty}\frac{|c_{k+1}|}{|c_k|} = L$$

so that

$$\lim_{k\to\infty}\frac{a_{k+1}}{a_k} = |x-a|\cdot L.$$

There are three possibilities for L : L can be 0 , it can be a finite positive value, or it can be infinite. Based on this value of L, we can determine for which values of x the original Taylor series converges.

- If $L=0$, then the Taylor series converges on $(-\infty,\infty)$.
- If L is infinite, then the Taylor series converges only at $x=a$.
- If L is finite and nonzero, then the Taylor series converges absolutely for all x that satisfy

  $$|x-a|\cdot L < 1$$

  or for all x such that

  $$|x-a| < \frac{1}{L},$$

  which is the interval

  $$\left(a-\frac{1}{L}, a+\frac{1}{L}\right).$$

Because the Ratio Test is inconclusive when the $|x-a| \cdot L = 1$, the endpoints $a \pm \frac{1}{L}$ have to be checked separately.

It is important to notice that the set of x values at which a Taylor series converges is always an interval centered at $x = a$. For this reason, the set on which a Taylor series converges is called the interval of convergence. Half the length of the interval of convergence is called the radius of convergence. If the interval of convergence of a Taylor series is infinite, then we say that the radius of convergence is infinite.

Activity 8.5.4

a. Use the Ratio Test to explicitly determine the interval of convergence of the Taylor series for $f(x) = \frac{1}{1-x}$ centered at $x = 0$.
b. Use the Ratio Test to explicitly determine the interval of convergence of the Taylor series for $f(x) = \cos(x)$ centered at $x = 0$.
c. Use the Ratio Test to explicitly determine the interval of convergence of the Taylor series for $f(x) = \sin(x)$ centered at $x = 0$.

The Ratio Test allows us to determine the set of x values for which a Taylor series converges absolutely. However, just because a Taylor series for a function f converges, we cannot be certain that the Taylor series actually converges to $f(x)$. To show why and where a Taylor series does in fact converge to the function f, we next consider the error that is present in Taylor polynomials.

Error Approximations for Taylor Polynomials

We now know how to find Taylor polynomials for functions such as $\sin(x)$, and how to determine the interval of convergence of the corresponding Taylor series. We next develop an error bound that will tell us how well an nth order Taylor polynomial $P_n(x)$ approximates its generating function $f(x)$. This error bound will also allow us to determine whether a Taylor series on its interval of convergence actually equals the function f from which the Taylor series is derived. Finally, we will be able to use the error bound to determine the order of the Taylor polynomial $P_n(x)$ that we will ensure that $P_n(x)$ approximates $f(x)$ to the desired degree of accuracy.

For this argument, we assume throughout that we center our approximations at 0 (but a similar argument holds for approximations centered at a). We define the exact error, $E_n(x)$, that results from approximating $f(x)$ with $P_n(x)$ by

$$E_n(x) = f(x) - P_n(x).$$

We are particularly interested in $|E_n(x)|$, the distance between P_n and f.
Because

$$P_n^{(k)}(0) = f^{(k)}(0)$$

for $0 \leq k \leq n$, we know that

$$E_n^{(k)}(0) = 0$$

for $0 \leq k \leq n$. Furthermore, since $P_n(x)$ is a polynomial of degree less than or equal to n, we know that

$$P_n^{(n+1)}(x) = 0.$$

Thus, since $E_n^{(n+1)}(x) = f^{(n+1)}(x) - P_n^{(n+1)}(x)$, it follows that

$$E_n^{(n+1)}(x) = f^{(n+1)}(x)$$

for all x.

Suppose that we want to approximate $f(x)$ at a number c close to 0 using $P_n(c)$. If we assume $\left|f^{(n+1)}(t)\right|$ is bounded by some number M on $[0,c]$, so that

$$\left|f^{(n+1)}(t)\right| \leq M$$

for all $0 \leq t \leq c$, then we can say that

$$\left|E_n^{(n+1)}(t)\right| = \left|f^{(n+1)}(t)\right| \leq M$$

for all t between 0 and c. Equivalently,

$$-M \leq E_n^{(n+1)}(t) \leq M \tag{8.5.6}$$

on $[0,c]$. Next, we integrate the three terms in Inequality $8.5.5$ from $t=0$ to $t=x$, and thus find that

$$\int_0^x -M dt \leq \int_0^x E_n^{(n+1)}(t)dt \leq \int_0^x M dt$$

for every value of x in $[0,c]$. Since $E_n^{(n)}(0)=0$, the First FTC tells us that

$$-Mx \leq E_n^{(n)}(x) \leq Mx$$

for every x in $[0,c]$.
Integrating this last inequality, we obtain

$$\int_0^x -Mtdt \leq \int_0^x E_n^{(n)}(t)dt \leq \int_0^x Mt\_dt$$

and thus

$$-M\frac{x^2}{2} \leq E_n^{(n-1)}(x) \leq M\frac{x^2}{2}$$

for all x in $[0,c]$.
Integrating n times, we arrive at

$$-M\frac{x^{n+1}}{(n+1)!} \leq E_n(x) \leq M\frac{x^{n+1}}{(n+1)!}$$

for all x in $[0,c]$. This enables us to conclude that

$$|E_n(x)| \leq M\frac{|x|^{n+1}}{(n+1)!}$$

for all x in $[0,c]$, and we have found a bound on the approximation's error, E_n.
Our work above was based on the approximation centered at $a=0$; the argument may be generalized to hold for any value of a, which results in the following theorem.

Definition: 8.5.3: The Lagrange Error Bound

The Lagrange Error Bound for $P_n(x)$.

Let f be a continuous function with $n+1$ continuous derivatives. Suppose that M is a positive real number such that $\left|f^{(n+1)}(x)\right| \leq M$ on the interval $[a,c]$. If $P_n(x)$ is the nth order Taylor polynomial for $f(x)$ centered at $x=a$, then

$$|P_n(c) - f(c)| \leq M\frac{|c-a|^{n+1}}{(n+1)!}.$$

We can use this error bound to tell us important information about Taylor polynomials and Taylor series, as we see in the following examples and activities.

Example 8.5.5

Determine how well the 10th order Taylor polynomial $P_{10}(x)$ for $\sin(x)$, centered at 0 , approximates $\sin(2)$.

Solution

To answer this question we use $f(x) = \sin(x), c = 2, a = 0$, and $n = 10$ in the Lagrange error bound formula. We also need to find an appropriate value for M. Note that the derivatives of $f(x) = \sin(x)$ are all equal to $\pm\sin(x)$ or $\pm\cos(x)$. Thus,

$$\left|f^{(n+1)}(x)\right| \leq 1$$

for any n and x. Therefore, we can choose M to be 1 . Then

$$|P_{10}(2) - f(2)| \leq (1)\frac{|2-0|^{11}}{(11)!} = \frac{2^{11}}{(11)!} \approx 0.00005130671797.$$

So $P_{10}(2)$ approximates $\sin(2)$ to within at most 0.00005130671797. A computer algebra system tells us that

$$P_{10}(2) \approx 0.9093474427 \text{ and } \sin(2) \approx 0.9092974268$$

with an actual difference of about 0.0000500159.

Activity 8.5.6

Let $P_n(x)$ be the nth order Taylor polynomial for $\sin(x)$ centered at $x = 0$. Determine how large we need to choose n so that $P_n(2)$ approximates $\sin(2)$ to 20 decimal places.

Example 8.5.6

Show that the Taylor series for $\sin(x)$ actually converges to $\sin(x)$ for all x.

Solution

Recall from the previous example that since $f(x) = \sin(x)$, we know

$$\left|f^{(n+1)}(x)\right| \leq 1$$

for any n and x. This allows us to choose $M = 1$ in the Lagrange error bound formula. Thus,

$$|P_n(x) - \sin(x)| \leq \frac{|x|^{n+1}}{(n+1)!} \tag{8.5.7}$$

for every x.
We showed in earlier work that the Taylor series $\sum_{k=0}^{\infty} \frac{x^k}{k!}$ converges for every value of x. Because the terms of any convergent series must approach zero, it follows that

$$\lim_{n\to\infty} \frac{x^{n+1}}{(n+1)!} = 0$$

for every value of x. Thus, taking the limit as $n \to \infty$ in the inequality 8.5.6, it follows that

$$\lim_{n\to\infty} |P_n(x) - \sin(x)| = 0.$$

As a result, we can now write

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$$

for every real number x.

Activity 8.5.7

a. Show that the Taylor series centered at 0 for $\cos(x)$ converges to $\cos(x)$ for every real number x.
b. Next we consider the Taylor series for e^x.
   i. Show that the Taylor series centered at 0 for e^x converges to e^x for every nonnegative value of x.
   ii. Show that the Taylor series centered at 0 for e^x converges to e^x for every negative value of x.
   iii. Explain why the Taylor series centered at 0 for e^x converges to e^x for every real number x. Recall that we earlier showed that the Taylor series centered at 0 for e^x converges for all x, and we have now completed the argument that the Taylor series for e^x actually converges to e^x for all x.
c. Let $P_n(x)$ be the nth order Taylor polynomial for e^x centered at 0 . Find a value of n so that $P_n(5)$ approximates e^5 correct to 8 decimal places.

Summary

- We can use Taylor polynomials to approximate functions. This allows us to approximate values of functions using only addition, subtraction, multiplication, and division of real numbers. The nth order Taylor polynomial centered at $x = a$ of a function f is

$$\begin{aligned} P_n(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k. \end{aligned}$$

- The Taylor series centered at $x = a$ for a function f is

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k. \tag{8.5.8}$$

  The nth order Taylor polynomial centered at a for f is the nth partial sum of its Taylor series centered at a. So the nth order Taylor polynomial for a function f is an approximation to f on the interval where the Taylor series converges; for the values of x for which the Taylor series converges to f we write

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k. \tag{8.5.9}$$

- The Lagrange Error Bound shows us how to determine the accuracy in using a Taylor polynomial to approximate a function. More specifically, if $P_n(x)$ is the nth order Taylor polynomial for f centered at $x = a$ and if M is an upper bound for $\left|f^{(n+1)}(x)\right|$ on the interval $[a, c]$, then

$$|P_n(c) - f(c)| \leq M\frac{|c-a|^{n+1}}{(n+1)!}. \tag{8.5.10}$$

This page titled 8.5: Taylor Polynomials and Taylor Series is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.8: The Tangent Line Approximation by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

8.6: Power Series

Motivating Questions

- What is a power series?
- What are some important uses of power series?
- What is the connection between power series and Taylor series?

We have noted in our work with Taylor polynomials and Taylor series that polynomial functions are the simplest possible functions in mathematics, in part because they require only addition and multiplication to evaluate. From the point of view of calculus, polynomials are especially nice: we can easily differentiate or integrate any polynomial. In light of our work in Section 8.5, we now know that several important non-polynomials have polynomial-like expansions. For example, for any real number x,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots.$$

There are two settings where other series like the one for e^x arise: we may be given an expression such as

$$1 + 2x + 3x^2 + 4x^3 + \cdots$$

and we seek the values of x for which the expression makes sense. Or we may be trying to find an unknown function f that has expression

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k + \cdots,$$

and we try to determine the values of the constants $a_0, a_1, \ldots$. The latter situation is explored in Preview Activity 8.6.1.

Preview Activity 8.6.1

In Chapter 7, we learned some of the many important applications of differential equations, and learned some approaches to solve or analyze them. Here, we consider an important approach that will allow us to solve a wider variety of differential equations.

Let's consider the familiar differential equation from exponential population growth given by

$$y' = ky, \tag{8.6.1}$$

where k is the constant of proportionality. While we can solve this differential equation using methods we have already learned, we take a different approach now that can be applied to a much larger set of differential equations. For the rest of this activity, let's assume that $k = 1$. We will use our knowledge of Taylor series to find a solution to the differential equation (8.6.1).

To do so, we assume that we have a solution $y = f(x)$ and that $f(x)$ has a Taylor series that can be written in the form

$$y = f(x) = \sum_{k=0}^{\infty} a_k x^k,$$

where the coefficients a_k are undetermined. Our task is to find the coefficients.

1. Assume that we can differentiate a power series term by term. By taking the derivative of $f(x)$ with respect to x and substituting the result into the differential equation (8.6.1), show that the equation

$$\sum_{k=1}^{\infty} k a_k x^{k-1} = \sum_{k=0}^{\infty} a_k x^k$$

must be satisfied in order for $f(x) = \sum_{k=0}^{\infty} a_k x^k$ to be a solution of the DE.

2. Two series are equal if and only if they have the same coefficients on like power terms. Use this fact to find a relationship between a_1 and a_0.
3. Now write a_2 in terms of a_1. Then write a_2 in terms of a_0.
4. Write a_3 in terms of a_2. Then write a_3 in terms of a_0.
5. Write a_4 in terms of a_3. Then write a_4 in terms of a_0.
6. Observe that there is a pattern in (b)-(e). Find a general formula for a_k in terms of a_0.
7. Write the series expansion for y using only the unknown coefficient a_0. From this, determine what familiar functions satisfy the differential equation (8.6.1). (*Hint*: Compare to a familiar Taylor series.)

Power Series

As Preview Activity 8.6.1 shows, it can be useful to treat an unknown function as if it has a Taylor series, and then determine the coefficients from other information. In other words, we define a function as an infinite series of powers of x and then determine the coefficients based on something besides a formula for the function. This method allows us to approximate solutions to many different types of differential equations, even if we cannot solve them explicitly. This is different from our work with Taylor series since we are not using an original function f to generate the coefficients of the series.

Definition 8.6.1: Power Series

A ***power series*** centered at $x = a$ is a function of the form

$$\sum_{k=0}^{\infty} c_k (x-a)^k \tag{8.6.2}$$

where $\{c_k\}$ is a sequence of real numbers and x is an independent variable.

A power series defines a function f whose domain is the set of x values for which the power series converges. We therefore write

$$f(x) = \sum_{k=0}^{\infty} c_k (x-a)^k.$$

It turns out that[1], on its interval of convergence, every power series is in fact the Taylor series of the function it defines, so all of the techniques we developed in the previous section can be applied to power series as well.

See Exercise 8.6.4 in this section.

Example 8.6.1

Consider the power series defined by

$$f(x) = \sum_{k=0}^{\infty} \frac{x^k}{2^k}.$$

What are $f(1)$ and $f\left(\frac{3}{2}\right)$? Find a general formula for $f(x)$ and determine the values for which this power series converges.

Answer available at the end of this section.

As we did for Taylor series, we define the interval of convergence of a power series (8.6.2) to be the set of values of x for which the series converges. And as we did with Taylor series, we typically use the Ratio Test to find the values of x for which the power series converges absolutely, and then check the endpoints separately if the radius of convergence is finite.

Example 8.6.2

Let $f(x) = \sum_{k=1}^{\infty} \frac{x^k}{k^2}$. Determine the interval of convergence of this power series.

Answer available at the end of this section.

Activity 8.6.2

Determine the interval of convergence of each power series.

1. $\displaystyle\sum_{k=1}^{\infty} \frac{(x-1)^k}{3k}$
2. $\displaystyle\sum_{k=1}^{\infty} kx^k$
3. $\displaystyle\sum_{k=1}^{\infty} \frac{k^2(x+1)^k}{4^k}$
4. $\displaystyle\sum_{k=1}^{\infty} \frac{x^k}{(2k)!}$
5. $\displaystyle\sum_{k=1}^{\infty} k!x^k$

Manipulating Power Series

We know power series expansions for important functions such as $\sin(x)$ and e^x. Often, we can use a known power series expansion to find a power series for a different, but related, function. The next activity demonstrates one way to do this.

Activity 8.6.3

Our goal in this activity is to find a power series expansion for $f(x) = \frac{1}{1+x^2}$ centered at $x = 0$.

While we could use the methods of Section 8.5 and differentiate $f(x) = \frac{1}{1+x^2}$ several times to look for patterns and find the Taylor series for $f(x)$, we seek an alternate approach because of how complicated the derivatives of $f(x)$ quickly become.

1. What is the Taylor series expansion for $g(x) = \frac{1}{1-x}$? What is the interval of convergence of this series?
2. How is $g(-x^2)$ related to $f(x)$? Explain, and hence substitute $-x^2$ for x in the power series expansion for $g(x)$. Given the relationship between $g(-x^2)$ and $f(x)$, how is the resulting series related to $f(x)$?
3. For which values of x will this power series expansion for $f(x)$ be valid? Why?

In a previous section we found several important Maclaurin series and their intervals of convergence. Here, we list these key functions and their corresponding expansions.

Note

$$\sin(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!} \qquad \text{for } -\infty < x < \infty$$

$$\cos(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!} \qquad \text{for } -\infty < x < \infty$$

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \qquad \text{for } -\infty < x < \infty$$

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \qquad \text{for } -1 < x < 1$$

As we saw in Activity 8.6.3, we can use these known series to find other power series expansions for related functions such as $\sin(x^2)$, e^{5x^3}, and $\cos(x^5)$.

Activity 8.6.4

Let f be the function given by the power series expansion

$$f(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}.$$

1. Assume that we can differentiate a power series term by term, just like we can differentiate a (finite) polynomial. Use the fact that

   $$f(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + (-1)^k \frac{x^{2k}}{(2k)!} + \cdots$$

   to find a power series expansion for $f'(x)$.

2. Observe that $f(x)$ and $f'(x)$ have familiar Taylor series. What familiar functions are these? What known relationship does our work demonstrate?
3. What is the series expansion for $f''(x)$? What familiar function is $f''(x)$?

Our work in Activity 8.6.3 holds more generally. The corresponding theorem, which we will not prove, states that we can differentiate a power series for a function f term by term and obtain the series expansion for f', and similarly we can integrate a series expansion for a function f term by term and obtain the series expansion for $\int f(x)\,dx$. For both, the radius of convergence of the resulting series is the same as the original, though it is possible that the convergence status of the various series may differ at the endpoints. The formal statement of the Power Series Differentiation and Integration Theorem follows.

Power Series Differentiation and Integration Theorem

Suppose $f(x)$ has a power series expansion

$$f(x) = \sum_{k=0}^{\infty} c_k x^k$$

and that the series converges absolutely to $f(x)$ on the interval $-r < x < r$. Then, the power series $\sum_{k=1}^{\infty} k c_k x^{k-1}$ obtained by differentiating the power series for $f(x)$ term by term converges absolutely to $f'(x)$ on the interval $-r < x < r$. That is,

$$f'(x) = \sum_{k=1}^{\infty} k c_k x^{k-1}, \text{ for } |x| < r.$$

Similarly, the power series $\sum_{k=0}^{\infty} c_k \frac{x^{k+1}}{k+1}$ obtained by integrating the power series for $f(x)$ term by term converges absolutely on the interval $-r < x < r$, and

$$\int f(x)\ dx = \sum_{k=0}^{\infty} c_k \frac{x^{k+1}}{k+1} + C, \text{ for } |x| < r.$$

This theorem validates the steps we took in Activity 8.6.4. It tells us that we can differentiate and integrate term by term on the interior of the interval of convergence, but it does not tell us what happens at the endpoints of this interval. We always need to check what happens at the endpoints separately. More importantly, we can use use the approach of differentiating or integrating a series term by term to find new series.

Example 8.6.3

Find a series expansion centered at $x = 0$ for $\arctan(x)$, as well as its interval of convergence.

Answer available at the end of this section.

Activity 8.6.5

Find a power series expansion for $\ln(1 + x)$ centered at $x = 0$ and determine its interval of convergence.

Summary

- A power series is a series of the form
$$\sum_{k=0}^{\infty} a_k x^k.$$
- We can often assume a solution to a given problem can be written as a power series, then use the information in the problem to determine the coefficients in the power series. This method allows us to approximate solutions to certain problems using partial sums of the power series; that is, we can find approximate solutions that are polynomials.
- The connection between power series and Taylor series is that they are essentially the same thing: on its interval of convergence a power series is the Taylor series of its sum.

Answers

Example 8.6.1

If we evaluate f at $x = 1$ we obtain the series

$$\sum_{k=0}^{\infty} \frac{1}{2^k}$$

which is a geometric series with ratio $\frac{1}{2}$. So we can sum this series and find that

$$f(1) = \frac{1}{1 - \frac{1}{2}} = 2.$$

Similarly,

$$f(3/2) = \sum_{k=0}^{\infty} \left(\frac{3}{4}\right)^k = \frac{1}{1 - \frac{3}{4}} = 4.$$

In general, $f(x)$ is a geometric series with ratio $\frac{x}{2}$, so

$$f(x) = \sum_{k=0}^{\infty} \left(\frac{x}{2}\right)^k = \frac{1}{1 - \frac{x}{2}} = \frac{2}{2 - x}$$

provided that $-1 < \frac{x}{2} < 1$ (which ensures that the ratio is less than 1 in absolute value). Thus, the power series that defines $f(x) = \frac{2}{2-x}$ converges for $-2 < x < 2$.

Example 8.6.2

First we will plot some of the partial sums of this power series to get an idea of the interval of convergence. Let

$$S_n(x) = \sum_{k=1}^{n} \frac{x^k}{k^2}$$

for each $n \geq 1$. Figure 8.6.1 shows plots of $S_{10}(x)$ (in red), $S_{25}(x)$ (in blue), and $S_{50}(x)$ (in green).

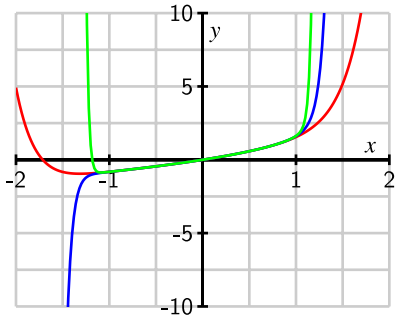


Figure 8.6.1. Graphs of some partial sums of the power series $\sum_{k=1}^{\infty} \frac{x^k}{k^2}$.

The behavior of S_{50} in particular suggests that $f(x)$ appears to be converging to a particular curve on the interval $(-1, 1)$, while growing without bound outside of that interval. Thus, the interval of convergence might be $-1 < x < 1$. To verify our conjecture, we apply the Ratio Test. Now,

$$a_k = \frac{x^k}{k^2},$$

so

$$\begin{aligned}
\lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|} &= \lim_{k\to\infty} \frac{\frac{|x|^{k+1}}{(k+1)^2}}{\frac{|x|^k}{k^2}} \\
&= \lim_{k\to\infty} |x| \left(\frac{k}{k+1}\right)^2 \\
&= |x| \lim_{k\to\infty} \left(\frac{k}{k+1}\right)^2 \\
&= |x|.
\end{aligned}$$

Therefore, the Ratio Test tells us that $f(x)$ converges absolutely when $|x| < 1$ and diverges when $|x| > 1$. Because the Ratio Test is inconclusive when $|x| = 1$, we need to check $x = 1$ and $x = -1$ individually.

When $x = 1$, observe that

$$f(1) = \sum_{k=1}^{\infty} \frac{1}{k^2}.$$

This is a p-series with $p > 1$, which we know converges. When $x = -1$, we have

$$f(-1) = \sum_{k=1}^{\infty} \frac{(-1)^k}{k^2}.$$

This is an alternating series, and since the sequence $\left\{\frac{1}{n^2}\right\}$ decreases to 0, the power series converges by the Alternating Series Test. Thus, the interval of convergence of this power series is $-1 \leq x \leq 1$.

Example 8.6.3

While we could differentiate $\arctan(x)$ repeatedly and look for patterns in the derivative values at $x = 0$ in an attempt to find the Maclaurin series for $\arctan(x)$ from the definition, it turns out to be far easier to use a known series in an insightful way. In Activity 8.6.3, we found that

$$\frac{1}{1+x^2} = \sum_{k=0}^{\infty} (-1)^k x^{2k}$$

for $-1 < x < 1$. Recall that

$$\frac{d}{dx}[\arctan(x)] = \frac{1}{1+x^2},$$

and therefore

$$\int \frac{1}{1+x^2}\, dx = \arctan(x) + C.$$

It follows that we can integrate the series for $\frac{1}{1+x^2}$ term by term to obtain the power series expansion for $\arctan(x)$. Doing so, we find that

$$\begin{aligned}
\arctan(x) &= \int \left(\sum_{k=0}^{\infty} (-1)^k x^{2k}\right) dx \\
&= \sum_{k=0}^{\infty} \left(\int (-1)^k x^{2k}\, dx\right) \\
&= \left(\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{2k+1}\right) + C.
\end{aligned}$$

The Power Series Differentiation and Integration Theorem tells us that this equality is valid for at least $-1 < x < 1$.

To find the value of the constant C, we can use the fact that $\arctan(0) = 0$. So

$$0 = \arctan(0) = \left(\sum_{k=0}^{\infty} (-1)^k \frac{0^{2k+1}}{2k+1}\right) + C = C,$$

and we must have $C = 0$. Therefore,

$$\arctan(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{2k+1} \tag{8.6.3}$$

for at least $-1 < x < 1$.

It is a straightforward exercise to check that the power series

$$\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{2k+1}$$

converges both when $x = -1$ and when $x = 1$; in each case, we have an alternating series with terms $\frac{1}{2k+1}$ that decrease to 0, and thus the interval of convergence for the series expansion for $\arctan(x)$ in Equation (8.6.3) is $-1 \leq x \leq 1$.

This page titled 8.6: Power Series is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.6: Power Series by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

8.7: Sequences and Series (Exercises)

8.1: Sequences

1. Limits of five sequences

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

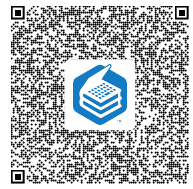
View this Interactive Content online



2. Formula for a sequence, given first terms

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Divergent or convergent sequences

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Terms of a sequence from sampling a signal

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5

Finding limits of convergent sequences can be a challenge. However, there is a useful tool we can adapt from our study of limits of continuous functions at infinity to use to find limits of sequences. We illustrate in this exercise with the example of the sequence

$$\frac{\ln(n)}{n}.$$

1. Calculate the first 10 terms of this sequence. Based on these calculations, do you think the sequence converges or diverges? Why?
2. For this sequence, there is a corresponding continuous function f defined by
$$f(x) = \frac{\ln(x)}{x}.$$
Draw the graph of $f(x)$ on the interval $[0, 10]$ and then plot the entries of the sequence on the graph. What conclusion do you think we can draw about the sequence $\left\{\frac{\ln(n)}{n}\right\}$ if $\lim_{x\to\infty} f(x) = L$? Explain.
3. Note that $f(x)$ has the indeterminate form $\frac{\infty}{\infty}$ as x goes to infinity. What idea from differential calculus can we use to calculate $\lim_{x\to\infty} f(x)$? Use this method to find $\lim_{x\to\infty} f(x)$. What, then, is $\lim_{n\to\infty} \frac{\ln(n)}{n}$?

6

We return to the example begun in Preview Activity 8.1.1 to see how to derive the formula for the amount of money in an account at a given time. We do this in a general setting. Suppose you invest P dollars (called the *principal*) in an account paying $r\%$ interest compounded monthly. In the first month you will receive $\frac{r}{12}$ (here r is in decimal form; e.g., if we have 8% interest, we write $\frac{0.08}{12}$) of the principal P in interest, so you earn
$$P\left(\frac{r}{12}\right)$$
dollars in interest. Assume that you reinvest all interest. Then at the end of the first month your account will contain the original principal P plus the interest, or a total of
$$P_1 = P + P\left(\frac{r}{12}\right) = P\left(1 + \frac{r}{12}\right)$$
dollars.

1. Given that your principal is now P_1 dollars, how much interest will you earn in the second month? If P_2 is the total amount of money in your account at the end of the second month, explain why
$$P_2 = P_1\left(1 + \frac{r}{12}\right) = P\left(1 + \frac{r}{12}\right)^2.$$
2. Find a formula for P_3, the total amount of money in the account at the end of the third month in terms of the original investment P.
3. There is a pattern to these calculations. Let P_n the total amount of money in the account at the end of the third month in terms of the original investment P. Find a formula for P_n.

7

Sequences have many applications in mathematics and the sciences. In a recent paper[3] the authors write

Hui H, Farilla L, Merkel P, Perfetti R. The short half-life of glucagon-like peptide-1 in plasma does not reflect its long-lasting beneficial effects, *Eur J Endocrinol* 2002 Jun;146(6):863-9.

> The incretin hormone glucagon-like peptide-1 (GLP-1) is capable of ameliorating glucose-dependent insulin secretion in subjects with diabetes. However, its very short half-life (1.5-5 min) in plasma represents a major limitation for its use in the clinical setting.

The half-life of GLP-1 is the time it takes for half of the hormone to decay in its medium. For this exercise, assume the half-life of GLP-1 is 5 minutes. So if A is the amount of GLP-1 in plasma at some time t, then only $\frac{A}{2}$ of the hormone will be present after $t + 5$ minutes. Suppose $A_0 = 100$ grams of the hormone are initially present in plasma.

1. Let A_1 be the amount of GLP-1 present after 5 minutes. Find the value of A_1.

2. Let A_2 be the amount of GLP-1 present after 10 minutes. Find the value of A_2.
3. Let A_3 be the amount of GLP-1 present after 15 minutes. Find the value of A_3.
4. Let A_4 be the amount of GLP-1 present after 20 minutes. Find the value of A_4.
5. Let A_n be the amount of GLP-1 present after $5n$ minutes. Find a formula for A_n.
6. Does the sequence $\{A_n\}$ converge or diverge? If the sequence converges, find its limit and explain why this value makes sense in the context of this problem.
7. Determine the number of minutes it takes until the amount of GLP-1 in plasma is 1 gram.

8

Continuous data is the basis for analog information, like music stored on old cassette tapes or vinyl records. A digital signal like on a CD or MP3 file is obtained by sampling an analog signal at some regular time interval and storing that information. For example, the sampling rate of a compact disk is 44,100 samples per second. So a digital recording is only an approximation of the actual analog information. Digital information can be manipulated in many useful ways that allow for, among other things, noisy signals to be cleaned up and large collections of information to be compressed and stored in much smaller space. While we won't investigate these techniques in this chapter, this exercise is intended to give an idea of the importance of discrete (digital) techniques.

Let f be the continuous function defined by $f(x) = \sin(4x)$ on the interval $[0, 10]$. A graph of f is shown in Figure 8.1.5.

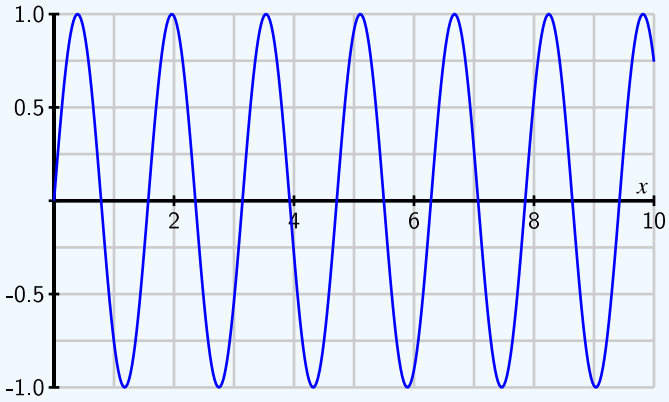


Figure 8.1.5. The graph of $f(x) = \sin(4x)$ on the interval $[0, 10]$

We approximate f by *sampling*, that is by partitioning the interval $[0, 10]$ into uniform subintervals and recording the values of f at the endpoints.

1. Ineffective sampling can lead to several problems in reproducing the original signal. As an example, partition the interval $[0, 10]$ into 8 equal length subintervals and create a list of points (the *sample*) using the endpoints of each subinterval. Plot your sample on graph of f in Figure Figure 8.1.5. What can you say about the period of your sample as compared to the period of the original function?
2. The sampling rate is the number of samples of a signal taken per second. As the part (a) illustrates, sampling at too small a rate can cause serious problems with reproducing the original signal (this problem of inefficient sampling leading to an inaccurate approximation is called *aliasing*). There is an elegant theorem called the Nyquist-Shannon Sampling Theorem that says that human perception is limited, which allows that replacement of a continuous signal with a digital one without any perceived loss of information. This theorem also provides the *lowest* rate at which a signal can be sampled (called the Nyquist rate) without such a loss of information. The theorem states that we should sample at double the maximum desired frequency so that every cycle of the original signal will be sampled at at least two points. Recall that the frequency of a sinusoidal function is the reciprocal of the period. Identify the frequency of the function f and determine the number of partitions of the interval $[0, 10]$ that give us the Nyquist rate.
3. Humans cannot typically pick up signals above 20 kHz. Explain why, then, that information on a compact disk is sampled at 44,100 Hz.

8.2: Geometric Series

1. Seventh term of a geometric sequence

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. A geometric series

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. A series that is not geometric

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Two sums of geometric sequences

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5

There is an old question that is often used to introduce the power of geometric growth. Here is one version. Suppose you are hired for a one month (30 days, working every day) job and are given two options to be paid.

Option 1. available at the end of this section.

1. How much will you be paid for the job in total under Option 1?
2. Complete Table 8.2.3 to determine the pay you will receive under Option 2 for the first 10 days.

Table 8.2.3. Option 2 payments

Day	Pay on this day	Total amount paid to date

1	\dollar0.01	\dollar0.01
2	\dollar0.02	\dollar0.03
3		
4		
5		
6		
7		
8		
9		
10		

3. Find a formula for the amount paid on day n, as well as for the total amount paid by day n. Use this formula to determine which option (1 or 2) you should take.

6

Suppose you drop a golf ball onto a hard surface from a height h. The collision with the ground causes the ball to lose energy and so it will not bounce back to its original height. The ball will then fall again to the ground, bounce back up, and continue. Assume that at each bounce the ball rises back to a height $\frac{3}{4}$ of the height from which it dropped. Let h_n be the height of the ball on the nth bounce, with $h_0 = h$. In this exercise we will determine the distance traveled by the ball and the time it takes to travel that distance.

1. Determine a formula for h_1 in terms of h.
2. Determine a formula for h_2 in terms of h.
3. Determine a formula for h_3 in terms of h.
4. Determine a formula for h_n in terms of h.
5. Write an infinite series that represents the total distance traveled by the ball. Then determine the sum of this series.
6. Next, let's determine the total amount of time the ball is in the air.
    1. When the ball is dropped from a height H, if we assume the only force acting on it is the acceleration due to gravity, then the height of the ball at time t is given by
    $$H - \frac{1}{2}gt^2.$$
    Use this formula to determine the time it takes for the ball to hit the ground after being dropped from height H.
    2. Use your work in the preceding item, along with that in (a)-(e) above to determine the total amount of time the ball is in the air.

7

Suppose you play a game with a friend that involves rolling a standard six-sided die. Before a player can participate in the game, he or she must roll a six with the die. Assume that you roll first and that you and your friend take alternate rolls. In this exercise we will determine the probability that you roll the first six.

1. Explain why the probability of rolling a six on any single roll (including your first turn) is $\frac{1}{6}$.
2. If you don't roll a six on your first turn, then in order for you to roll the first six on your second turn, both you and your friend had to fail to roll a six on your first turns, and then you had to succeed in rolling a six on your second turn. Explain why the probability of this event is

$$\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right) = \left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right).$$

3. Now suppose you fail to roll the first six on your second turn. Explain why the probability is

$$\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right) = \left(\frac{5}{6}\right)^4\left(\frac{1}{6}\right)$$

that you to roll the first six on your third turn.

4. The probability of you rolling the first six is the probability that you roll the first six on your first turn plus the probability that you roll the first six on your second turn plus the probability that your roll the first six on your third turn, and so on. Explain why this probability is

$$\frac{1}{6} + \left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right) + \left(\frac{5}{6}\right)^4\left(\frac{1}{6}\right) + \cdots.$$

Find the sum of this series and determine the probability that you roll the first six.

8

The goal of a federal government stimulus package is to positively affect the economy. Economists and politicians quote numbers like "k million jobs and a net stimulus to the economy of n billion of dollars." Where do they get these numbers? Let's consider one aspect of a stimulus package: tax cuts. Economists understand that tax cuts or rebates can result in long-term spending that is many times the amount of the rebate. For example, assume that for a typical person, 75% of her entire income is spent (that is, put back into the economy). Further, assume the government provides a tax cut or rebate that totals P dollars for each person.

1. The tax cut of P dollars is income for its recipient. How much of this tax cut will be spent?
2. In this simple model, we will say that the spent portion of the tax cut/rebate from part (a) then becomes income for another person who, in turn, spends 75% of this income. After this ``second round" of spent income, how many total dollars have been added to the economy as a result of the original tax cut/rebate?
3. This second round of spending becomes income for another group who spend 75% of this income, and so on. In economics this is called the *multiplier effect*. Explain why an original tax cut/rebate of P dollars will result in multiplied spending of

$$0.75P(1 + 0.75 + 0.75^2 + \cdots).$$

dollars.

4. Based on these assumptions, how much stimulus will a 200 billion dollar tax cut/rebate to consumers add to the economy, assuming consumer spending remains consistent forever.

9

Like stimulus packages, home mortgages and foreclosures also impact the economy. A problem for many borrowers is the adjustable rate mortgage, in which the interest rate can change (and usually increases) over the duration of the loan, causing the monthly payments to increase beyond the ability of the borrower to pay. Most financial analysts recommend fixed rate loans, ones for which the monthly payments remain constant throughout the term of the loan. In this exercise we will analyze fixed rate loans.

When most people buy a large ticket item like car or a house, they have to take out a loan to make the purchase. The loan is paid back in monthly installments until the entire amount of the loan, plus interest, is paid. With a loan, we borrow money, say P dollars (called the principal), and pay off the loan at an interest rate of r%. To pay back the loan we make regular monthly payments, some of which goes to pay off the principal and some of which is charged as interest. In most cases, the interest is computed based on the amount of principal that remains at the beginning of the month. We assume a fixed rate loan, that is one in which we make a constant monthly payment M on our loan, beginning in the original month of the loan.

Suppose you want to buy a house. You have a certain amount of money saved to make a down payment, and you will borrow the rest to pay for the house. Of course, for the privilege of loaning you the money, the bank will charge you interest on this loan, so the amount you pay back to the bank is more than the amount you borrow. In fact, the amount you ultimately pay depends on three things: the amount you borrow (called the *principal*), the interest rate, and the length of time you have to pay off the loan plus interest (called the *duration* of the loan). For this example, we assume that the interest rate is fixed at r %.

To pay off the loan, each month you make a payment of the same amount (called *installments*). Suppose we borrow P dollars (our principal) and pay off the loan at an interest rate of r% with regular monthly installment payments of M dollars. So in month 1 of the loan, before we make any payments, our principal is P dollars. Our goal in this exercise is to find a formula that relates these three parameters to the time duration of the loan.

We are charged interest every month at an annual rate of r%, so each month we pay $\frac{r}{12}$% interest on the principal that remains. Given that the original principal is P dollars, we will pay $\left(\frac{0.0r}{12}\right)P$ dollars in interest on our first payment. Since we paid M dollars in total for our first payment, the remainder of the payment $(M-\left(\frac{r}{12}\right)P)$ goes to pay down the principal. So the principal remaining after the first payment (let's call it P_1) is the original principal minus what we paid on the principal, or

$$P_1 = P - \left(M - \left(\frac{r}{12}\right)P\right) = \left(1+\frac{r}{12}\right)P - M.$$

As long as P_1 is positive, we still have to keep making payments to pay off the loan.

1. Recall that the amount of interest we pay each time depends on the principal that remains. How much interest, in terms of P_1 and r, do we pay in the second installment?
2. How much of our second monthly installment goes to pay off the principal? What is the principal P_2, or the balance of the loan, that we still have to pay off after making the second installment of the loan? Write your response in the form $P_2 = (\ )P_1 - (\ )M$, where you fill in the parentheses.
3. Show that $P_2 = \left(1+\frac{r}{12}\right)^2 P - \left[1+\left(1+\frac{r}{12}\right)\right]M$.
4. Let P_3 be the amount of principal that remains after the third installment. Show that
$$P_3 = \left(1+\frac{r}{12}\right)^3 P - \left[1+\left(1+\frac{r}{12}\right)+\left(1+\frac{r}{12}\right)^2\right]M.$$
5. If we continue in the manner described in the problems above, then the remaining principal of our loan after n installments is
$$P_n = \left(1+\frac{r}{12}\right)^n P - \left[\sum_{k=0}^{n-1}\left(1+\frac{r}{12}\right)^k\right]M. \tag{8.2.7}$$
This is a rather complicated formula and one that is difficult to use. However, we can simplify the sum if we recognize part of it as a partial sum of a geometric series. Find a formula for the sum
$$\sum_{k=0}^{n-1}\left(1+\frac{r}{12}\right)^k. \tag{8.2.8}$$
and then a general formula for P_n that does not involve a sum.
6. It is usually more convenient to write our formula for P_n in terms of years rather than months. Show that $P(t)$, the principal remaining after t years, can be written as
$$P(t) = \left(P - \frac{12M}{r}\right)\left(1+\frac{r}{12}\right)^{12t} + \frac{12M}{r}. \tag{8.2.9}$$
7. Now that we have analyzed the general loan situation, we apply formula (8.2.9) to an actual loan. Suppose we charge \$1,000 on a credit card for holiday expenses. If our credit card charges 20% interest and we pay only the minimum payment of \$25 each month, how long will it take us to pay off the \$1,000 charge? How much in total will we have paid on this \$1,000 charge? How much total interest will we pay on this loan?

8. Now we consider larger loans, e.g., automobile loans or mortgages, in which we borrow a specified amount of money over a specified period of time. In this situation, we need to determine the amount of the monthly payment we need to make to pay off the loan in the specified amount of time. In this situation, we need to find the monthly payment M that will take our outstanding principal to 0 in the specified amount of time. To do so, we want to know the value of M that makes $P(t) = 0$ in formula (8.2.9). If we set $P(t) = 0$ and solve for M, it follows that

$$M = \frac{rP\left(1 + \frac{r}{12}\right)^{12t}}{12\left(\left(1 + \frac{r}{12}\right)^{12t} - 1\right)}.$$

   1. Suppose we want to borrow \$15,000 to buy a car. We take out a 5 year loan at 6.25%. What will our monthly payments be? How much in total will we have paid for this \$15,000 car? How much total interest will we pay on this loan?
   2. Suppose you charge your books for winter semester on your credit card. The total charge comes to \$525. If your credit card has an interest rate of 18% and you pay \$20 per month on the card, how long will it take before you pay off this debt? How much total interest will you pay?
   3. Say you need to borrow \$100,000 to buy a house. You have several options on the loan:
      - 30 years at 6.5%
      - 25 years at 7.5%
      - 15 years at 8.25%.
      1. What are the monthly payments for each loan?
      2. Which mortgage is ultimately the best deal (assuming you can afford the monthly payments)? In other words, for which loan do you pay the least amount of total interest?

8.3: Series of Real Numbers

1. Convergence of a sequence and its series

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Two partial sums

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Convergence of a series and its sequence

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Convergence of an integral and a related series

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5

In this exercise we investigate the sequence $\left\{\frac{b^n}{n!}\right\}$ for any constant b.

1. Use the Ratio Test to determine if the series $\sum \frac{10^k}{k!}$ converges or diverges.
2. Now apply the Ratio Test to determine if the series $\sum \frac{b^k}{k!}$ converges for any constant b.
3. Use your result from (b) to decide whether the sequence $\left\{\frac{b^n}{n!}\right\}$ converges or diverges. If the sequence $\left\{\frac{b^n}{n!}\right\}$ converges, to what does it converge? Explain your reasoning.

6

There is a test for convergence similar to the Ratio Test called the *Root Test*. Suppose we have a series $\sum a_k$ of positive terms so that $a_n \to 0$ as $n \to \infty$.

1. Assume

$$\sqrt[n]{a_n} \to r$$

as n goes to infinity. Explain why this tells us that $a_n \approx r^n$ for large values of n.

2. Using the result of part (a), explain why $\sum a_k$ looks like a geometric series when n is big. What is the ratio of the geometric series to which $\sum a_k$ is comparable?
3. Use what we know about geometric series to determine that values of r so that $\sum a_k$ converges if $\sqrt[n]{a_n} \to r$ as $n \to \infty$.

7

The associative and distributive laws of addition allow us to add finite sums in any order we want. That is, if $\sum_{k=0}^{n} a_k$ and $\sum_{k=0}^{n} b_k$ are finite sums of real numbers, then

$$\sum_{k=0}^{n} a_k + \sum_{k=0}^{n} b_k = \sum_{k=0}^{n} (a_k + b_k).$$

However, we do need to be careful extending rules like this to infinite series.

1. Let $a_n = 1 + \frac{1}{2^n}$ and $b_n = -1$ for each nonnegative integer n.
    1. Explain why the series $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ both diverge.
    2. Explain why the series $\sum_{k=0}^{\infty} (a_k + b_k)$ converges.
    3. Explain why

$$\sum_{k=0}^{\infty} a_k + \sum_{k=0}^{\infty} b_k \neq \sum_{k=0}^{\infty} (a_k + b_k).$$

    This shows that it is possible to have to two divergent series $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ but yet have the series $\sum_{k=0}^{\infty} (a_k + b_k)$ converge.

2. While part (a) shows that we cannot add series term by term in general, we can under reasonable conditions. The problem in part (a) is that we tried to add divergent series. In this exercise we will show that if $\sum a_k$ and $\sum b_k$ are convergent series, then $\sum(a_k + b_k)$ is a convergent series and
$$\sum(a_k + b_k) = \sum a_k + \sum b_k.$$
   1. Let A_n and B_n be the nth partial sums of the series $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$, respectively. Explain why
$$A_n + B_n = \sum_{k=1}^{n}(a_k + b_k).$$
   2. Use the previous result and properties of limits to show that
$$\sum_{k=1}^{\infty}(a_k + b_k) = \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} b_k.$$
   (Note that the starting point of the sum is irrelevant in this problem, so it doesn't matter where we begin the sum.)
3. Use the prior result to find the sum of the series $\sum_{k=0}^{\infty} \frac{2^k+3^k}{5^k}$.

8

In the Limit Comparison Test we compared the behavior of a series to one whose behavior we know. In that test we use the limit of the ratio of corresponding terms of the series to determine if the comparison is valid. In this exercise we see how we can compare two series directly, term by term, without using a limit of sequence. First we consider an example.

1. Consider the series
$$\sum \frac{1}{k^2} \text{ and } \sum \frac{1}{k^2+k}.$$
   We know that the series $\sum \frac{1}{k^2}$ is a p-series with $p = 2 > 1$ and so $\sum \frac{1}{k^2}$ converges. In this part of the exercise we will see how to use information about $\sum \frac{1}{k^2}$ to determine information about $\sum \frac{1}{k^2+k}$. Let $a_k = \frac{1}{k^2}$ and $b_k = \frac{1}{k^2+k}$.
   1. Let S_n be the nth partial sum of $\sum \frac{1}{k^2}$ and T_n the nth partial sum of $\sum \frac{1}{k^2+k}$. Which is larger, S_1 or T_1? Why?
   2. Recall that
$$S_2 = S_1 + a_2 \text{ and } T_2 = T_1 + b_2.$$
      Which is larger, a_2 or b_2? Based on that answer, which is larger, S_2 or T_2?
   3. Recall that
$$S_3 = S_2 + a_3 \text{ and } T_3 = T_2 + b_3.$$
      Which is larger, a_3 or b_3? Based on that answer, which is larger, S_3 or T_3?
   4. Which is larger, a_n or b_n? Explain. Based on that answer, which is larger, S_n or T_n?
   5. Based on your response to the previous part of this exercise, what relationship do you expect there to be between $\sum \frac{1}{k^2}$ and $\sum \frac{1}{k^2+k}$? Do you expect $\sum \frac{1}{k^2+k}$ to converge or diverge? Why?
2. The example in the previous part of this exercise illustrates a more general result. Explain why the Direct Comparison Test, stated here, works.

The Direct Comparison Test

If
$$0 \le b_k \le a_k$$
for every k, then we must have

$$0 \leq \sum b_k \leq \sum a_k$$

1. If $\sum a_k$ converges, then $\sum b_k$ converges.
2. If $\sum b_k$ diverges, then $\sum a_k$ diverges.

Note 8.3.7

This comparison test applies only to series with nonnegative terms.

1. Use the Direct Comparison Test to determine the convergence or divergence of the series $\sum \frac{1}{k-1}$. Hint: Compare to the harmonic series.
2. Use the Direct Comparison Test to determine the convergence or divergence of the series $\sum \frac{k}{k^3+1}$.

8.4: Alternating Series

1. Testing convergence for an alternating series

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

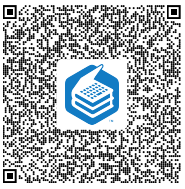
View this Interactive Content online



2. Estimating the sum of an alternating series

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3. Estimating the sum of a different alternating series

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Estimating the sum of one more alternating series

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5

Conditionally convergent series converge very slowly. As an example, consider the famous formula [1]

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{1}{2k+1}. \tag{8.4.3}$$

We will derive this formula in upcoming work.

In theory, the partial sums of this series could be used to approximate π.

1. Show that the series in (8.4.3) converges conditionally.
2. Let S_n be the nth partial sum of the series in (8.4.3). Calculate the error in approximating $\frac{\pi}{4}$ with S_{100} and explain why this is not a very good approximation.
3. Determine the number of terms it would take in the series (8.4.3) to approximate $\frac{\pi}{4}$ to 10 decimal places. (The fact that it takes such a large number of terms to obtain even a modest degree of accuracy is why we say that conditionally convergent series converge very slowly.)

6

We have shown that if $\sum (-1)^{k+1} a_k$ is a convergent alternating series, then the sum S of the series lies between any two consecutive partial sums S_n. This suggests that the average $\frac{S_n + S_{n+1}}{2}$ is a better approximation to S than is S_n.

1. Show that $\frac{S_n + S_{n+1}}{2} = S_n + \frac{1}{2}(-1)^{n+2} a_{n+1}$.
2. Use this revised approximation in (a) with $n = 20$ to approximate $\ln(2)$ given that

$$\ln(2) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k}.$$

   Compare this to the approximation using just S_{20}. For your convenience, $S_{20} = \frac{155685007}{232792560}$.

7

In this exercise, we examine one of the conditions of the Alternating Series Test. Consider the alternating series

$$1 - 1 + \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{9} + \frac{1}{4} - \frac{1}{16} + \cdots,$$

where the terms are selected alternately from the sequences $\left\{\frac{1}{n}\right\}$ and $\left\{-\frac{1}{n^2}\right\}$.

1. Explain why the nth term of the given series converges to 0 as n goes to infinity.
2. Rewrite the given series by grouping terms in the following manner:

$$(1 - 1) + \left(\frac{1}{2} - \frac{1}{4}\right) + \left(\frac{1}{3} - \frac{1}{9}\right) + \left(\frac{1}{4} - \frac{1}{16}\right) + \cdots.$$

   Use this regrouping to determine if the series converges or diverges.

3. Explain why the condition that the sequence $\{a_n\}$ *decreases* to a limit of 0 is included in the Alternating Series Test.

8

Conditionally convergent series exhibit interesting and unexpected behavior. In this exercise we examine the conditionally convergent alternating harmonic series $\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}$ and discover that addition is not commutative for conditionally convergent series. We will also encounter Riemann's Theorem concerning rearrangements of conditionally convergent series. Before we begin, we remind ourselves that

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} = \ln(2),$$

a fact which will be verified in a later section.

1. First we make a quick analysis of the positive and negative terms of the alternating harmonic series.
   1. Show that the series $\sum_{k=1}^{\infty} \frac{1}{2k}$ diverges.
   2. Show that the series $\sum_{k=1}^{\infty} \frac{1}{2k+1}$ diverges.
   3. Based on the results of the previous parts of this exercise, what can we say about the sums $\sum_{k=C}^{\infty} \frac{1}{2k}$ and $\sum_{k=C}^{\infty} \frac{1}{2k+1}$ for any positive integer C? Be specific in your explanation.
2. Recall addition of real numbers is commutative; that is

   $$a + b = b + a$$

   for any real numbers a and b. This property is valid for any sum of finitely many terms, but does this property extend when we add infinitely many terms together?

   The answer is no, and something even more odd happens. Riemann's Theorem (after the nineteenth-century mathematician Georg Friedrich Bernhard Riemann) states that a conditionally convergent series can be rearranged to converge to *any* prescribed sum. More specifically, this means that if we choose any real number S, we can rearrange the terms of the alternating harmonic series $\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}$ so that the sum is S. To understand how Riemann's Theorem works, let's assume for the moment that the number S we want our rearrangement to converge to is positive. Our job is to find a way to order the sum of terms of the alternating harmonic series to converge to S.

   1. Explain how we know that, regardless of the value of S, we can find a partial sum P_1

      $$P_1 = \sum_{k=1}^{n_1} \frac{1}{2k+1} = 1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{2n_1+1}$$

      of the positive terms of the alternating harmonic series that equals or exceeds S. Let

      $$S_1 = P_1.$$

   2. Explain how we know that, regardless of the value of S_1, we can find a partial sum N_1

      $$N_1 = -\sum_{k=1}^{m_1} \frac{1}{2k} = -\frac{1}{2} - \frac{1}{4} - \frac{1}{6} - \cdots - \frac{1}{2m_1}$$

      so that

      $$S_2 = S_1 + N_1 \leq S.$$

   3. Explain how we know that, regardless of the value of S_2, we can find a partial sum P_2

      $$P_2 = \sum_{k=n_1+1}^{n_2} \frac{1}{2k+1} = \frac{1}{2(n_1+1)+1} + \frac{1}{2(n_1+2)+1} + \cdots + \frac{1}{2n_2+1}$$

      of the remaining positive terms of the alternating harmonic series so that

      $$S_3 = S_2 + P_2 \geq S.$$

   4. Explain how we know that, regardless of the value of S_3, we can find a partial sum

$$N_2 = -\sum_{k=m_1+1}^{m_2} \frac{1}{2k} = -\frac{1}{2(m_1+1)} - \frac{1}{2(m_1+2)} - \cdots - \frac{1}{2m_2}$$

of the remaining negative terms of the alternating harmonic series so that

$$S_4 = S_3 + N_2 \leq S.$$

5. Explain why we can continue this process indefinitely and find a sequence $\{S_n\}$ whose terms are partial sums of a rearrangement of the terms in the alternating harmonic series so that $\lim_{n\to\infty} S_n = S$.

8.5: Taylor Polynomials and Taylor Series

1. Determining Taylor polynomials from a function formula

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. Determining Taylor polynomials from given derivative values

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

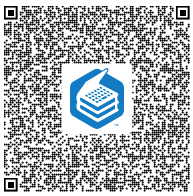
View this Interactive Content online



3. Finding the Taylor series for a given rational function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Finding the Taylor series for a given trigonometric function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



5. Finding the Taylor series for a given logarithmic function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

6

In this exercise we investigation the Taylor series of polynomial functions.

1. Find the 3rd order Taylor polynomial centered at $a = 0$ for $f(x) = x^3 - 2x^2 + 3x - 1$. Does your answer surprise you? Explain.
2. Without doing any additional computation, find the 4th, 12th, and 100th order Taylor polynomials (centered at $a = 0$) for $f(x) = x^3 - 2x^2 + 3x - 1$. Why should you expect this?
3. Now suppose $f(x)$ is a degree m polynomial. Completely describe the nth order Taylor polynomial (centered at $a = 0$) for each n.

7

The examples we have considered in this section have all been for Taylor polynomials and series centered at 0, but Taylor polynomials and series can be centered at any value of a. We look at examples of such Taylor polynomials in this exercise.

1. Let $f(x) = \sin(x)$. Find the Taylor polynomials up through order four of f centered at $x = \frac{\pi}{2}$. Then find the Taylor series for $f(x)$ centered at $x = \frac{\pi}{2}$. Why should you have expected the result?
2. Let $f(x) = \ln(x)$. Find the Taylor polynomials up through order four of f centered at $x = 1$. Then find the Taylor series for $f(x)$ centered at $x = 1$.
3. Use your result from (b) to determine which Taylor polynomial will approximate $\ln(2)$ to two decimal places. Explain in detail how you know you have the desired accuracy.

8

We can use known Taylor series to obtain other Taylor series, and we explore that idea in this exercise, as a preview of work in the following section.

1. Calculate the first four derivatives of $\sin(x^2)$ and hence find the fourth order Taylor polynomial for $\sin(x^2)$ centered at $a = 0$.
2. Part (a) demonstrates the brute force approach to computing Taylor polynomials and series. Now we find an easier method that utilizes a known Taylor series. Recall that the Taylor series centered at 0 for $f(x) = \sin(x)$ is

$$\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}. \tag{8.5.7}$$

    1. Substitute x^2 for x in the Taylor series (8.5.7). Write out the first several terms and compare to your work in part (a). Explain why the substitution in this problem should give the Taylor series for $\sin(x^2)$ centered at 0.
    2. What should we expect the interval of convergence of the series for $\sin(x^2)$ to be? Explain in detail.

9

Based on the examples we have seen, we might expect that the Taylor series for a function f always converges to the values $f(x)$ on its interval of convergence. We explore that idea in more detail in this exercise. Let $f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$

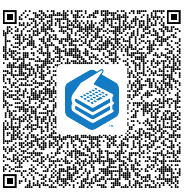
1. Show, using the definition of the derivative, that $f'(0) = 0$.
2. It can be shown that $f^{(n)}(0) = 0$ for all $n \geq 2$. Assuming that this is true, find the Taylor series for f centered at 0.
3. What is the interval of convergence of the Taylor series centered at 0 for f? Explain. For which values of x the interval of convergence of the Taylor series does the Taylor series converge to $f(x)$?

8.6: Power Series

1. Finding coefficients in a power series expansion of a rational function

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



2. Finding coefficients in a power series expansion of a function involving $\arctan(x)$

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



3.

We can use power series to approximate definite integrals to which known techniques of integration do not apply. We will illustrate this in this exercise with the definite integral $\int_0^1 \sin(x^2)\ ds$.

1. Use the Taylor series for $\sin(x)$ to find the Taylor series for $\sin(x^2)$. What is the interval of convergence for the Taylor series for $\sin(x^2)$? Explain.
2. Integrate the Taylor series for $\sin(x^2)$ term by term to obtain a power series expansion for $\int \sin(x^2)\ dx$.
3. Use the result from part (b) to explain how to evaluate $\int_0^1 \sin(x^2)\ dx$. Determine the number of terms you will need to approximate $\int_0^1 \sin(x^2)\ dx$ to 3 decimal places.

4

There is an important connection between power series and Taylor series. Suppose f is defined by a power series centered at 0 so that

$$f(x) = \sum_{k=0}^{\infty} a_k x^k.$$

1. Determine the first 4 derivatives of f evaluated at 0 in terms of the coefficients a_k.
2. Show that $f^{(n)}(0) = n!a_n$ for each positive integer n.
3. Explain how the result of (b) tells us the following:

   On its interval of convergence, a power series is the Taylor series of its sum.

5

In this exercise we will begin with a strange power series and then find its sum. The Fibonacci sequence $\{f_n\}$ is a famous sequence whose first few terms are

$$f_0 = 0, f_1 = 1, f_2 = 1, f_3 = 2, f_4 = 3, f_5 = 5, f_6 = 8, f_7 = 13, \cdots,$$

where each term in the sequence after the first two is the sum of the preceding two terms. That is, $f_0 = 0,\ f_1 = 1$ and for $n \geq 2$ we have

$$f_n = f_{n-1} + f_{n-2}.$$

Now consider the power series

$$F(x) = \sum_{k=0}^{\infty} f_k x^k.$$

We will determine the sum of this power series in this exercise.

1. Explain why each of the following is true.
   1. $xF(x) = \sum_{k=1}^{\infty} f_{k-1}x^k$
   2. $x^2F(x) = \sum_{k=2}^{\infty} f_{k-2}x^k$
2. Show that
   $$F(x) - xF(x) - x^2F(x) = x.$$
3. Now use the equation
   $$F(x) - xF(x) - x^2F(x) = x$$
   to find a simple form for $F(x)$ that doesn't involve a sum.
4. Use a computer algebra system or some other method to calculate the first 8 derivatives of $\frac{x}{1-x-x^2}$ evaluated at 0. Why shouldn't the results surprise you?

6

Airy's equation [2]

$$y'' - xy = 0, \tag{8.6.4}$$

can be used to model an undamped vibrating spring with spring constant x (note that y is an unknown function of x). So the solution to this differential equation will tell us the behavior of a spring-mass system as the spring ages (like an automobile shock absorber). Assume that a solution $y = f(x)$ has a Taylor series that can be written in the form

$$y = \sum_{k=0}^{\infty} a_k x^k,$$

where the coefficients are undetermined. Our job is to find the coefficients.

The general differential equations of the form $y'' \pm k^2 xy = 0$ is called Airy's equation. These equations arise in many problems, such as the study of diffraction of light, diffraction of radio waves around an object, aerodynamics, and the buckling of a uniform column under its own weight.

1. Differentiate the series for y term by term to find the series for y'. Then repeat to find the series for y''.
2. Substitute your results from part (a) into the Airy equation and show that we can write Equation (8.6.4) in the form
$$\sum_{k=2}^{\infty}(k-1)ka_k x^{k-2} - \sum_{k=0}^{\infty} a_k x^{k+1} = 0. \tag{8.6.5}$$
3. At this point, it would be convenient if we could combine the series on the left in (8.6.5), but one written with terms of the form x^{k-2} and the other with terms in the form x^{k+1}. Explain why
$$\sum_{k=2}^{\infty}(k-1)ka_k x^{k-2} = \sum_{k=0}^{\infty}(k+1)(k+2)a_{k+2}x^k. \tag{8.6.6}$$
4. Now show that
$$\sum_{k=0}^{\infty} a_k x^{k+1} = \sum_{k=1}^{\infty} a_{k-1}x^k. \tag{8.6.7}$$
5. We can now substitute (8.6.6) and (8.6.7) into (8.6.5) to obtain
$$\sum_{n=0}^{\infty}(n+1)(n+2)a_{n+2}x^n - \sum_{n=1}^{\infty} a_{n-1}x^n = 0. \tag{8.6.8}$$
Combine the like powers of x in the two series to show that our solution must satisfy
$$2a_2 + \sum_{k=1}^{\infty}\left[(k+1)(k+2)a_{k+2} - a_{k-1}\right]x^k = 0. \tag{8.6.9}$$
6. Use equation (8.6.9) to show the following:
   1. $a_{3k+2} = 0$ for every positive integer k,
   2. $a_{3k} = \frac{1}{(2)(3)(5)(6)\cdots(3k-1)(3k)}a_0$ for $k \geq 1$,
   3. $a_{3k+1} = \frac{1}{(3)(4)(6)(7)\cdots(3k)(3k+1)}a_1$ for $k \geq 1$.
7. Use the previous part to conclude that the general solution to the Airy equation (8.6.4) is
$$y = a_0\left(1 + \sum_{k=1}^{\infty}\frac{x^{3k}}{(2)(3)(5)(6)\cdots(3k-1)(3k)}\right) + a_1\left(x + \sum_{k=1}^{\infty}\frac{x^{3k+1}}{(3)(4)(6)(7)\cdots(3k)(3k+1)}\right).$$
Any values for a_0 and a_1 then determine a specific solution that we can approximate as closely as we like using this series solution.

Answers

5

You can be paid $500 per day or

You can be paid 1 cent the first day, 2 cents the second day, 4 cents the third day, 8 cents the fourth day, and so on, doubling the amount you are paid each day.

This page titled 8.7: Sequences and Series (Exercises) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.E: Sequences and Series (Exercises) has no license indicated.
- 8.1: Sequences by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 8.2: Geometric Series by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 8.3: Series of Real Numbers by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 8.4: Alternating Series by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 8.5: Taylor Polynomials and Taylor Series by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.
- 8.6: Power Series by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.

9: MULTIVARIABLE AND VECTOR FUNCTIONS

http://scholarworks.gvsu.edu/books/14/

9.1: Functions of Several Variables and Three Dimensional Space

9.2: Vectors

9.3: Dot Product

9.4: The Cross Product

9.5: Lines and Planes in Space

9.6: Vector-Valued Functions

9.7: Derivatives and Integrals of Vector-Valued Functions

9.8: Arc Length and Curvature

This page titled 9: Multivariable and Vector Functions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 9: Multivariable and Vector Functions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

9.1: Functions of Several Variables and Three Dimensional Space

Motivating Questions

- What is a function of several variables? What do we mean by the domain of a function of several variables?
- How do we find the distance between two points in $\mathbb{R}^3$? What is the equation of a sphere in $\mathbb{R}^3$?
- What is a trace of a function of two variables? What does a trace tell us about a function?
- What is a level curve of a function of two variables? What does a level curve tell us about a function?

Throughout our mathematical careers we have studied functions of a single variable. We define a function of one variable as a rule that assigns exactly one output to each input. We analyze these functions by looking at their graphs, calculating limits, differentiating, integrating, and more. Functions of several variables will be the main focus of Chapters 10 and 11, where we will analyze these functions by looking at their graphs, calculating limits, differentiating, integrating, and more. We will see that many of the ideas from single variable calculus translate well to functions of several variables, but we will have to make some adjustments as well. In this chapter we introduce functions of several variables and then discuss some of the tools (vectors and vector-valued functions) that will help us understand and analyze functions of several variables.

Preview Activity 9.1.1

Suppose you invest money in an account that pays 5% interest compounded continuously. If you invest P dollars in the account, the amount A of money in the account after t years is given by

$$A = Pe^{0.05t}.$$

The variables P and t are independent of each other, so using functional notation we write

$$A(P, t) = Pe^{0.05t}.$$

1. Find the amount of money in the account after 7 years if you originally invest 1000 dollars.
2. Evaluate $A(5000, 8)$. Explain in words what this calculation represents.
3. Now consider only the situation where the amount invested is fixed at 1000 dollars. Calculate the amount of money in the account after t years as indicated in Table 9.1.1. Round payments to the nearest penny.

   Table 9.1.1. Amount of money in an account with an initial investment of 1000 dollars.

Duration (in years)	2	3	4	5	6
Amount (dollars)					

4. Now consider the situation where we want to know the amount of money in the account after 10 years given various initial investments. Calculate the amount of money in the account as indicated in Table 9.1.2. Round payments to the nearest penny.

   Table 9.1.2. Amount of money in an account after 10 years.

Initial investment (dollars)	500	1000	5000	7500	10000
Amount (dollars)					

5. Describe as best you can the combinations of initial investments and time that result in an account containing $10,000.

Functions of Several Variables

Up to this point we have been concerned with functions of a single variable. What defined such a function is that every input in the domain produced a unique output in the range. We saw similar behavior in Preview Activity 9.1.1, where each pair (P, t) of inputs produces a unique output $A(P, t)$. Additionally, the two variables P and t had no real relation to each other. That is, we could choose any value of P without considering what value t might have, and we could select any value of t to use without regard to what value P might have. For that reason we say that the variables t and P are *independent* of each other. Thus, we call $A = A(P, t)$ a function of the two independent variables P and t. This is the key idea in defining a function of two independent variables.

Definition 9.1.1

A ***function f of two independent variables*** is a rule that assigns to each ordered pair (x, y) in some set D exactly one real number $f(x, y)$.

There is, of course, no reason to restrict ourselves to functions of only two variables—we can use any number of variables we like. For example,

$$f(x, y, z) = x^2 - 2xz + \cos(y)$$

defines f as a function of the three variables x, y, and z. In general, a function of n independent variables is a rule that assigns to an ordered n-tuple $(x_1, x_2, \ldots, x_n)$ in some set D exactly one real number.

As with functions of a single variable, it is important to understand the set of inputs for which the function is defined.

Definition 9.1.2

The ***domain*** of a function f is the set of all inputs at which the function is defined.

Activity 9.1.2

Identify the domain of each of the following functions. Draw a picture of each domain in the xy-plane.

1. $f(x, y) = x^2 + y^2$
2. $f(x, y) = \sqrt{x^2 + y^2}$
3. $Q(x, y) = \dfrac{x + y}{x^2 - y^2}$
4. $s(x, y) = \dfrac{1}{\sqrt{1 - xy^2}}$

Representing Functions of Two Variables

One of the techniques we use to study functions of one variable is to create a table of values. We can do the same for functions of two variables, except that our tables will have to allow us to keep track of both input variables. We can do this with a 2-dimensional table, where we list the x-values down the first column and the y-values across the first row. As an example, suppose we launch a projectile, using a golf club, a cannon, or some other device, from ground level. Under ideal conditions (ignoring wind resistance, spin, or any other forces except the force of gravity) the horizontal distance the object will travel depends on the initial velocity x the object is given, and the angle y at which it is launched. If we let f represent the horizontal distance the object travels, then f is a function of the two variables x and y, and we represent f in functional notation by

$$f(x, y) = \frac{x^2 \sin(2y)}{g},$$

where g is the acceleration due to gravity. (Note that g is constant, 32 feet per second squared. We will derive this equation in a

later section.) To create a table of values for f, we list the x-values down the first column and the y-values across the first row. The value $f(x, y)$ is then displayed in the location where the x row intersects the y column, as shown in Table 9.1.3 (where we measure x in feet per second and y in radians).

Table 9.1.3. *Values of* $f(x, y) = \frac{x^2 \sin(2y)}{g}$.

$x \backslash y$	0.2	0.4	0.6	0.8	1.0	1.2	1.4
25	7.6	14.0	18.2	19.5	17.8	13.2	6.5
50	30.4	56.0	72.8	78.1	71.0		26.2
75	68.4		163.8	175.7	159.8	118.7	58.9
100	121.7	224.2	291.3	312.4	284.2	211.1	104.7
125	190.1	350.3	455.1		444.0	329.8	163.6
150	273.8	504.4	655.3	702.8	639.3	474.9	235.5
175	372.7	686.5	892.0	956.6	870.2	646.4	
200	486.8	896.7	1165.0	1249.5	1136.6	844.3	418.7
225	616.2	1134.9	1474.5	1581.4	1438.5	1068.6	530.0
250	760.6	1401.1		1952.3	1776.0	1319.3	654.3

Activity 9.1.3

Complete Table 9.1.3 by filling in the missing values of the function f. Round entries to the nearest tenth.

If f is a function of a single variable x, then we define the graph of f to be the set of points of the form $(x, f(x))$, where x is in the domain of f. We then plot these points using the coordinate axes in order to visualize the graph. We can do a similar thing with functions of several variables. Table 9.1.5 identifies points of the form $(x, y, f(x, y))$, and we define the graph of f to be the set of these points.

Definition 9.1.3: Graph

The ***graph*** of a function $f = f(x, y)$ is the set of points of the form $(x, y, f(x, y))$, where the point (x, y) is in the domain of f.

We also often refer to the graph of a function f of two variables as the *surface* generated by f. Points in the form $(x,\ y,\ f(x, y))$ are in three dimensions, so plotting these points takes a bit more work than graphs of functions in two dimensions. To plot these three-dimensional points, we need to set up a coordinate system with three mutually perpendicular axes — the x-axis, the y-axis, and the z-axis (called the *coordinate axes*). There are essentially two different ways we could set up a 3D coordinate system, as shown in Figure 9.1.1; thus, before we can proceed, we need to establish a convention.

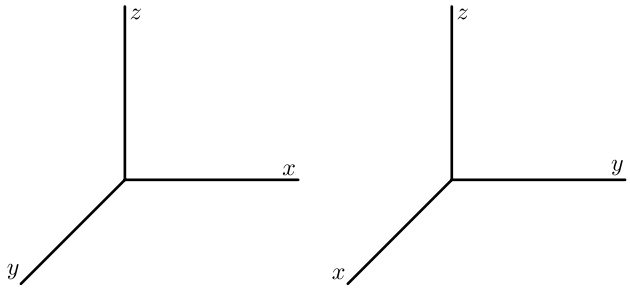


Figure 9.1.1. Left: A left hand system. Right: A right hand system

The distinction between these two figures is subtle, but important. In the coordinate system shown at left in Figure 9.1.7, imagine that you are sitting on the positive z-axis next to the label "z." Looking down at the x- and y-axes, you see that the y-axis is obtained by rotating the x-axis by 90° in the *clockwise* direction. Again sitting on the positive z-axis in the coordinate system at right in Figure 9.1.7, you see that the y-axis is obtained by rotating the x-axis by 90° in the *counterclockwise* direction.

We call the coordinate system at right in Figure 9.1.7 a *right-hand system*; if we point the index finger of our *right* hand along the positive x-axis and our middle finger along the positive y-axis, then our thumb points in the direction of the positive z-axis. Following mathematical conventions, we choose to use a right-hand system throughout this book.

Now that we have established a convention for a right-hand system, we can draw a graph of the distance function defined by $f(x,y) = \frac{x^2 \sin(2y)}{g}$. Note that the function f is continuous in both variables, so when we plot these points in the right hand coordinate system, we can connect them all to form a surface in 3-space. The graph of the distance function f is shown in Figure 9.1.2.

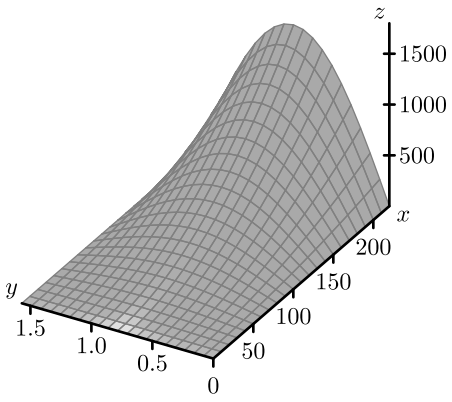


Figure 9.1.2. The distance surface.

There are many graphing tools available for drawing three-dimensional surfaces as indicated in the Preface (see Links to interactive graphics in Features of the Text). Since we will be able to visualize graphs of functions of two independent variables, but not functions of more than two variables, we will primarily deal with functions of two variables in this text. It is important to note, however, that the techniques we develop apply to functions of any number of variables.

Notation: We let $\mathbb{R}^2$ denote the set of all ordered pairs of real numbers in the plane (two copies of the real number system) and let $\mathbb{R}^3$ represent the set of all ordered triples of real numbers (which constitutes three-space).

Some Standard Equations in Three-Space

In addition to graphing functions, we will also want to understand graphs of some simple equations in three dimensions. For example, in $\mathbb{R}^2$, the graphs of the equations $x = a$ and $y = b$, where a and b are constants, are lines parallel to the coordinate axes. In the next activity we consider their three-dimensional analogs.

Activity 9.1.4

1. Consider the set of points (x, y, z) that satisfy the equation $x = 2$. Describe this set as best you can.
2. Consider the set of points (x, y, z) that satisfy the equation $y = -1$. Describe this set as best you can.
3. Consider the set of points (x, y, z) that satisfy the equation $z = 0$. Describe this set as best you can.

Activity 9.1.4 shows that the equations where one independent variable is constant lead to planes parallel to ones that result from a pair of the coordinate axes. When we make the constant 0, we get the *coordinate planes*. The xy-plane satisfies $z = 0$, the xz -plane satisfies $y = 0$, and the yz-plane satisfies $x = 0$ (see Figure 9.1.3).

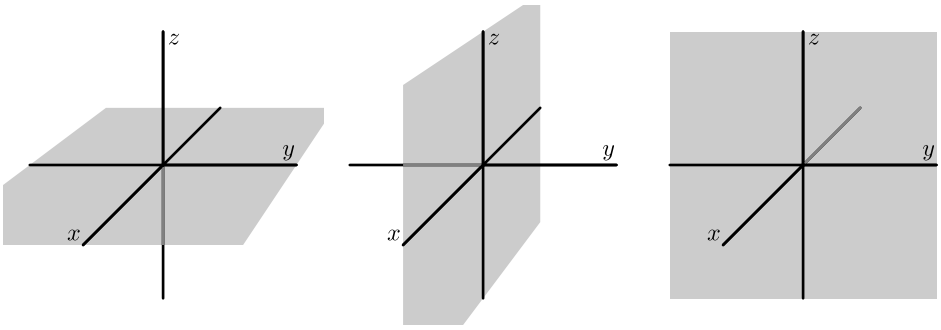


Figure 9.1.3. The coordinate planes.

On a related note, we define a circle in $\mathbb{R}^2$ as the set of all points equidistant from a fixed point. In $\mathbb{R}^3$, we call the set of all points equidistant from a fixed point a *sphere*. To find the equation of a sphere, we need to understand how to calculate the distance between two points in three-space, and we explore this idea in the next activity.

Activity 9.1.5

Let $P = (x_0, y_0, z_0)$ and $Q = (x_1, y_1, z_1)$ be two points in $\mathbb{R}^3$. These two points form opposite vertices of a rectangular box whose sides are planes parallel to the coordinate planes as illustrated in Figure 9.1.10, and the distance between P and Q is the length of the blue diagonal shown in Figure 9.1.4.

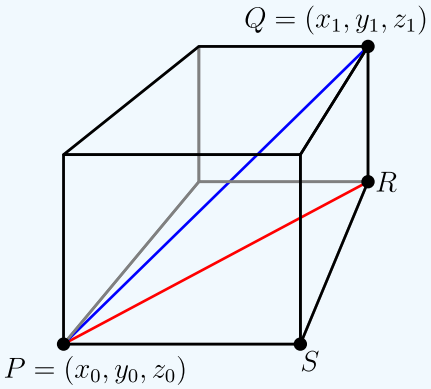


Figure 9.1.4. The distance formula in $\mathbb{R}^3$.

1. Consider the right triangle PRS in the base of the box whose hypotenuse is shown as the red line in Figure 9.1.10. What are the coordinates of the vertices of this triangle? Since this right triangle lies in a plane, we can use the Pythagorean Theorem to find a formula for the length of the hypotenuse of this triangle. Find such a formula, which will be in terms of x_0, y_0, x_1, and y_1.
2. Now notice that the triangle PRQ whose hypotenuse is the blue segment connecting the points P and Q with a leg as the hypotenuse PR of the triangle found in part (a) lies entirely in a plane, so we can again use the Pythagorean Theorem to find the length of its hypotenuse. Explain why the length of this hypotenuse, which is the distance between the points P and Q, is

$$\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}.$$

The formula developed in Activity 9.1.5 is important to remember.

The distance between points

The distance between points $P = (x_0, y_0, z_0)$ and $Q = (x_1, y_1, z_1)$ (denoted as $|PQ|$) in $\mathbb{R}^3$ is given by the formula

$$|PQ| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}. \tag{9.1.1}$$

Equation $(9.1.1)$ can be used to derive the formula for a sphere centered at a point (x_0, y_0, z_0) with radius r. Since the distance from any point (x, y, z) on such a sphere to the point (x_0, y_0, z_0) is r, the point (x, y, z) will satisfy the equation

$$\sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2} = r$$

Squaring both sides, we come to the standard equation for a sphere.

The equation of a sphere

The equation of a sphere with center (x_0, y_0, z_0) and radius r is

$$(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = r^2.$$

This makes sense if we compare this equation to its two-dimensional analogue, the equation of a circle of radius r in the plane centered at (x_0, y_0):

$$(x - x_0)^2 + (y - y_0)^2 = r^2.$$

Traces

When we study functions of several variables we are often interested in how each individual variable affects the function in and of itself. In Preview Activity $9.1.1$, we saw that the amount of money in an account depends on the amount initially invested and the duration of the investment. However, if we fix the initial investment, the amount of money in the account depends only on the duration of the investment, and if we set the duration of the investment constant, then the amount of money in the account depends only on the initial investment. This idea of keeping one variable constant while we allow the other to change will be an important tool for us when studying functions of several variables.

As another example, consider again the distance function f defined by

$$f(x, y) = \frac{x^2 \sin(2y)}{g}$$

where x is the initial velocity of an object in feet per second, y is the launch angle in radians, and g is the acceleration due to gravity (32 feet per second squared). If we hold the launch angle constant at $y = 0.6$ radians, we can consider f a function of the initial velocity alone. In this case we have

$$f(x) = \frac{x^2}{32}\sin(2 \cdot 0.6).$$

We can plot this curve on the surface by tracing out the points on the surface when $y = 0.6$, as shown at left in Figure $9.1.11$. The formula clearly shows that f is quadratic in the x-direction. More descriptively, as we increase the launch velocity while keeping the launch angle constant, the horizontal distance the object travels increases proportional to the square of the initial velocity.

Similarly, if we fix the initial velocity at 150 feet per second, we can consider the distance as a function of the launch angle only. In this case we have

$$f(y) = \frac{150^2 \sin(2y)}{32}.$$

We can again plot this curve on the surface by tracing out the points on the surface when $x = 150$, as shown at right in Figure

9.1.5. The formula clearly show that f is sinusoidal in the y-direction. More descriptively, as we increase the launch angle while keeping the initial velocity constant, the horizontal distance traveled by the object is proportional to the sine of twice the launch angle.

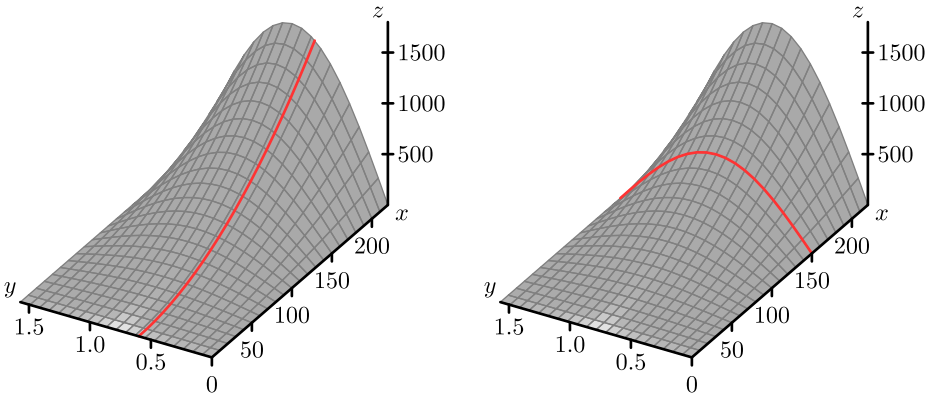


Figure 9.1.51. Left: The trace with $y = 0.6$. Right: The trace with $x = 150$.

The curves we define when we fix one of the independent variables in our two variable function are called *traces*.

Definition 9.1.4

A ***trace*** of a function f of two independent variables x and y in the x direction is a curve of the form $z = f(x, c)$, where c is a constant. Similarly, a ***trace*** of a function f of two independent variables x and y in the y direction is a curve of the form $z = f(c, y)$, where c is a constant.

Understanding trends in the behavior of functions of two variables can be challenging, as can sketching their graphs; traces help us with each of these tasks.

Activity 9.1.6

In the following questions, we investigate the use of traces to better understand a function through both tables and graphs.

a. Identify the $y = 0.6$ trace for the distance function f defined by $f(x, y) = \frac{x^2 \sin(2y)}{g}$ by highlighting or circling the appropriate cells in Table 9.1.3. Write a sentence to describe the behavior of the function along this trace.

b. Identify the $x = 150$ trace for the distance function by highlighting or circling the appropriate cells in Table 9.1.3. Write a sentence to describe the behavior of the function along this trace.

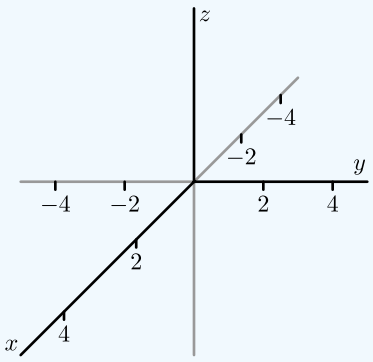


Figure 9.1.6. Coordinate axes to sketch traces.

c. For the function g defined by $g(x,y) = x^2 + y^2 + 1$, explain the type of function that each trace in the x direction will be (keeping y constant). Plot the $y = -4$, $y = -2$, $y = 0$, $y = 2$, and $y = 4$ traces in 3-dimensional coordinate system provided in Figure 9.1.13.

d. For the function g defined by $g(x,y) = x^2 + y^2 + 1$, explain the type of function that each trace in the y direction will be (keeping x constant). Plot the $x = -4$, $x = -2$, $x = 0$, $x = 2$, and $x = 4$ traces in 3-dimensional coordinate system in Figure 9.1.6.

e. Describe the surface generated by the function g.

Contour Maps and Level Curves

We have all seen topographic maps such as the one of the Porcupine Mountains in the upper peninsula of Michigan shown in Figure 9.1.7.[1] The curves on these maps show the regions of constant altitude. The contours also depict changes in altitude: contours that are close together signify steep ascents or descents, while contours that are far apart indicate only slight changes in elevation. Thus, contour maps tell us a lot about three-dimensional surfaces. Mathematically, if $f(x,y)$ represents the altitude at the point (x,y), then each contour is the graph of an equation of the form $f(x,y) = k$, for some constant k.

Map source: Michigan Department of Natural Resources, with permission of the Michigan DNR and Bob Wild.

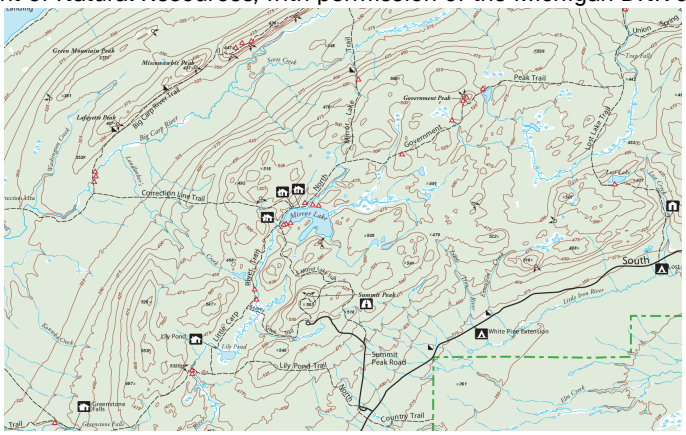


Figure 9.1.7. Contour map of the Porcupine Mountains.

Activity 9.1.7

On the topographical map of the Porcupine Mountains in Figure 9.1.7,

1. identify the highest and lowest points you can find;
2. from a point of your choice, determine a path of steepest ascent that leads to the highest point;
3. from that same initial point, determine the least steep path that leads to the highest point.

Curves on a surface that describe points at the same height or level are called *level curves*.

Definition 9.1.5

A ***level curve (or contour)*** of a function f of two independent variables x and y is a curve of the form $k = f(x,y)$, where k is a constant.

Topographical maps can be used to create a three-dimensional surface from the two-dimensional contours or level curves. For

example, level curves of the distance function defined by $f(x,y) = \frac{x^2 \sin(2y)}{32}$ plotted in the xy-plane are shown at left in Figure 9.1.8. If we lift these contours and plot them at their respective heights, then we get a picture of the surface itself, as illustrated at right in Figure 9.1.8.

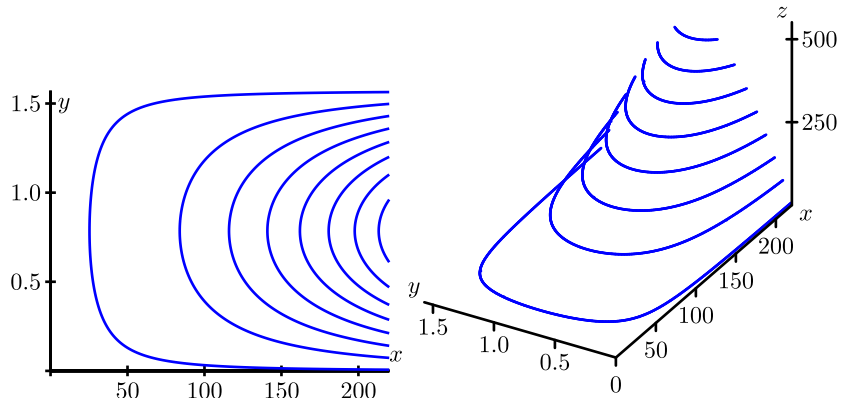


Figure 9.1.8. Left: Level curves. Right: Level curves at appropriate heights.

The use of level curves and traces can help us construct the graph of a function of two variables.

Activity 9.1.8

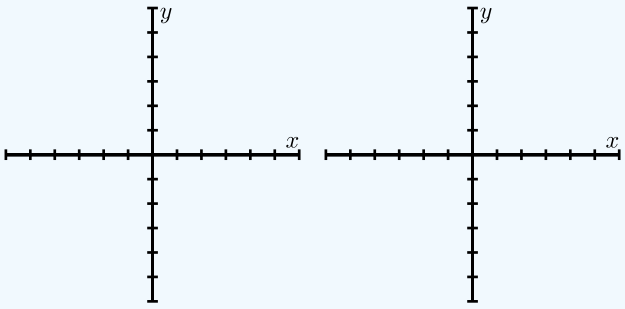


Figure 9.1.9. Left: Level curves for $f(x,y) = x^2 + y^2$. Right: Level curves for $g(x,y) = \sqrt{x^2 + y^2}$.

1. Let $f(x,y) = x^2 + y^2$. Draw the level curves $f(x,y) = k$ for $k = 1$, $k = 2$, $k = 3$, and $k = 4$ on the left set of axes given in Figure 9.1.17. (You decide on the scale of the axes.) Explain what the surface defined by f looks like.
2. Let $g(x,y) = \sqrt{x^2 + y^2}$. Draw the level curves $g(x,y) = k$ for $k = 1$, $k = 2$, $k = 3$, and $k = 4$ on the right set of axes given in Figure 9.1.17. (You decide on the scale of the axes.) Explain what the surface defined by g looks like.
3. Compare and contrast the graphs of f and g. How are they alike? How are they different? Use traces for each function to help answer these questions.

The traces and level curves of a function of two variables are curves in space. In order to understand these traces and level curves better, we will first spend some time learning about vectors and vector-valued functions in the next few sections and return to our study of functions of several variables once we have those more mathematical tools to support their study.

A gallery of functions

We end this section by considering a collection of functions and illustrating their graphs and some level curves.

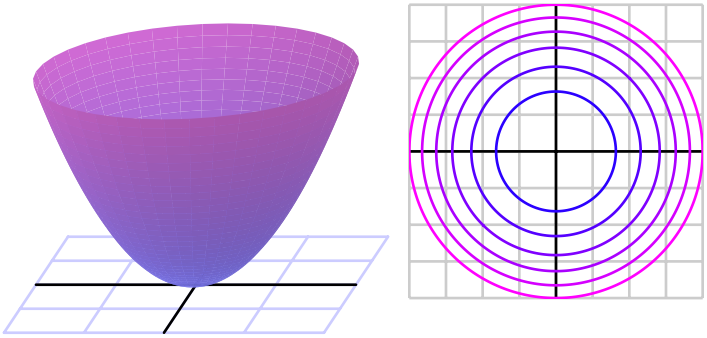


Figure 9.1.10. $z = x^2 + y^2$

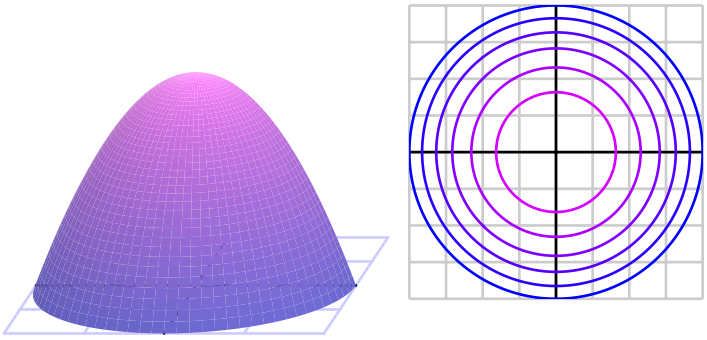


Figure 9.1.11. $z = 4 - (x^2 + y^2)$

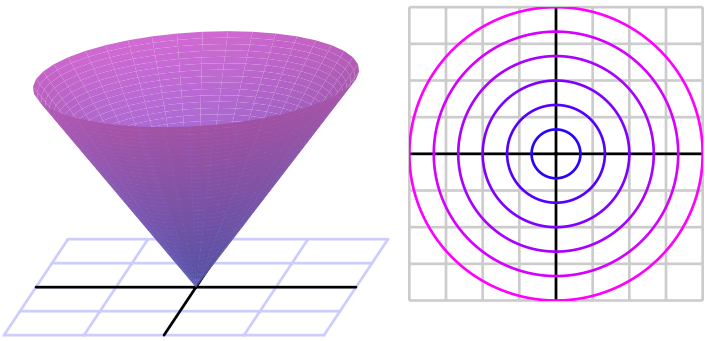


Figure 9.1.12. $z = \sqrt{z^2 + y^2}$

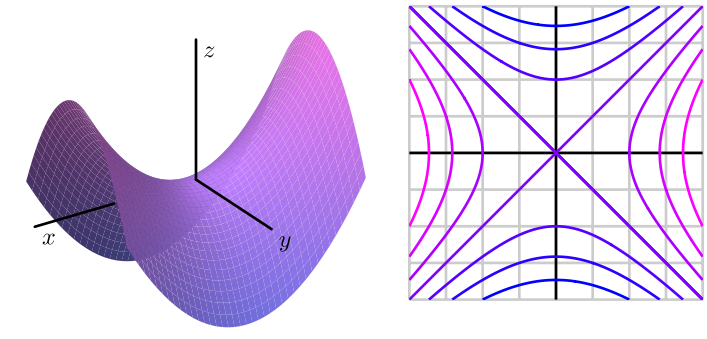


Figure 9.1.13. $z = x^2 - y^2$

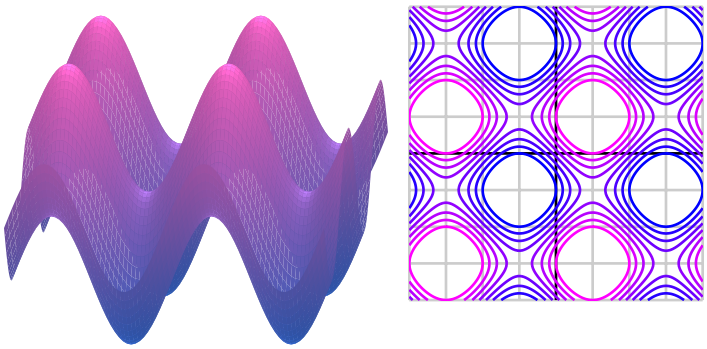


Figure 9.1.14. $z = \sin(x) + \sin(y)$

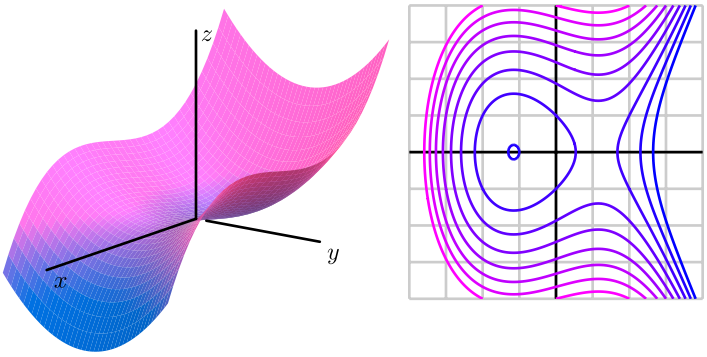


Figure 9.1.15. $z = y^2 - x^3 + x$

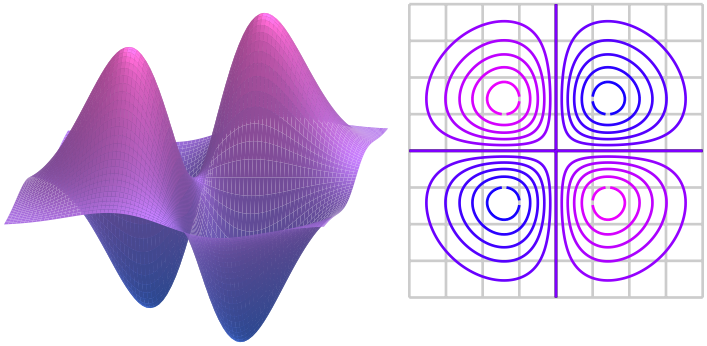


Figure 9.1.16. $z = xye^{-x^2-y^2}$

Summary

- A function f of several variables is a rule that assigns a unique number to an ordered collection of independent inputs. The domain of a function of several variables is the set of all inputs for which the function is defined.
- In $\mathbb{R}^3$, the distance between points $P = (x_0, y_0, z_0)$ and $Q = (x_1, y_1, z_1)$ (denoted as $|PQ|$) is given by the formula

$$|PQ| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}.$$

  and thus the equation of a sphere with center (x_0, y_0, z_0) and radius r is

$$(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = r^2.$$

- A trace of a function f of two independent variables x and y is a curve of the form $z = f(c, y)$ or $z = f(x, c)$, where c is a constant. A trace tells us how the function depends on a single independent variable if we treat the other independent variable as a constant.

- A level curve of a function f of two independent variables x and y is a curve of the form $k = f(x, y)$, where k is a constant. A level curve describes the set of inputs that lead to a specific output of the function.

This page titled 9.1: Functions of Several Variables and Three Dimensional Space is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 9.1: Functions of Several Variables and Three Dimensional Space by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

9.2: Vectors

Motivating Questions

- What is a vector?
- What does it mean for two vectors to be equal?
- How do we add two vectors together and multiply a vector by a scalar?
- How do we determine the magnitude of a vector? What is a unit vector, and how do we find a unit vector in the direction of a given vector?

Quantities like length, speed, area, and mass are all measured by numbers (called *scalars*). Other quantities, like velocity, force, and displacement, have two attributes: magnitude and direction. These quantities are represented by *vectors* and are the study of this section. For example, we will use vectors to calculate work done by a constant force, calculate torque, determine direction vectors for lines and normal vectors for planes, define curvature, and determine the direction of greatest increase on a surface. For most of these applications, we will be interested in using vectors to measure direction and/or speed. Vectors will be a major tool for us in determining the behavior of functions of several variables.

If we are at a point x in the domain of a function of one variable, there are only two directions in which we can move: in the positive or negative x-direction. If, however, we are at a point (x, y) in the domain of a function of two variables, there are many directions in which we can move. Thus, it is important for us to have a means to indicate direction, and we will do so using vectors. This notion of direction in space will be critical for us to find direction vectors for lines, tangent lines to curves, normal vectors to planes, and to determine direction of motion.

Preview Activity 9.2.1

Postscript is a programming language whose primary purpose is to describe the appearance of text or graphics. A simple set of Postscript commands that produces the triangle in the plane with vertices $(0, 0)$, $(1, 1)$, and $(1, -1)$ is the following:

```
(0,0) moveto
(1,1) lineto stroke
(1,-1) lineto stroke
(0,0) lineto stroke
```

The key idea in these commands is that we start at the origin, then tell Postscript that we want to start at the point $(0, 0)$, draw a line from the point $(0, 0)$ to the point $(1, 1)$ (this is what the lineto and stroke commands do), then draw lines from $(1, 1)$ to $(1, -1)$ and $(1, -1)$ back to the origin. Each of these commands encodes two important pieces of information: a direction in which to move and a distance to move. Mathematically, we can capture this information succinctly in a vector. To do so, we record the movement on the map in a pair $\langle x, y\rangle$ (this pair $\langle x, y\rangle$ is a *vector*), where x is the horizontal displacement and y the vertical displacement from one point to another. So, for example, the vector from the origin to the point $(1, 1)$ is represented by the vector $\langle 1, 1\rangle$.

1. What is the vector $\mathbf{v}_1 = \langle x, y\rangle$ that describes the displacement from the point $(1, 1)$ to the point $(1, -1)$? How can we use this vector to determine the distance from the point $(1, 1)$ to the point $(1, -1)$?
2. Suppose we want to draw the triangle with vertices $A = (2, 3)$, $B = (-3, 1)$, and $C = (4, -2)$. As a shorthand notation, we will denote the vector from the point A to the point B as $\overrightarrow{AB}$
   1. Determine the vectors $\overrightarrow{AB}$, $\overrightarrow{BC}$, and $\overrightarrow{AC}$.
   2. What relationship do you see among the vectors $\overrightarrow{AB}$, $\overrightarrow{BC}$, and $\overrightarrow{AC}$? Explain why this relationship should hold.

Representations of Vectors

Preview Activity $9.2.1$ shows how we can record the magnitude and direction of a change in position using an ordered pair of numbers $\langle x, y \rangle$. There are many other quantities, such as force and velocity, that possess the attributes of magnitude and direction, and we will call each such quantity a *vector*.

Definition $9.2.1$

A ***vector*** is a quantity that possesses the attributes of magnitude and direction.

We can represent a vector geometrically as a directed line segment, with the magnitude as the length of the segment and an arrowhead indicating direction, as shown at left in Figure $9.2.1$.

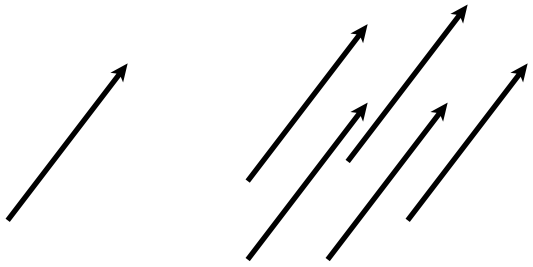


Figure $9.2.1$. Left: A vector. Right: Representations of the same vector.

According to the definition, a vector possesses the attributes of length (magnitude) and direction; the vector's position, however, is not mentioned. Consequently, we regard as equal any two vectors having the same magnitude and direction, as shown at right in Figure $9.2.2$. In other words, two vectors are ***equal*** provided they have the same magnitude and direction.

This means that the same vector may be drawn in the plane in many different ways. For instance, suppose that we would like to draw the vector $\langle 3, 4 \rangle$, which represents a horizontal change of three units and a vertical change of four units. We may place the *tail* of the vector (the point from which the vector originates) at the origin and the *tip* (the terminal point of the vector) at $(3, 4)$, as illustrated at left in Figure $9.2.3$. A vector with its tail at the origin is said to be in *standard position*.

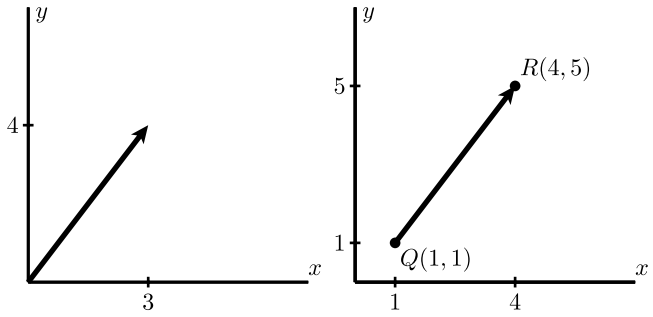


Figure $9.2.3$. Left: Standard position. Right: A vector between two points.

Alternatively, we may place the tail of the vector $\langle 3, 4 \rangle$ at another point, such as $Q(1, 1)$. After a displacement of three units to the right and four units up, the tip of the vector is at the point $R(4, 5)$ (see the vector at right in Figure $9.2.3$).

In this example, the vector led to the directed line segment from Q to R, which we denote as $\overrightarrow{QR}$. We may also turn the situation around: given the two points Q and R, we obtain the vector $\langle 3, 4 \rangle$ because we move horizontally three units and vertically four units to get from Q to R. In other words, $\overrightarrow{QR} = \langle 3, 4 \rangle$. In general, the vector $\overrightarrow{QR}$ from the point $Q = (q_1, q_2)$ to $R = (r_1, r_2)$ is found by taking the difference of coordinates, so that

$$\overrightarrow{QR} = \langle r_1 - q_1, r_2 - q_2 \rangle.$$

We will use boldface letters to represent vectors, such as $v = \langle 3, 4 \rangle$, to distinguish them from scalars. The entries of a vector are called its *components*; in the vector $\langle 3, 4 \rangle$, the x component is 3 and the y component is 4. We use pointed brackets $\langle\ ,\ \rangle$ and the term *components* to distinguish a vector from a point $(\ ,\ )$ and its *coordinates*. There is, however, a close connection between vectors and points. Given a point P, we will frequently consider the vector $\overrightarrow{OP}$ from the origin O to P. For instance, if $P = (3, 4)$, then $\overrightarrow{OP} = \langle 3, 4 \rangle$ as in Figure 9.2.4. In this way, we think of a point P as defining a vector $\overrightarrow{OP}$ whose components agree with the coordinates of P. The vector $\overrightarrow{OP}$ is called the *position vector* of P.

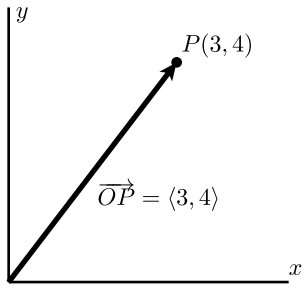


Figure 9.2.4. A point defines a vector

While we often illustrate vectors in the plane since it is easier to draw pictures, different situations call for the use of vectors in three or more dimensions. For instance, a vector $\mathbf{v}$ in n-dimensional space, $\mathbb{R}^n$, has n components and may be represented as

$$\mathbf{v} = \langle v_1, v_2, v_3, \ldots, v_n \rangle.$$

The next activity will help us to become accustomed to vectors and operations on vectors in three dimensions.

Activity 9.2.2

An article by C.Kenneth Tanner of the University of Georgia argues that, due to the concept of social distance, a secondary school classroom for 20 students should have 1344 square feet of floor space. Suppose a classroom is 32 feet by 42 feet by 8 feet. Set the origin O of the classroom to be its center. In this classroom, a student is sitting on a chair whose seat is at location $A = (9, -6, -1.5)$, an overhead projector is located at position $B = (0, 1, 3)$, and the teacher is standing at point $C = (-2, 20, -4)$, all distances measured in feet. Determine the components of the indicated vectors and explain in context what each represents.

a. $\overrightarrow{OA}$	b. $\overrightarrow{OB}$	c. $\overrightarrow{OC}$	d. $\overrightarrow{AB}$	e. $\overrightarrow{AC}$	f. $\overrightarrow{BC}$

Equality of Vectors

Because location is not mentioned in the definition of a vector, any two vectors that have the same magnitude and direction are equal. It is helpful to have an algebraic way to determine when this occurs. That is, if we know the components of two vectors $\mathbf{u}$ and $\mathbf{v}$, we will want to be able to determine algebraically when $\mathbf{u}$ and $\mathbf{v}$ are equal. There is an obvious set of conditions that we use.

Equality of Vectors

Two vectors $\mathbf{u} = \langle u_1, u_2 \rangle$ and $\mathbf{v} = \langle v_1, v_2 \rangle$ in $\mathbb{R}^2$ are equal if and only if their corresponding components are equal: $u_1 = v_1$ and $u_2 = v_2$. More generally, two vectors $\mathbf{u} = \langle u_1, u_2, \ldots, u_n \rangle$ and $\mathbf{v} = \langle v_1, v_2, \ldots, v_n \rangle$ in $\mathbb{R}^n$ are equal if and only if $u_i = v_i$ for each possible value of i.

Operations on Vectors

Vectors are not numbers, but we can now represent them with components that are real numbers. As such, we naturally wonder if it is possible to add two vectors together, multiply two vectors, or combine vectors in any other ways. In this section, we will study two operations on vectors: vector addition and scalar multiplication. To begin, we investigate a natural way to add two vectors together, as well as to multiply a vector by a scalar.

Activity 9.2.3

Let $\mathbf{u} = \langle 2, 3\rangle$, $\mathbf{v} = \langle -1, 4\rangle$.

1. Using the two specific vectors above, what is the natural way to define the vector sum $\mathbf{u} + \mathbf{v}$?
2. In general, how do you think the vector sum $\mathbf{a} + \mathbf{b}$ of vectors $\mathbf{a} = \langle a_1, a_2\rangle$ and $\mathbf{b} = \langle b_1, b_2\rangle$ in $\mathbb{R}^2$ should be defined? Write a formal definition of a vector sum based on your intuition.
3. In general, how do you think the vector sum $\mathbf{a} + \mathbf{b}$ of vectors $\mathbf{a} = \langle a_1, a_2, a_3\rangle$ and $\mathbf{b} = \langle b_1, b_2, b_3\rangle$ in $\mathbb{R}^3$ should be defined? Write a formal definition of a vector sum based on your intuition.
4. Returning to the specific vector $\mathbf{v} = \langle -1, 4\rangle$ given above, what is the natural way to define the scalar multiple $\frac{1}{2}\mathbf{v}$?
5. In general, how do you think a scalar multiple of a vector $\mathbf{a} = \langle a_1, a_2\rangle$ in $\mathbb{R}^2$ by a scalar c should be defined? how about for a scalar multiple of a vector $\mathbf{a} = \langle a_1, a_2, a_3\rangle$ in $\mathbb{R}^3$ by a scalar c? Write a formal definition of a scalar multiple of a vector based on your intuition.

We can now add vectors and multiply vectors by scalars, and thus we can add together scalar multiples of vectors. This allows us to define *vector subtraction*, $v - u$, as the sum of v and -1 times u, so that

$$\mathbf{v} - \mathbf{u} = \mathbf{v} + (-1)\mathbf{u}.$$

Using vector addition and scalar multiplication, we will often represent vectors in terms of the special vectors $\mathbf{i} = \langle 1, 0\rangle$ and $\mathbf{j} = \langle 0, 1\rangle$. For instance, we can write the vector $\langle a, b\rangle$ in $\mathbb{R}^2$ as

$$\langle a, b\rangle = a\langle 1, 0\rangle + b\langle 0, 1\rangle = a\mathbf{i} + b\mathbf{j},$$

which means that

$$\langle 2, -3\rangle = 2\mathbf{i} - 3\mathbf{j}.$$

In the context of $\mathbb{R}^3$, we let $\mathbf{i} = \langle 1, 0, 0\rangle$, $\mathbf{j} = \langle 0, 1, 0\rangle$, and $\mathbf{k} = \langle 0, 0, 1\rangle$, and we can write the vector $\langle a, b, c\rangle$ in $\mathbb{R}^3$ as

$$\langle a, b, c\rangle = a\langle 1, 0, 0\rangle + b\langle 0, 1, 0\rangle + c\langle 0, 0, 1\rangle = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}.$$

The vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are called the *standard unit vectors* (as we will learn momentarily, unit vectors have length 1), and are important in the physical sciences.

Properties of Vector Operations

We know that the scalar sum $1 + 2$ is equal to the scalar sum $2 + 1$. This is called the *commutative* property of scalar addition. Any time we define operations on objects (like addition of vectors) we usually want to know what kinds of properties the operations have. For example, is addition of vectors a commutative operation? To answer this question we take two *arbitrary* vectors $\mathbf{v}$ and $\mathbf{u}$ and add them together and see what happens. Let $\mathbf{v} = \langle v_1, v_2\rangle$ and $\mathbf{u} = \langle u_1, u_2\rangle$. Now we use the fact that v_1, v_2, u_1, and u_2 are scalars, and that the addition of scalars is commutative to see that

$$\begin{aligned}\mathbf{v} + \mathbf{u} &= \langle v_1, v_2\rangle + \langle u_1, u_2\rangle \\ &= \langle v_1 + u_1, v_2 + u_2\rangle \\ &= \langle u_1 + v_1, u_2 + v_2\rangle \\ &= \langle u_1, u_2\rangle + \langle v_1, v_2\rangle \\ &= \mathbf{u} + \mathbf{v}.\end{aligned}$$

So the vector sum is a commutative operation. Similar arguments can be used to show the following properties of vector addition and scalar multiplication.

Properties of vector operations

Let $\mathbf{v}$, $\mathbf{u}$, and $\mathbf{w}$ be vectors in $\mathbb{R}^n$ and let a and b be scalars. Then

1. $\mathbf{v}+\mathbf{u}=\mathbf{u}+\mathbf{v}$
2. $(\mathbf{v}+\mathbf{u})+\mathbf{w}=\mathbf{v}+(\mathbf{u}+\mathbf{w})$
3. The vector $\mathbf{0}=\langle 0,0,\ldots,0\rangle$ has the property that $\mathbf{v}+\mathbf{0}=\mathbf{v}$. The vector $\mathbf{0}$ is called the *zero vector*.
4. $(-1)\mathbf{v}+\mathbf{v}=\mathbf{0}$. The vector $(-1)\mathbf{v}=-\mathbf{v}$ is called the *additive inverse* of the vector $\mathbf{v}$.
5. $(a+b)\mathbf{v}=a\mathbf{v}+b\mathbf{v}$
6. $a(\mathbf{v}+\mathbf{u})=a\mathbf{v}+a\mathbf{u}$
7. $(ab)\mathbf{v}=a(b\mathbf{v})$
8. $1\mathbf{v}=\mathbf{v}$.

We verified the first property for vectors in $\mathbb{R}^2$; it is straightforward to verify that the rest of the eight properties just noted hold for all vectors in $\mathbb{R}^n$.

Geometric Interpretation of Vector Operations

Next, we explore a geometric interpretation of vector addition and scalar multiplication that allows us to visualize these operations. Let $\mathbf{u}=\langle 4,6\rangle$ and $\mathbf{v}=\langle 3,-2\rangle$. Then $\mathbf{w}=\mathbf{u}+\mathbf{v}=\langle 7,4\rangle$, as shown on the left in Figure $9.2.5$.

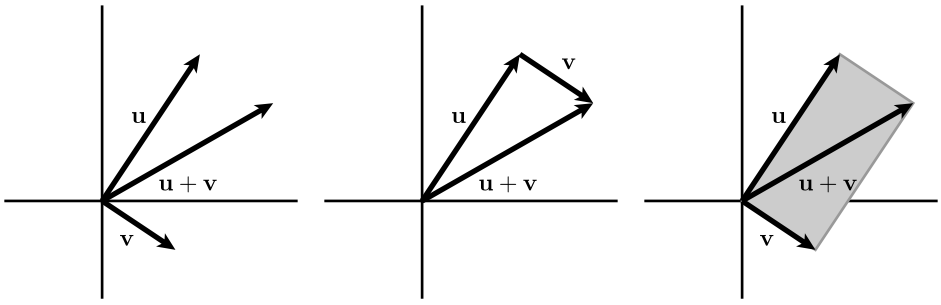


Figure $9.2.5$. A vector sum (left), summing displacements (center), the parallelogram law (right).

If we think of these vectors as displacements in the plane, we find a geometric way to envision vector addition. For instance, the vector $\mathbf{u}+\mathbf{v}$ will represent the displacement obtained by following the displacement $\mathbf{u}$ with the displacement $\mathbf{v}$. We may picture this by placing the tail of $\mathbf{v}$ at the tip of $\mathbf{u}$, as seen in the center of Figure $9.2.5$.

Of course, vector addition is commutative so we obtain the same sum if we place the tail of $\mathbf{u}$ at the tip of $\mathbf{v}$. We therefore see that $\mathbf{u}+\mathbf{v}$ appears as the diagonal of the parallelogram determined by $\mathbf{u}$ and $\mathbf{v}$, as shown at right in Figure $9.2.5$.

Vector subtraction has a similar interpretation. At left in Figure $9.2.6$ we see vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}=\mathbf{u}+\mathbf{v}$. If we rewrite $\mathbf{v}=\mathbf{w}-\mathbf{u}$, we have the arrangement shown at right in Figure $9.2.6$. In other words, to form the difference $\mathbf{w}-\mathbf{u}$, we draw a vector from the tip of $\mathbf{u}$ to the tip of $\mathbf{w}$.

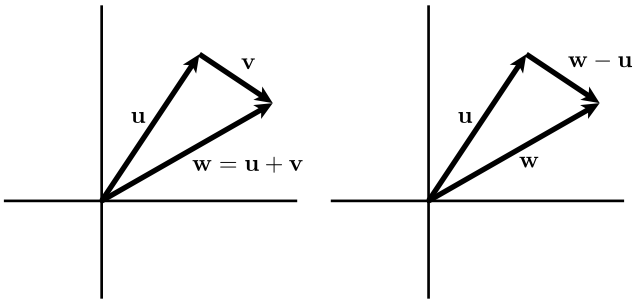


Figure 9.2.6. Left: Vector addition. Right: Vector subtraction.

In a similar way, we may geometrically represent a scalar multiple of a vector. For instance, if $\mathbf{v} = \langle 2, 3 \rangle$, then $2\mathbf{v} = \langle 4, 6 \rangle$. As shown in Figure 9.2.7, multiplying $\mathbf{v}$ by 2 leaves the direction unchanged, but stretches $\mathbf{v}$ by 2. Also, $-2\mathbf{v} = \langle -4, -6 \rangle$, which shows that multiplying by a negative scalar gives a vector pointing in the opposite direction of $\mathbf{v}$.

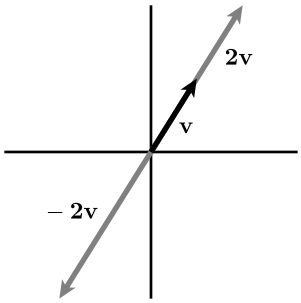


Figure 9.2.7. Scalar multiplication

Activity 9.2.4

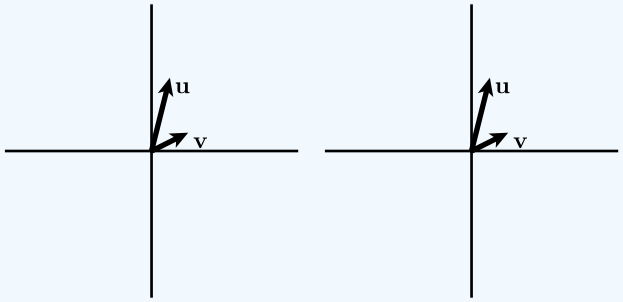


Figure 9.2.8. Left: Sketch sums. Right: Sketch multiples.

Suppose that $\mathbf{u}$ and $\mathbf{v}$ are the vectors shown in Figure 9.2.8.

1. On the axes at left in Figure 9.2.8, sketch the vectors $\mathbf{u} + \mathbf{v}$, $\mathbf{v} - \mathbf{u}$, $2\mathbf{u}$, $-2\mathbf{u}$, and $-3\mathbf{v}$.
2. What is $0\mathbf{v}$?
3. On the axes at right in Figure 9.2.8, sketch the vectors $-3\mathbf{v}$, $-2\mathbf{v}$, $-1\mathbf{v}$, $2\mathbf{v}$, and $3\mathbf{v}$.
4. Give a geometric description of the set of terminal points of the vectors $t\mathbf{v}$ where t is any scalar.
5. On the set of axes at right in Figure 9.2.8, sketch the vectors $\mathbf{u} - 3\mathbf{v}$, $\mathbf{u} - 2\mathbf{v}$, $\mathbf{u} - \mathbf{v}$, $\mathbf{u} + \mathbf{v}$, and $\mathbf{u} + 2\mathbf{v}$.
6. Give a geometric description of the set of terminal points of the vectors $\mathbf{u} + t\mathbf{v}$ where t is any scalar.

The Magnitude of a Vector

By definition, vectors have both direction and magnitude (or length). We now investigate how to calculate the magnitude of a vector. Since a vector $\mathbf{v}$ can be represented by a directed line segment, we can use the distance formula to calculate the length of the segment. This length is the *magnitude* of the vector $\mathbf{v}$ and is denoted $|\mathbf{v}|$.

Activity 9.2.5

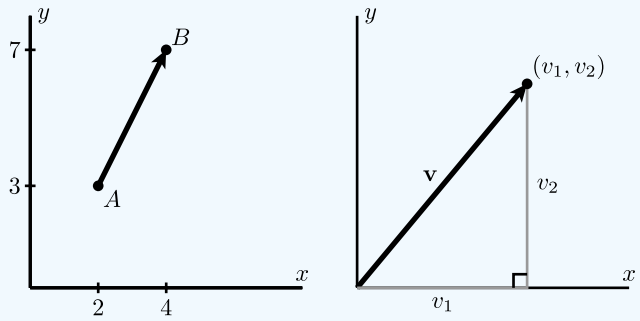


Figure 9.2.9. Left: $\overrightarrow{AB}$. Right: An arbitrary vector, $\mathbf{v}$.

1. Let $A = (2, 3)$ and $B = (4, 7)$, as shown at left in Figure 9.2.9. Compute $|\overrightarrow{AB}|$.
2. Let $\mathbf{v} = \langle v_1, v_2 \rangle$ be the vector in $\mathbb{R}^2$ with components v_1 and v_2 as shown at right in Figure 9.2.9. Use the distance formula to find a general formula for $|\mathbf{v}|$.
3. Let $\mathbf{v} = \langle v_1, v_2, v_3 \rangle$ be a vector in $\mathbb{R}^3$. Use the distance formula to find a general formula for $|\mathbf{v}|$.
4. Suppose that $\mathbf{u} = \langle 2, 3 \rangle$ and $\mathbf{v} = \langle -1, 2 \rangle$. Find $|\mathbf{u}|$, $|\mathbf{v}|$, and $|\mathbf{u} + \mathbf{v}|$. Is it true that $|\mathbf{u} + \mathbf{v}| = |\mathbf{u}| + |\mathbf{v}|$?
5. Under what conditions will $|\mathbf{u} + \mathbf{v}| = |\mathbf{u}| + |\mathbf{v}|$? (Hint: Think about how $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{u} + \mathbf{v}$ form the sides of a triangle.)
6. With the vector $\mathbf{u} = \langle 2, 3 \rangle$, find the lengths of $2\mathbf{u}$, $3\mathbf{u}$, and $-2\mathbf{u}$, respectively, and use proper notation to label your results.
7. If t is any scalar, how is $|t\mathbf{u}|$ related to $|\mathbf{u}|$?
8. A *unit vector* is a vector whose magnitude is 1. Of the vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{i} + \mathbf{j}$, which are unit vectors?
9. Find a unit vector $\mathbf{v}$ whose direction is the same as $\mathbf{u} = \langle 2, 3 \rangle$. (Hint: Consider the result of part (g).)

Summary

- A vector is an object that possesses the attributes of magnitude and direction. Examples of vector quantities are position, velocity, acceleration, and force.
- Two vectors are equal if they have the same direction and magnitude. Notice that position is not considered, so a vector is independent of its location.
- If $\mathbf{u} = \langle u_1, u_2, \ldots, u_n \rangle$ and $\mathbf{v} = \langle v_1, v_2, \ldots, v_n \rangle$ are two vectors in $\mathbb{R}^n$, then their vector sum is the vector
$$\mathbf{u} + \mathbf{v} = \langle u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n \rangle.$$
- If $\mathbf{u} = \langle u_1, u_2, \ldots, u_n \rangle$ is a vector in $\mathbb{R}^n$ and c is a scalar, then the scalar multiple $c\mathbf{u}$ is the vector
$$c\mathbf{u} = \langle cu_1, cu_2, \ldots, cu_n \rangle.$$
- The magnitude of the vector $\mathbf{v} = \langle v_1, v_2, \ldots, v_n \rangle$ in $\mathbb{R}^n$ is the scalar
$$|\mathbf{v}| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}.$$
- A vector $\mathbf{u}$ is a unit vector provided that $|\mathbf{u}| = 1$. If $\mathbf{v}$ is a nonzero vector, then the vector $\frac{v}{|v|}$ is a unit vector with the same direction as $\mathbf{v}$.

This page titled 9.2: Vectors is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 9.2: Vectors by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

9.3: Dot Product

Motivating Questions

- How is the dot product of two vectors defined and what geometric information does it tell us?
- How can we tell if two vectors in $\mathbb{R}^n$ are perpendicular?
- How do we find the projection of one vector onto another?

In the last section, we considered vector addition and scalar multiplication and found that each operation had a natural geometric interpretation. In this section, we will introduce a means of multiplying vectors.

Preview Activity 9.3.1

For two-dimensional vectors $\mathbf{u} = \langle u_1, u_2 \rangle$ and $\mathbf{v} = \langle v_1, v_2 \rangle$, the dot product is simply the scalar obtained by

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2.$$

a. If $\mathbf{u} = \langle 3, 4 \rangle$ and $\mathbf{v} = \langle -2, 1 \rangle$, find the dot product $\mathbf{u} \cdot \mathbf{v}$.

b. Find $\mathbf{i} \cdot \mathbf{j}$ and $\mathbf{i} \cdot \mathbf{j}$.

c. If $\mathbf{u} = \langle 3, 4 \rangle$, find $\mathbf{u} \cdot \mathbf{u}$. How is this related to $|\mathbf{u}|$?

d. On the axes in Figure 9.3.1, plot the vectors $\mathbf{u} = \langle 1, 3 \rangle$ and $\mathbf{v} = \langle -3, 1 \rangle$. Then, find $\mathbf{u} \cdot \mathbf{v}$. What is the angle between these vectors?

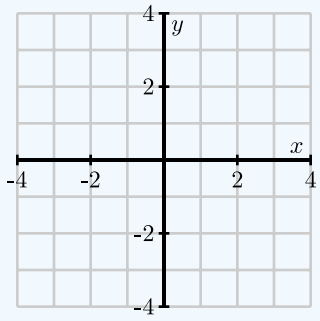


Figure 9.3.1. For part (d)

e. On the axes in Figure 9.3.2, plot the vector $\mathbf{u} = \langle 1, 3 \rangle$.

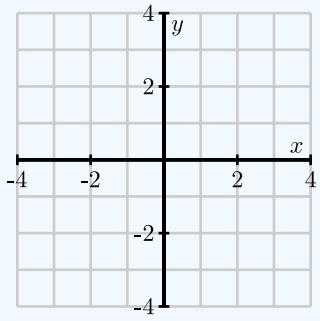


Figure 9.3.2. For part (e)

For each of the following vectors $\mathbf{v}$, plot the vector on Figure 9.3.2 and then compute the dot product $\mathbf{u} \cdot \mathbf{v}$.

- $\mathbf{v} = \langle 3, 2 \rangle$.
- $\mathbf{v} = \langle 3, 0 \rangle$.
- $\mathbf{v} = \langle 3, -1 \rangle$.
- $\mathbf{v} = \langle 3, -2 \rangle$.
- $\mathbf{v} = \langle 3, -4 \rangle$.

f. Based upon the previous part of this activity, what do you think is the sign of the dot product in the following three cases shown in Figure 9.3.3?

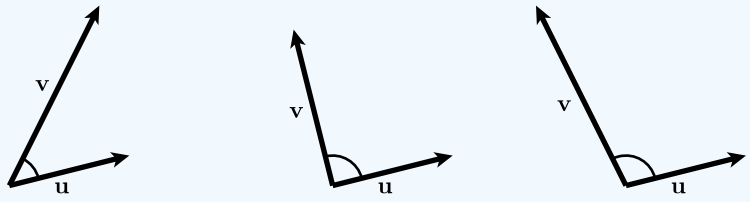


Figure 9.3.3. For part (f)

The Dot Product

The definition of the dot product for vectors in $\mathbb{R}^2$ given in Preview Activity 9.3.1 can be extended to vectors in $\mathbb{R}^n$.

Definition 9.3.4

The *dot product* of vectors $\mathbf{u} = \langle u_1, u_2, \ldots, u_n \rangle$ and $\mathbf{v} = \langle v_1, v_2, \ldots, v_n \rangle$ in $\mathbb{R}^n$ is the scalar

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n.$$

(As we will see shortly, the dot product arises in physics to calculate the work done by a vector force in a given direction. It might be more natural to define the dot product in this context, but it is more convenient from a mathematical perspective to define the dot product algebraically and then view work as an application of this definition.)

For instance, we find that

$$\langle 3, 0, 1 \rangle \cdot \langle -2, 1, 4 \rangle = 3 \cdot (-2) + 0 \cdot 1 + 1 \cdot 4 = -6 + 0 + 4 = -2.$$

Notice that the resulting quantity is a scalar. Our work in Preview Activity 9.3.1 examined dot products of two-dimensional vectors.

Activity 9.3.2

Determine each of the following.

1. $\langle 1, 2, -3 \rangle \cdot \langle 4, -2, 0 \rangle$.
2. $\langle 0, 3, -2, 1 \rangle \cdot \langle 5, -6, 0, 4 \rangle$

The dot product is a natural way to define a product of two vectors. In addition, it behaves in ways that are similar to the product of, say, real numbers.

Properties of the dot product

Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ be vectors in $\mathbb{R}^n$. Then

1. $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$ (the dot product is *commutative*), and
2. $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = (\mathbf{u} \cdot \mathbf{w}) + (\mathbf{v} \cdot \mathbf{w})$.
3. if c is a scalar, then $(c\mathbf{u}) \cdot \mathbf{w} = c(\mathbf{u} \cdot \mathbf{w})$.

Moreover, the dot product gives us valuable geometric information about the vectors and their relative orientation. For instance, let's consider what happens when we dot a vector with itself:

$$\mathbf{u} \cdot \mathbf{u} = \langle u_1, u_2, \ldots, u_n \rangle \cdot \langle u_1, u_2, \ldots, u_n \rangle = u_1^2 + u_2^2 + \ldots + u_n^2 = |\mathbf{u}|^2.$$

In other words, the dot product of a vector with itself gives the square of the length of the vector: $\mathbf{u} \cdot \mathbf{u} = |\mathbf{u}|^2$.

The angle between vectors

The dot product can help us understand the angle between two vectors. For instance, if we are given two vectors $\mathbf{u}$ and $\mathbf{v}$, there are two angles that these vectors create, as depicted at left in Figure 9.3.4. We will call θ, the smaller of these angles, the *angle between these vectors*. Notice that θ lies between 0 and π.

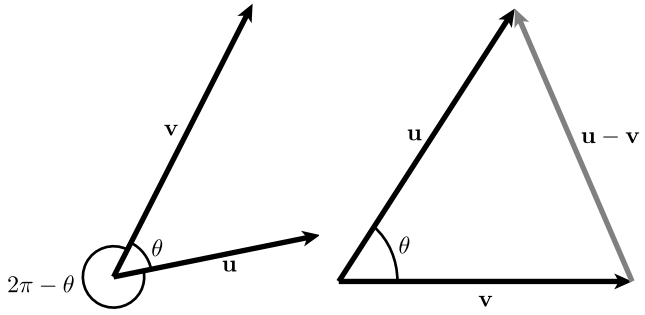


Figure 9.3.4. Left: The angle between $\mathbf{u}$ and $\mathbf{v}$. Right: The triangle formed by $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{u} - \mathbf{v}$.

To determine this angle, we may apply the Law of Cosines to the triangle shown at right in Figure 9.3.4.

Using the fact that the dot product of a vector with itself gives us the square of its length, together with the properties of the dot product, we find:

$$\begin{aligned}
|\mathbf{u} - \mathbf{v}|^2 &= |\mathbf{u}|^2 + |\mathbf{v}|^2 - 2|\mathbf{u}||\mathbf{v}|\cos(\theta) \\
(\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v}) &= \mathbf{u} \cdot \mathbf{u} + \mathbf{v} \cdot \mathbf{v} - 2|\mathbf{u}||\mathbf{v}|\cos(\theta) \\
\mathbf{u} \cdot (\mathbf{u} - \mathbf{v}) - \mathbf{v} \cdot (\mathbf{u} - \mathbf{v}) &= \mathbf{u} \cdot \mathbf{u} + \mathbf{v} \cdot \mathbf{v} - 2|\mathbf{u}||\mathbf{v}|\cos(\theta) \\
\mathbf{u} \cdot \mathbf{u} - 2\mathbf{u} \cdot \mathbf{v} + \mathbf{v} \cdot \mathbf{v} &= \mathbf{u} \cdot \mathbf{u} + \mathbf{v} \cdot \mathbf{v} - 2|\mathbf{u}||\mathbf{v}|\cos(\theta) \\
-2\mathbf{u} \cdot \mathbf{v} &= -2|\mathbf{u}||\mathbf{v}|\cos(\theta) \\
\mathbf{u} \cdot \mathbf{v} &= |\mathbf{u}||\mathbf{v}|\cos(\theta).
\end{aligned}$$

To summarize, we have the important relationship

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n = |\mathbf{u}||\mathbf{v}|\cos(\theta). \tag{9.3.1}$$

It is sometimes useful to think of Equation (9.3.1) as giving us an expression for the angle between two vectors:

$$\theta = \cos^{-1}\left(\frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|}\right).$$

The real beauty of this expression is this: the dot product is a very simple algebraic operation to perform yet it provides us with important geometric information — namely the angle between the vectors — that would be difficult to determine otherwise.

Activity 9.3.3

Determine each of the following.

1. The length of the vector $\mathbf{u} = \langle 1, 2, -3 \rangle$ using the dot product.
2. The angle between the vectors $\mathbf{u} = \langle 1, 2 \rangle$ and $\mathbf{v} = \langle 4, -1 \rangle$ to the nearest tenth of a degree.
3. The angle between the vectors $\mathbf{y} = \langle 1, 2, -3 \rangle$ and $\mathbf{z} = \langle -2, 1, 1 \rangle$ to the nearest tenth of a degree.
4. If the angle between the vectors $\mathbf{u}$ and $\mathbf{v}$ is a right angle, what does the expression $\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos(\theta)$ say about their dot product?
5. If the angle between the vectors $\mathbf{u}$ and $\mathbf{v}$ is acute—that is, less than $\pi/2$—what does the expression $\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos(\theta)$ say about their dot product?
6. If the angle between the vectors $\mathbf{u}$ and $\mathbf{v}$ is obtuse—that is, greater than $\pi/2$—what does the expression $\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos(\theta)$ say about their dot product?

The Dot Product and Orthogonality

When the angle between two vectors is a right angle, it is frequently the case that something important is happening. In this case, we say the vectors are *orthogonal*. For instance, orthogonality often plays a role in optimization problems; to determine the shortest path from a point in $\mathbb{R}^3$ to a given plane, we move along a line orthogonal to the plane.

As Activity 9.3.3 indicates, the dot product provides a simple means to determine whether two vectors are orthogonal to one another. In this case, $\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos(\pi/2) = 0$, so we make the following important observation.

The dot product and orthogonality

Two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$ are orthogonal to each other if $\mathbf{u} \cdot \mathbf{v} = 0$.

More generally, the sign of the dot product gives us useful information about the relative orientation of the vectors. If we remember that

$$\begin{aligned} \cos(\theta) > 0 \quad & \text{if } \theta \text{ is an acute angle,} \\ \cos(\theta) = 0 \quad & \text{if } \theta \text{ is a right angle, and} \\ \cos(\theta) < 0 \quad & \text{if } \theta \text{ is an obtuse angle,} \end{aligned}$$

we see that for nonzero vectors $\mathbf{u}$ and $\mathbf{v}$,

$$\begin{aligned} \mathbf{u} \cdot \mathbf{v} > 0 \quad & \text{if } \theta \text{ is an acute angle,} \\ \mathbf{u} \cdot \mathbf{v} = 0 \quad & \text{if } \theta \text{ is a right angle, and} \\ \mathbf{u} \cdot \mathbf{v} < 0 \quad & \text{if } \theta \text{ is an obtuse angle.} \end{aligned}$$

This is illustrated in Figure 9.3.5.

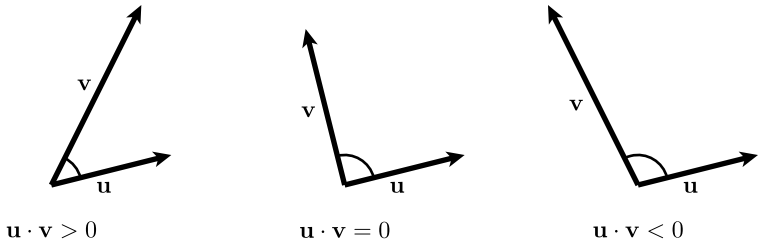


Figure 9.3.5. The orientation of vectors

Work, Force, and Displacement

In physics, work is a measure of the energy required to apply a force to an object through a displacement. For instance, Figure 9.3.6 shows a force $\mathbf{F}$ displacing an object from point A to point B. The displacement is then represented by the vector $\overrightarrow{AB}$.

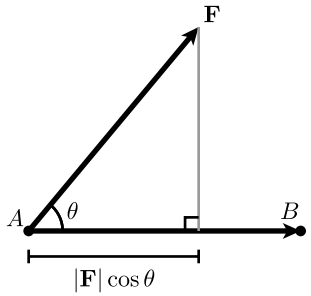


Figure 9.3.6. A force $\mathbf{F}$ displacing an object.

It turns out that the work required to displace the object is

$$W = \mathbf{F} \cdot \overrightarrow{AB} = |\mathbf{F}||\overrightarrow{AB}| \cos(\theta).$$

This means that the work is determined only by the magnitude of the force applied parallel to the displacement. Consequently, if we are given two vectors $\mathbf{u}$ and $\mathbf{v}$, we would like to write $\mathbf{u}$ as a sum of two vectors, one of which is parallel to $\mathbf{v}$ and one of which is orthogonal to $\mathbf{v}$. We take up this task after the next activity.

Activity 9.3.4

Determine the work done by a 25 pound force acting at a 30° angle to the direction of the object's motion, if the object is pulled 10 feet. In addition, is more work or less work done if the angle to the direction of the object's motion is 60°?

Projections

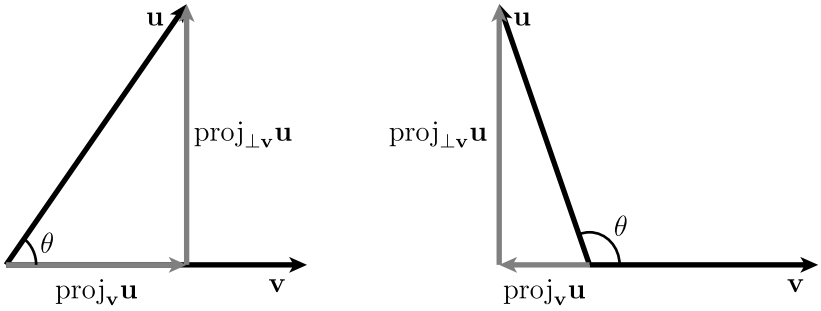


Figure 9.3.7. Left: $proj_{\mathbf{v}}\mathbf{u}$. Right: $proj_{\mathbf{v}}\mathbf{u}$ when $\theta > \frac{\pi}{2}$.

Suppose we are given two vectors $\mathbf{u}$ and $\mathbf{v}$ as shown at left in Figure 9.3.7. Motivated by our discussion of work, we would like to write $\mathbf{u}$ as a sum of two vectors, one of which is parallel to $\mathbf{v}$ and one of which is orthogonal. That is, we would like to write

$$\mathbf{u} = proj_{\mathbf{v}}\mathbf{u} + proj_{\perp\mathbf{v}}\mathbf{u}, \tag{9.3.2}$$

where $proj_{\mathbf{v}}\mathbf{u}$ is parallel to $\mathbf{v}$ and $proj_{\perp\mathbf{v}}\mathbf{u}$ is orthogonal to $\mathbf{v}$. We call the vector $proj_{\mathbf{v}}\mathbf{u}$ the *projection* of $\mathbf{u}$ onto $\mathbf{v}$. Note that, as the diagram at right in Figure 9.3.8 illustrates, it is also possible to create a projection even if the angle between the vectors $\mathbf{u}$ and $\mathbf{v}$ exceeds $\frac{\pi}{2}$.

To find the vector $proj_{\mathbf{v}}\mathbf{u}$, we will dot both sides of Equation (9.3.2) with the vector $\mathbf{v}$, to find that

$$\begin{aligned}\mathbf{u}\cdot\mathbf{v} &= (proj_{\mathbf{v}}\mathbf{u} + proj_{\perp\mathbf{v}}\mathbf{u})\cdot\mathbf{v}\\ &= (proj_{\mathbf{v}}\mathbf{u})\cdot\mathbf{v} + (proj_{\perp\mathbf{v}}\mathbf{u})\cdot\mathbf{v}\\ &= (proj_{\mathbf{v}}\mathbf{u})\cdot\mathbf{v}.\end{aligned}$$

Notice that $(proj_{\perp\mathbf{v}}\mathbf{u})\cdot\mathbf{v} = 0$ since $proj_{\perp\mathbf{v}}\mathbf{u}$ is orthogonal to $\mathbf{v}$. Also, $proj_{\mathbf{v}}\mathbf{u}$ must be a scalar multiple of $\mathbf{v}$ since it is parallel to $\mathbf{v}$, so we will write $proj_{\mathbf{v}}\mathbf{u} = s\mathbf{v}$. It follows that

$$\mathbf{u}\cdot\mathbf{v} = (proj_{\mathbf{v}}\mathbf{u})\cdot\mathbf{v} = s\mathbf{v}\cdot\mathbf{v},$$

which means that

$$s = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}}$$

and hence

$$proj_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}}\mathbf{v} = \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v}$$

It is sometimes useful to write $proj_{\mathbf{v}}\mathbf{u}$ as a scalar times a unit vector in the direction of $\mathbf{v}$. We call this scalar the *component of* $\mathbf{u}$ *along* $\mathbf{v}$ and denote it as $comp_{\mathbf{v}}\mathbf{u}$. We therefore have

$$proj_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v} = \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{v}|}\frac{\mathbf{v}}{|\mathbf{v}|} = comp_{\mathbf{v}}\mathbf{u}\frac{\mathbf{v}}{|\mathbf{v}|},$$

so that

$$comp_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{v}|}.$$

The dot product and projections

Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in $\mathbb{R}^n$. The component of $\mathbf{u}$ in the direction of $\mathbf{v}$ is the scalar

$$comp_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{v}|},$$

and the projection of $\mathbf{u}$ onto $\mathbf{v}$ is the vector

$$proj_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}}\mathbf{v}.$$

Moreover, since

$$\mathbf{u} = proj_{\mathbf{v}}\mathbf{u} + proj_{\perp\mathbf{v}}\mathbf{u},$$

it follows that

$$proj_{\perp\mathbf{v}}\mathbf{u} = \mathbf{u} - proj_{\mathbf{v}}\mathbf{u}.$$

This shows that once we have computed $proj_{\mathbf{v}}\mathbf{u}$, we can find $proj_{\perp\mathbf{v}}\mathbf{u}$ simply by calculating the difference of two known vectors.

Activity 9.3.5

Let $\mathbf{u} = \langle 2, 6\rangle$.

1. Let $\mathbf{v} = \langle 4, -8\rangle$. Find $comp_{\mathbf{v}}\mathbf{u}$, $proj_{\mathbf{v}}\mathbf{u}$ and $proj_{\perp\mathbf{v}}\mathbf{u}$, and draw a picture to illustrate. Finally, express $\mathbf{u}$ as the sum of two vectors where one is parallel to $\mathbf{v}$ and the other is perpendicular to $\mathbf{v}$.
2. Now let $\mathbf{v} = \langle -2, 4\rangle$. Without doing any calculations, find $proj_{\mathbf{v}}\mathbf{u}$. Explain your reasoning. (Hint: Refer to the picture you drew in part (a).)
3. Find a vector $\mathbf{w}$ not parallel to $\mathbf{z} = \langle 3, 4\rangle$ such that $proj_{\mathbf{z}}\mathbf{w}$ has length 10. Note that there are infinitely many different answers.

Summary

- The dot product of two vectors in $\mathbb{R}^n$, $\mathbf{u} = \langle u_1, u_2, \ldots, u_n\rangle$ and $\mathbf{v} = \langle v_1, v_2, \ldots, v_n\rangle$, is the scalar
$$\mathbf{u}\cdot\mathbf{v} = u_1v_1 + u_2v_2 + \cdots + u_nv_n.$$
- The dot product is related to the length of a vector since $\mathbf{u}\cdot\mathbf{u} = |\mathbf{u}|^2$.
- The dot product provides us with information about the angle between the vectors since
$$\mathbf{u}\cdot\mathbf{v} = |\mathbf{u}|\ |\mathbf{v}|\cos(\theta),$$
where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$.
- Two vectors are orthogonal if the angle between them is $\pi/2$. In terms of the dot product, the vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if and only if $\mathbf{u}\cdot\mathbf{v} = 0$.
- The projection of a vector $\mathbf{u}$ in $\mathbb{R}^n$ onto a vector $\mathbf{v}$ in $\mathbb{R}^n$ is the vector
$$proj_{\mathbf{v}}\mathbf{u} = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}}\mathbf{v}.$$

This page titled 9.3: Dot Product is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 9.3: Dot Product by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

9.4: The Cross Product

Motivating Questions

- How and when is the cross product of two vectors defined?
- What geometric information does the cross product provide?

The last two sections have introduced some basic algebraic operations on vectors—addition, scalar multiplication, and the dot product—with useful geometric interpretations. In this section, we will meet a final algebraic operation, the *cross product*, which again conveys important geometric information.

To begin, we must emphasize that the cross product is only defined for vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^3$. Also, remember that we use a right-hand coordinate system, as described in Section 9.1. In particular, recall that the vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are oriented as shown below in Figure 9.4.1. Earlier, we noticed that if we point the index finger of our *right* hand in the direction of $\mathbf{i}$ and our middle finger in the direction of $\mathbf{j}$, then our thumb points in the direction of $\mathbf{k}$.

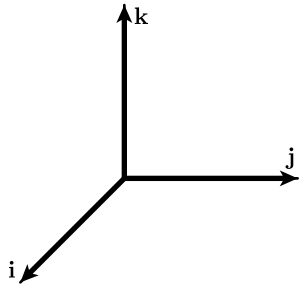


Figure 9.4.1. Basis vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$.

Preview Activity 9.4.1

The cross product of two vectors, $\mathbf{u}$ and $\mathbf{v}$, will itself be a *vector* denoted $\mathbf{u} \times \mathbf{v}$. The direction of $\mathbf{u} \times \mathbf{v}$ is determined by the right-hand rule: if we point the index finger of our right hand in the direction of $\mathbf{u}$ and our middle finger in the direction of $\mathbf{v}$, then our thumb points in the direction of $\mathbf{u} \times \mathbf{v}$.

1. We begin by defining the cross products using the vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$. Referring to Figure 9.4.1, explain why $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ in that order form a right-hand system. We then define $\mathbf{i} \times \mathbf{j}$ to be $\mathbf{k}$ — that is $\mathbf{i} \times \mathbf{j} = \mathbf{k}$.
2. Now explain why $\mathbf{i}$, $\mathbf{k}$, and $-\mathbf{j}$ in that order form a right-hand system. We then define $\mathbf{i} \times \mathbf{k}$ to be $-\mathbf{j}$ — that is $\mathbf{i} \times \mathbf{k} = -\mathbf{j}$.
3. Continuing in this way, complete the missing entries in Table 9.4.2.

   Table 9.4.1. Table of cross products involving $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$.

$\mathbf{i} \times \mathbf{j} = \mathbf{k}$			$\mathbf{i} \times \mathbf{k} = -\mathbf{j}$			$\mathbf{j} \times \mathbf{k} =$
$\mathbf{j} \times \mathbf{i} =$			$\mathbf{k} \times \mathbf{i} =$			$\mathbf{k} \times \mathbf{j} =$

4. Up to this point, the products you have seen, such as the product of real numbers and the dot product of vectors, have been commutative, meaning that the product does not depend on the order of the terms. For instance, $2 \cdot 5 = 5 \cdot 2$. The table above suggests, however, that the cross product is *anti-commutative*: for any vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^3$, $\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$. If we consider the case when $\mathbf{u} = \mathbf{v}$, this shows that $\mathbf{v} \times \mathbf{v} = -(\mathbf{v} \times \mathbf{v})$. What does this tell us about $\mathbf{v} \times \mathbf{v}$; in particular, what vector is unchanged by scalar multiplication by -1?

5. It is not difficult to show that the cross product interacts with scalar multiplication and vector addition as one would expect: that is

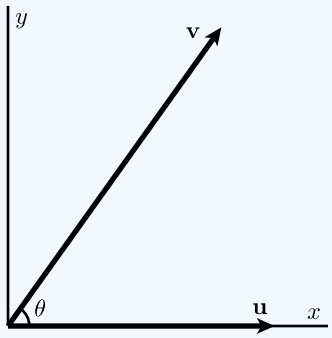
$$(c\mathbf{u}) \times \mathbf{v} = c(\mathbf{u} \times \mathbf{v})$$
$$(\mathbf{u} + \mathbf{v}) \times \mathbf{w} = (\mathbf{u} \times \mathbf{w}) + (\mathbf{v} \times \mathbf{w})$$

   We can combine these properties to make cross product calculations a bit easier. For example,

$$\begin{aligned}(2\mathbf{i} + \mathbf{j}) \times \mathbf{k} &= (2\mathbf{i} \times \mathbf{k}) + (\mathbf{j} \times \mathbf{k}) \\ &= 2(\mathbf{i} \times \mathbf{k}) + (\mathbf{j} \times \mathbf{k}) \\ &= -2\mathbf{j} + \mathbf{i}.\end{aligned}$$

   Using these properties along with Table 9.4.2, find the cross product $\mathbf{u} \times \mathbf{v}$ if $\mathbf{u} = 2\mathbf{i} + 3\mathbf{j}$ and $\mathbf{v} = -\mathbf{i} + \mathbf{k}$.

6. Verify that the cross product $\mathbf{u} \times \mathbf{v}$ you just found in part (e) is orthogonal to both $\mathbf{u}$ and $\mathbf{v}$.
7. Consider the vectors $\mathbf{u}$ and $\mathbf{v}$ in the xy-plane as shown below in Figure 9.4.2.



1.

   Figure 9.4.2. Two vectors in the xy-plane

   Explain why $\mathbf{u} = |\mathbf{u}|\mathbf{i}$ and $\mathbf{v} = |\mathbf{v}|\cos(\theta)\mathbf{i} + |\mathbf{v}|\sin(\theta)\mathbf{j}$. Then compute the length of $|\mathbf{u} \times \mathbf{v}|$.

Computing the cross product

As we have seen in Preview Activity 9.4.1, the cross product $\mathbf{u} \times \mathbf{v}$ is defined for two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^3$ and produces another vector in $\mathbb{R}^3$. Using the right-hand rule, we saw that

- $\mathbf{i} \times \mathbf{j} = \mathbf{k}$
- $\mathbf{j} \times \mathbf{i} = -\mathbf{k}$
- $\mathbf{i} \times \mathbf{k} = -\mathbf{j}$
- $\mathbf{k} \times \mathbf{i} = \mathbf{j}$
- $\mathbf{j} \times \mathbf{k} = \mathbf{i}$
- $\mathbf{k} \times \mathbf{j} = -\mathbf{i}$

If, in addition, we assume the cross product behaves like we think a product should (e.g., the cross product distributes over vector addition), we can compute the cross product in terms of the components of general vectors to find a formula for the cross product. Doing so we see that

$$\begin{aligned}\mathbf{u}\times\mathbf{v} &= (u_1\mathbf{i}+u_2\mathbf{j}+u_3\mathbf{k})\times(v_1\mathbf{i}+v_2\mathbf{j}+v_3\mathbf{k})\\ &= u_1\mathbf{i}\times(v_1\mathbf{i}+v_2\mathbf{j}+v_3\mathbf{k})+u_2\mathbf{j}\times(v_1\mathbf{i}+v_2\mathbf{j}+v_3\mathbf{k})\\ &\quad +u_3\mathbf{k}\times(v_1\mathbf{i}+v_2\mathbf{j}+v_3\mathbf{k})\\ &= u_1v_1\mathbf{i}\times\mathbf{i}+u_1v_2\mathbf{i}\times\mathbf{j}+u_1v_3\mathbf{i}\times\mathbf{k}+u_2v_1\mathbf{j}\times\mathbf{i}+u_2v_2\mathbf{j}\times\mathbf{j}\\ &\quad +u_2v_3\mathbf{j}\times\mathbf{k}+u_3v_1\mathbf{k}\times\mathbf{i}+u_3v_2\mathbf{k}\times\mathbf{j}+u_3v_3\mathbf{k}\times\mathbf{k}\\ &= u_1v_2\mathbf{k}-u_1v_3\mathbf{j}-u_2v_1\mathbf{k}+u_2v_3\mathbf{i}+u_3v_1\mathbf{j}-u_3v_2\mathbf{i}\\ &= (u_2v_3-u_3v_2)\mathbf{i}-(u_1v_3-u_3v_1)\mathbf{j}+(u_1v_2-u_2v_1)\mathbf{k}.\end{aligned}$$

(Like the dot product, the cross product arises in physical applications, e.g., torque, but it is more convenient mathematically to begin from an algebraic perspective.)

The previous calculations lead us to define the cross product of vectors in $\mathbb{R}^3$ as follows.

Definition 9.4.1: Cross Product

The *cross product* $\mathbf{u}\times\mathbf{v}$ of vectors $\mathbf{u}=u_1\mathbf{i}+u_2\mathbf{j}+u_3\mathbf{k}$ and $\mathbf{v}=v_1\mathbf{i}+v_2\mathbf{j}+v_3\mathbf{k}$ in $\mathbb{R}^3$ is the vector

$$(u_2v_3-u_3v_2)\mathbf{i}-(u_1v_3-u_3v_1)\mathbf{j}+(u_1v_2-u_2v_1)\mathbf{k}. \tag{9.4.1}$$

At first, this may look intimidating and difficult to remember. However, if we rewrite the expression in Equation (9.4.1) using determinants, important structure emerges. The determinant of a 2×2 matrix is

$$\begin{vmatrix} a & b\\ c & d\end{vmatrix} = ad-bc.$$

It follows that we can thus rewrite Equation (9.4.1) in the form

$$\mathbf{u}\times\mathbf{v} = \begin{vmatrix} u_2 & u_3\\ v_2 & v_3\end{vmatrix}\mathbf{i}-\begin{vmatrix} u_1 & u_3\\ v_1 & v_3\end{vmatrix}\mathbf{j}+\begin{vmatrix} u_1 & u_2\\ v_1 & v_2\end{vmatrix}\mathbf{k}.$$

For those familiar with the determinant of a 3×3 matrix, we write the mnemonic as

$$\mathbf{u}\times\mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k}\\ u_1 & u_2 & u_3\\ v_1 & v_2 & v_3\end{vmatrix}.$$

Activity 9.4.2

Suppose $\mathbf{u}=\langle 0,1,3\rangle$ and $\mathbf{v}=\langle 2,-1,0\rangle$. Use the formula (9.4.1) for the following.

1. Find the cross product $\mathbf{u}\times\mathbf{v}$.
2. Evaluate the dot products $\mathbf{u}\cdot(\mathbf{u}\times\mathbf{v})$ and $\mathbf{v}\cdot(\mathbf{u}\times\mathbf{v})$. What does this tell you about the geometric relationship among $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{u}\times\mathbf{v}$?
3. Find the cross product $\mathbf{v}\times\mathbf{i}$.
4. Multiplication of real numbers is *associative*, which means, for instance, that $(2\cdot 5)\cdot 3 = 2\cdot(5\cdot 3)$. Is it true that the cross product of vectors is associative? For instance, is it true that $(\mathbf{u}\times\mathbf{v})\times\mathbf{i}=\mathbf{u}\times(\mathbf{v}\times\mathbf{i})$?
5. Find the cross product $\mathbf{u}\times\mathbf{u}$.

The cross product satisfies the following properties, some of which were illustrated in Preview Activity 9.4.1 and may be easily verified from the definition (9.4.1).

Properties of the cross product

Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ be vectors in $\mathbb{R}^3$, and let c be a scalar. Then

1. $\mathbf{u} \times \mathbf{v} = -(\mathbf{v} \times \mathbf{u})$
2. $(\mathbf{u} + \mathbf{v}) \times \mathbf{w} = (\mathbf{u} \times \mathbf{w}) + (\mathbf{v} \times \mathbf{w})$
3. $(c\mathbf{u}) \times \mathbf{w} = c(\mathbf{u} \times \mathbf{w}) = \mathbf{u} \times (c\mathbf{w})$
4. $\mathbf{u} \times \mathbf{v} = \mathbf{0}$ if $\mathbf{u}$ and $\mathbf{v}$ are parallel.
5. The cross product is not associative; that is, in general
$$(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} \neq \mathbf{u} \times (\mathbf{v} \times \mathbf{w}).$$

Just as we found for the dot product, the cross product provides us with useful geometric information. In particular, both the length and direction of the cross product $\mathbf{u} \times \mathbf{v}$ encode information about the geometric relationship between $\mathbf{u}$ and $\mathbf{v}$.

The Length of u x v

We may ask whether the length $|\mathbf{u} \times \mathbf{v}|$ has any relationship to the lengths of $\mathbf{u}$ and $\mathbf{v}$. To investigate, we will compute the square of the length $|\mathbf{u} \times \mathbf{v}|^2$ and denote by θ the angle between $\mathbf{u}$ and $\mathbf{v}$, as in Section 9.3. Doing so, we find through some significant algebra that

$$\begin{aligned}
|\mathbf{u} \times \mathbf{v}|^2 &= (u_2v_3 - u_3v_2)^2 + (u_1v_3 - u_3v_1)^2 + (u_1v_2 - u_2v_1)^2 \\
&= u_2^2v_3^2 - 2u_2u_3v_2v_3 + u_3^2v_2^2 + u_1^2v_3^2 - 2u_1u_3v_1v_3 + u_3^2v_1^2 \\
&\quad + u_1^2v_2^2 - 2u_1u_2v_1v_2 + u_2^2v_1^2 \\
&= u_1^2(v_2^2 + v_3^2) + u_2^2(v_1^2 + v_3^2) + u_3^2(v_1^2 + v_2^2) \\
&\quad - 2(u_1u_2v_1v_2 + u_1u_3v_1v_3 + u_2u_3v_2v_3) \\
&= u_1^2(v_1^2 + v_2^2 + v_3^2) + u_2^2(v_1^2 + v_2^2 + v_3^2) + u_3^2(v_1^2 + v_2^2 + v_3^2) \\
&\quad - (u_1^2v_1^2 + u_2^2v_2^2 + u_3^2v_3^2 + 2(u_1u_2v_1v_2 + u_1u_3v_1v_3 + u_2u_3v_2v_3)) \\
&= (u_1^2 + u_2^2 + u_3^2)(v_1^2 + v_2^2 + v_3^2) - (u_1v_1 + u_2v_2 + u_3v_3)^2 \\
&= |\mathbf{u}|^2|\mathbf{v}|^2 - (\mathbf{u} \cdot \mathbf{v})^2 \\
&= |\mathbf{u}|^2|\mathbf{v}|^2(1 - \cos^2(\theta)) \\
&= |\mathbf{u}|^2|\mathbf{v}|^2 \sin^2(\theta).
\end{aligned}$$

Therefore, we have found $|\mathbf{u} \times \mathbf{v}|^2 = |\mathbf{u}|^2|\mathbf{v}|^2 \sin^2(\theta)$, which means that

$$|\mathbf{u} \times \mathbf{v}| = |\mathbf{u}||\mathbf{v}| \sin(\theta). \tag{9.4.2}$$

Note that the third property stated above says that $\mathbf{u} \times \mathbf{v} = \mathbf{0}$ if $\mathbf{u}$ and $\mathbf{v}$ are parallel. This is reflected in Equation (9.4.2) since $\sin(\theta) = 0$ if $\mathbf{u}$ and $\mathbf{v}$ are parallel, which implies that $\mathbf{u} \times \mathbf{v} = \mathbf{0}$.

Equation (9.4.2) also has a geometric implication. Consider the parallelogram formed by two vectors $\mathbf{u}$ and $\mathbf{v}$, as shown in Figure 9.4.3.

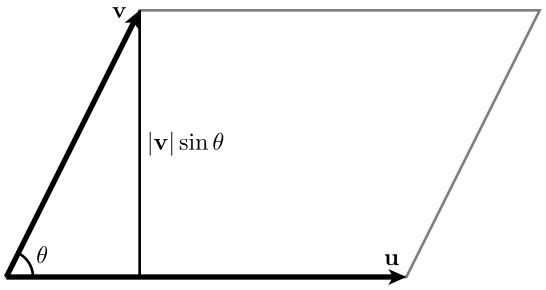


Figure 9.4.3. The parallelogram formed by $\mathbf{u}$ and $\mathbf{v}$

Remember that the area of a parallelogram is the product of its base and height. As shown in the figure, we may consider the base of the parallelogram to be $|\mathbf{u}|$ and the height to be $|\mathbf{v}|\sin(\theta)$. This means that the area of the parallelogram formed by $\mathbf{u}$ and $\mathbf{v}$ is

$$|\mathbf{u}||\mathbf{v}|\sin(\theta) = |\mathbf{u}\times\mathbf{v}|.$$

This leads to the following interesting fact.

The length of the cross product

The length, $|\mathbf{u}\times\mathbf{v}|$, of the cross product of vectors $\mathbf{u}$ and $\mathbf{v}$ is the area of the parallelogram determined by $\mathbf{u}$ and $\mathbf{v}$.

Note also that if $\mathbf{u} = u_1\mathbf{i} + u_2\mathbf{j} + 0\mathbf{k}$ and $\mathbf{v} = v_1\mathbf{i} + v_2\mathbf{j} + 0\mathbf{k}$ are vectors in the xy-plane, then Equation (9.4.1) shows that the area of the parallelogram determined by $\mathbf{u}$ and $\mathbf{v}$ is $|\mathbf{u}\times\mathbf{v}| = |u_1v_2 - u_2v_1|$ is the absolute value of the 2×2 determinant $\begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix}$. So the absolute value of a determinant of a 2×2 matrix is also the area of a parallelogram.

Activity 9.4.3

1. Find the area of the parallelogram formed by the vectors $\mathbf{u} = \langle 1, 3, -2\rangle$ and $\mathbf{v} = \langle 3, 0, 1\rangle$.
2. Find the area of the parallelogram in $\mathbb{R}^3$ whose vertices are $(1, 0, 1)$, $(0, 0, 1)$, $(2, 1, 0)$, and $(1, 1, 0)$. (Hint: It might be helpful to draw a picture to see how the vertices are arranged so you can determine which vectors you might use.)

The Direction of $\mathbf{u}\times\mathbf{v}$

Now that we understand the length of $\mathbf{u}\times\mathbf{v}$, we will investigate its direction. Remember from Preview Activity 9.4.1 that cross products involving the vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ resulted in vectors that are orthogonal to the two terms. We will see that this holds more generally.

We begin by computing $\mathbf{u}\cdot(\mathbf{u}\times\mathbf{v})$, and see that

$$\begin{aligned}\mathbf{u}\cdot(\mathbf{u}\times\mathbf{v}) &= u_1(u_2v_3 - u_3v_2) - u_2(u_1v_3 - u_3v_1) + u_3(u_1v_2 - u_2v_1)\\ &= u_1u_2v_3 - u_1u_3v_2 - u_2u_1v_3 + u_2u_3v_1 + u_3u_1v_2 - u_3u_2v_1\\ &= 0\end{aligned}$$

To summarize, we have $\mathbf{u}\cdot(\mathbf{u}\times\mathbf{v}) = 0$, which implies that $\mathbf{u}$ is orthogonal to $\mathbf{u}\times\mathbf{v}$. In the same way, we can show that $\mathbf{v}$ is orthogonal to $\mathbf{u}\times\mathbf{v}$. The net effect is that $\mathbf{u}\times\mathbf{v}$ is a vector that is perpendicular to both $\mathbf{u}$ and $\mathbf{v}$, and hence $\mathbf{u}\times\mathbf{v}$ is perpendicular to the plane determined by $\mathbf{u}$ and $\mathbf{v}$. Moreover, the direction of $\mathbf{u}\times\mathbf{v}$ is determined by applying the right-hand rule to $\mathbf{u}$ and $\mathbf{v}$, as we saw in Preview Activity 9.4.1. In light of our earlier work that showed $|\mathbf{u}||\mathbf{v}|\sin(\theta) = |\mathbf{u}\times\mathbf{v}|.$, we may now express $\mathbf{u}\times\mathbf{v}$ in the following different way.

The cross product as normal vector

Suppose that $\mathbf{u}$ and $\mathbf{v}$ are not parallel and that $\mathbf{n}$ is the unit vector perpendicular to the plane containing $\mathbf{u}$ and $\mathbf{v}$ determined by the right-hand rule. Then

$$\mathbf{u}\times\mathbf{v} = |\mathbf{u}||\mathbf{v}|\sin(\theta)\ \mathbf{n}.$$

There is yet one more geometric implication we may draw from this result. Suppose $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ are vectors in $\mathbb{R}^3$ that are not coplanar and that form a three-dimension parallelepiped as shown in Figure 9.4.4.

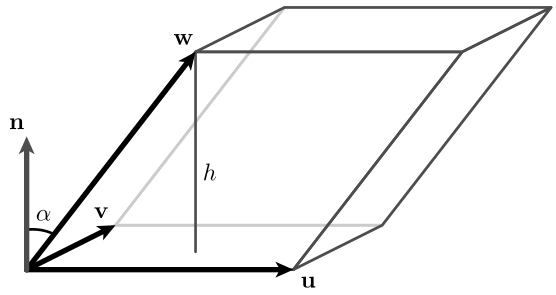


Figure 9.4.4. The parallelepiped determined by $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$

The volume of the parallelepiped is determined by multiplying A, the area of the base, by the height h. As we have just seen, the area of the base is $|\mathbf{u} \times \mathbf{v}|$. Moreover, the height $h = |\mathbf{w}| \cos(\alpha)$ where α is the angle between $\mathbf{w}$ and the vector $\mathbf{n}$, which is orthogonal to the plane formed by $\mathbf{u}$ and $\mathbf{v}$. Since $\mathbf{n}$ is parallel to $\mathbf{u} \times \mathbf{v}$, the angle between $\mathbf{w}$ and $\mathbf{u} \times \mathbf{v}$ is also α. This shows that

$$|(\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w}| = |\mathbf{u} \times \mathbf{v}||\mathbf{w}| \cos(\alpha) = Ah,$$

and therefore

The cross product and the volume of a parallelepiped

The volume of the parallelepiped determined by $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ is $|(\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w}|$.

As a dot product of two vectors, the quantity $(\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w}$ is a scalar and is called the *triple scalar product*.

Activity 9.4.4

Suppose $\mathbf{u} = \langle 3, 5, -1 \rangle$ and $\mathbf{v} = \langle 2, -2, 1 \rangle$.

1. Find two unit vectors orthogonal to both $\mathbf{u}$ and $\mathbf{v}$.
2. Find the volume of the parallelepiped formed by the vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w} = \langle 3, 3, 1 \rangle$.
3. Find a vector orthogonal to the plane containing the points $(0, 1, 2)$, $(4, 1, 0)$, and $(-2, 2, 2)$.
4. Given the vectors $\mathbf{u}$ and $\mathbf{v}$ shown below in Figure 9.4.5, sketch the cross products $\mathbf{u} \times \mathbf{v}$ and $\mathbf{v} \times \mathbf{u}$.

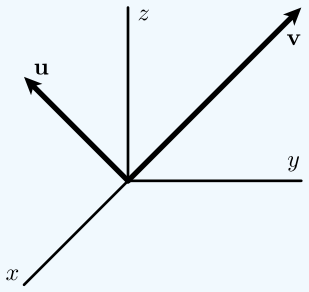


Figure 9.4.5. Vectors $\mathbf{u}$ and $\mathbf{v}$

5. Do the vectors $\mathbf{a} = \langle 1, 3, -2 \rangle$,$\mathbf{b} = \langle 2, 1, -4 \rangle$, and $\mathbf{c} = \langle 0, 1, 0 \rangle$ in standard position lie in the same plane? Use the concepts from this section to explain.

Torque is measured by a cross product

We have seen that the cross product enables us to produce a vector perpendicular to two given vectors, to measure the area of a parallelogram, and to measure the volume of a parallelepiped. Besides these geometric applications, the cross product also

enables us to describe a physical quantity called *torque*.

Suppose that we would like to turn a bolt using a wrench as shown in Figure 9.4.6. If a force $\mathbf{F}$ is applied to the wrench and $\mathbf{r}$ is the vector from the position on the wrench at which the force is applied to center of the bolt, we define the torque, τ, to be

$$\tau = \mathbf{F} \times \mathbf{r}.$$

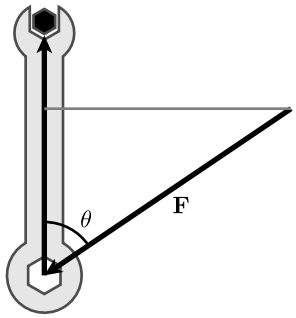


Figure 9.4.6. A force applied to a wrench

When a force is applied to an object, Newton's Second Law tells us that the force is equal to the rate of change of the object's linear momentum. Similarly, the torque applied to an object is equal to the rate of change of the object's angular momentum. In other words, torque will cause the bolt to rotate.

In many industrial applications, bolts are required to be tightened using a specified torque. Of course, the magnitude of the torque is $|\tau| = |\mathbf{F} \times \mathbf{r}| = |\mathbf{F}||\mathbf{r}||\sin(\theta)$. Thus, to produce a larger torque, we can increase either $|\mathbf{F}|$ or $|\mathbf{r}|$, which you may know if you have ever removed lug nuts when changing a flat tire. The ancient Greek mathematician Archimedes said: “Give me a lever long enough and a fulcrum on which to place it, and I shall move the world.” A modern spin on this statement is: “Allow me to make $|\mathbf{r}|$ large enough, and I shall produce a torque large enough to move the world.”

Comparing the dot and cross products

Finally, it is worthwhile to compare and contrast the dot and cross products.

- $\mathbf{u} \cdot \mathbf{v}$ is a scalar, while $\mathbf{u} \times \mathbf{v}$ is a vector.
- $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$, while $\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$
- $\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}|\cos(\theta)$, while $|\mathbf{u} \times \mathbf{v}| = |\mathbf{u}||\mathbf{v}|\sin(\theta)$.
- $\mathbf{u} \cdot \mathbf{v} = 0$ if $\mathbf{u}$ and $\mathbf{v}$ are perpendicular, while $\mathbf{u} \times \mathbf{v} = \mathbf{0}$ if $\mathbf{u}$ and $\mathbf{v}$ are parallel.

Summary

- The cross product is defined *only* for vectors in $\mathbb{R}^3$. The cross product of vectors $\mathbf{u} = u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k}$ and $\mathbf{v} = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$ in $\mathbb{R}^3$ is the vector

  $$\mathbf{u} \times \mathbf{v} = (u_2v_3 - u_3v_2)\mathbf{i} - (u_1v_3 - u_3v_1)\mathbf{j} + (u_1v_2 - u_2v_1)\mathbf{k}.$$

- Geometrically, the cross product is

  $$\mathbf{u} \times \mathbf{v} = |\mathbf{u}|\ |\mathbf{v}|\ \sin(\theta)\ \mathbf{n},$$

  where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$ and $\mathbf{n}$ is a unit vector perpendicular to both $\mathbf{u}$ and $\mathbf{v}$ as determined by the right-hand rule.

- The cross product of vectors $\mathbf{u}$ and $\mathbf{v}$ is a vector perpendicular to both $\mathbf{u}$ and $\mathbf{v}$.
- The magnitude $|\mathbf{u} \times \mathbf{v}|$ of the cross product of the vectors $\mathbf{u}$ and $\mathbf{v}$ gives the area of the parallelogram determined by $\mathbf{u}$ and $\mathbf{v}$. Also, the scalar triple product $|(\mathbf{u} \times \mathbf{v}) \cdot \mathbf{w}|$ gives the volume of the parallelepiped determined by $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$.

This page titled 9.4: The Cross Product is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 9.4: The Cross Product by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

9.5: Lines and Planes in Space

Motivating Questions

- How are lines in $\mathbb{R}^3$ similar to and different from lines in $\mathbb{R}^2$?
- What is the role that vectors play in representing equations of lines, particularly in $\mathbb{R}^3$?
- How can we think of a plane as a set of points determined by a point and a vector?
- How do we find the equation of a plane through three given non-collinear points?

In single variable calculus, we learn that a differentiable function is *locally linear*. In other words, if we zoom in on the graph of a differentiable function at a point, the graph looks like the tangent line to the function at that point. Linear functions played important roles in single variable calculus, useful in approximating differentiable functions, in approximating roots of functions (see Newton's Method), and approximating solutions to first order differential equations (see Euler's Method). In multivariable calculus, we will soon study curves in space; differentiable curves turn out to be locally linear as well. In addition, as we study functions of two variables, we will see that such a function is locally linear at a point if the surface defined by the function looks like a plane (the tangent plane) as we zoom in on the graph.

Consequently, it is important for us to understand both lines and planes in space, as these define the linear functions in $\mathbb{R}^2$ and $\mathbb{R}^3$. (Recall that a function is linear if it is a polynomial function whose terms all have degree less than or equal to 1. For example, x defines a single variable linear function and $x + y$ a two variable linear function, but xy is not linear since it has degree two, the sum of the degress of its factors.) We begin our work by considering some familiar ideas in $\mathbb{R}^2$ but from a new perspective.

Preview Activity $9.5.1$

We are familiar with equations of lines in the plane in the form $y = mx + b$, where m is the slope of the line and $(0, b)$ is the y-intercept. In this activity, we explore a more flexible way of representing lines that we can use not only in the plane, but in higher dimensions as well.

To begin, consider the line through the point $(2, -1)$ with slope $\frac{2}{3}$ as shown in Figure $9.5.1$.

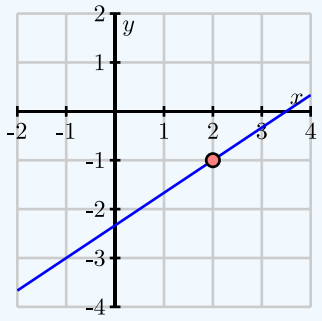


Figure $9.5.1$. The line through $(2, -1)$ with slope $\frac{2}{3}$.

1. Suppose we increase x by 1 from the point $(2, -1)$. How does the y-value change? What is the point on the line with x-coordinate 3?
2. Suppose we decrease x by 3.25 from the point $(2, -1)$. How does the y-value change? What is the point on the line with x-coordinate -1.25?
3. Now, suppose we increase x by some arbitrary value $3t$ from the point $(2, -1)$. How does the y-value change? What is the point on the line with x-coordinate $2 + 3t$?
4. Observe that the slope of the line is related to any vector whose y-component divided by the x-component is the slope of the line. For the line in this exercise, we might use the vector $\langle 3, 2\rangle$, which describes the direction of the line. Explain why the terminal points of the vectors $\mathbf{r}(t)$, where

$$\mathbf{r}(t) = \langle 2, -1\rangle + \langle 3, 2\rangle t,$$

trace out the graph of the line through the point $(2, -1)$ with slope $\frac{2}{3}$.

5. Now we extend this vector approach to $\mathbb{R}^3$ and consider a second example. Let $\mathcal{L}$ be the line in $\mathbb{R}^3$ through the point $(1, 0, 2)$ in the direction of the vector $\langle 2, -1, 4\rangle$. Find the coordinates of three distinct points on line $\mathcal{L}$. Explain your thinking.
6. Find a vector in the form

$$\mathbf{r}(t) = \langle x_0, y_0, z_0\rangle + \langle a, b, c\rangle t$$

whose terminal points trace out the line $\mathcal{L}$ that is described in (e). That is, you should be able to locate any point on the line by determining a corresponding value of t.

Lines in Space

In two-dimensional space, a non-vertical line is defined to be the set of points satisfying the equation

$$y = mx + b,$$

for some constants m and b. The value of m (the slope) tells us how the dependent variable changes for every one unit increase in the independent variable, while the point $(0, b)$ is the y-intercept and anchors the line to a location on the y-axis. Alternatively, we can think of the slope as being related to the vector $\langle 1, m\rangle$, which tells us the direction of the line, as shown on the left in Figure 9.5.2. Thus, we can identify a line in space by fixing a point P and a direction $\mathbf{v}$, as shown on the right. Since we also have vectors in space to provide direction, this same idea of a point and a direction determining a line works in $\mathbb{R}^n$ for any n.

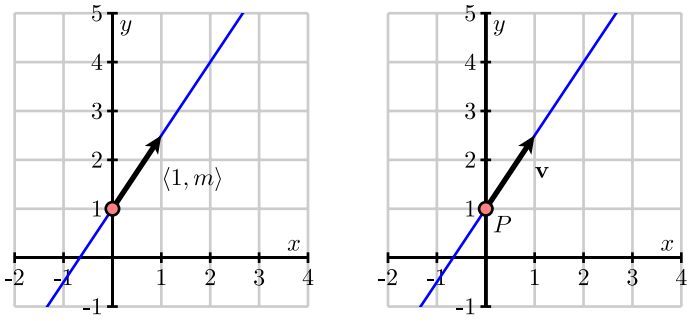


Figure 9.5.2. A vector description of a line

Definition 9.5.1

A *line* in space is the set of terminal points of vectors emanating from a given point P that are parallel to a fixed vector $\mathbf{v}$.

The fixed vector $\mathbf{v}$ in the definition is called a *direction vector* for the line. As we saw in Preview Activity 9.5.1, to find an equation for a line through point P in the direction of vector $\mathbf{v}$, observe that any vector parallel to $\mathbf{v}$ will have the form $t\mathbf{v}$ for some scalar t. So, any vector emanating from the point P in a direction parallel to the vector $\mathbf{v}$ will be of the form

$$\overrightarrow{OP} + \mathbf{v}t \tag{9.5.1}$$

for some scalar t (where O is the origin).

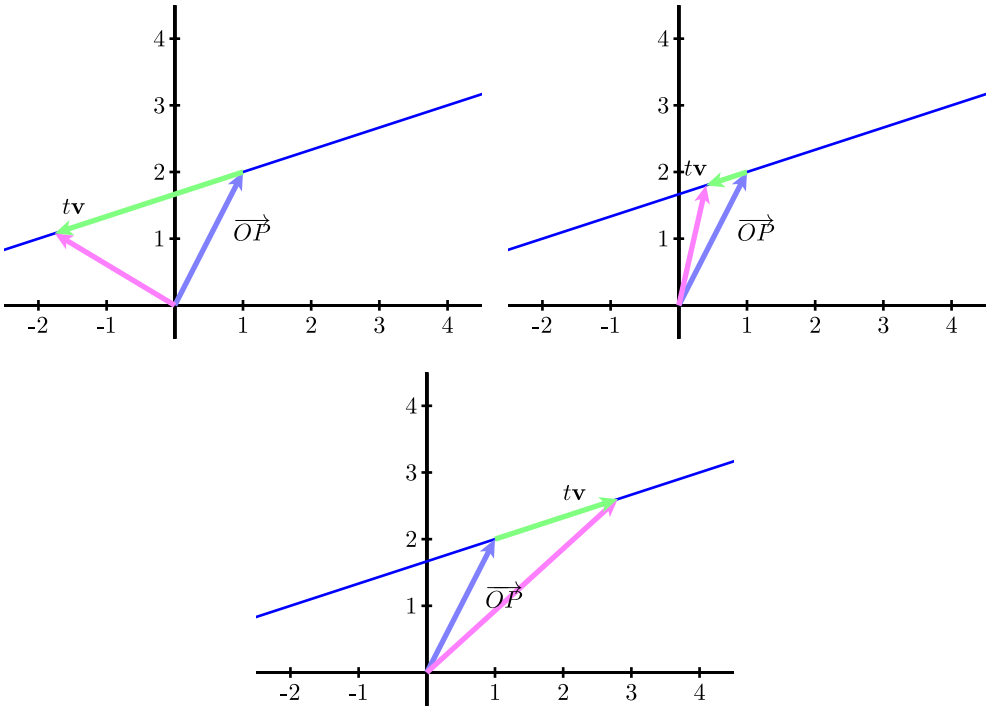


Figure 9.5.3. A line in 2-space.

Figure $9.5.3$ presents three images of a line in two-space in which we can identify the vector $\overrightarrow{OP}$ and the vector $t\mathbf{v}$ as in Equation $(9.5.1)$. Here, $\overrightarrow{OP}$ is the fixed vector shown in blue, while the direction vector $\mathbf{v}$ is the vector parallel to the vector shown in green (that is, the green vector represents $t\mathbf{v}$, and the line is traced out by the terminal points of the magenta vector). In other words, the tips (terminal points) of the magenta vectors (the vectors of the form $\overrightarrow{OP} + t\mathbf{v}$) trace out the line as t changes.

In particular, the terminal points of the vectors of the form in $(9.5.1)$ define a linear function $\mathbf{r}$ in space of the following form, which is valid in any dimension.

The vector form of a line

The *vector form* of a line through the point P in the direction of the vector $\mathbf{v}$ is

$$\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{v}, \tag{9.5.2}$$

where $\mathbf{r}_0$ is the position vector $\overrightarrow{OP}$ from the origin to the point P.

Of course, it is common to represent lines in the plane using the slope-intercept equation $y = mx + b$. The vector form of the line, described above, is an alternative way to represent lines that has the following two advantages. First, in two dimensions, we are able to represent vertical lines, whose slope m is not defined, using a vertical direction vector, such as $\mathbf{v} = \langle 0, 1 \rangle$. Second, this description of lines works in any dimension even though there is no concept of the slope of a line in more than two dimensions.

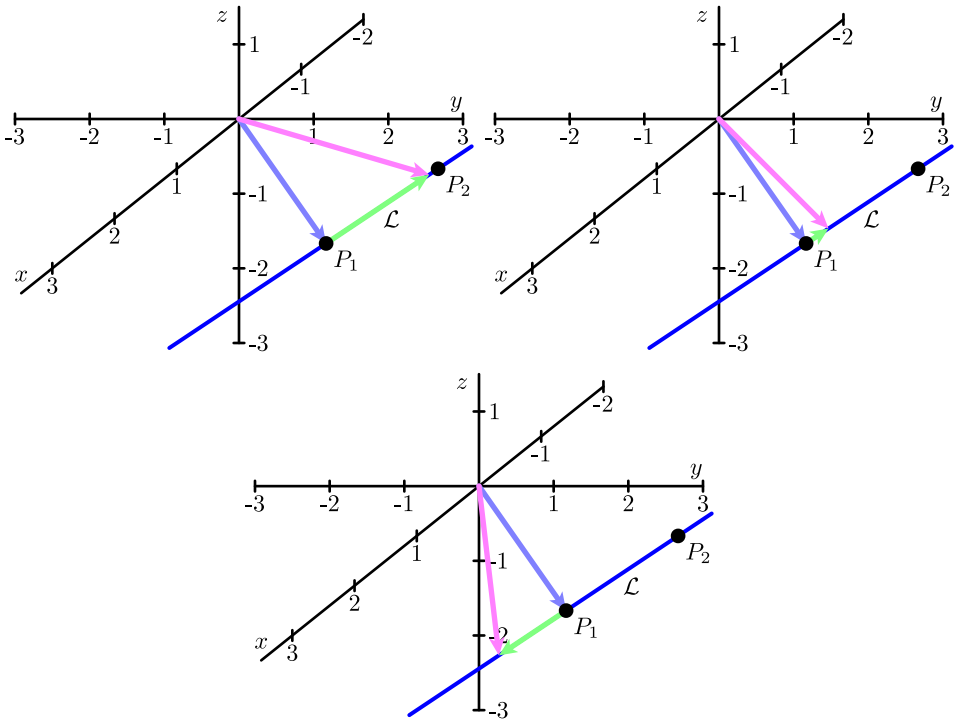


Figure 9.5.4. A line in 3-space.

Activity 9.5.2

Let $P_1 = (1, 2, -1)$ and $P_2 = (-2, 1, -2)$. Let $\mathcal{L}$ be the line in $\mathbb{R}^3$ through P_1 and P_2, and note that three snapshots of this line are shown in Figure 9.5.5.

1. Find a direction vector for the line $\mathcal{L}$.
2. Find a vector equation of $\mathcal{L}$ in the form $\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{v}$.
3. Consider the vector equation $\mathbf{s}(t) = \langle -5, 0, -3\rangle + t\langle 6, 2, 2\rangle$. What is the direction of the line given by $\mathbf{s}(t)$? Is this new line parallel to line $\mathcal{L}$?
4. Do $\mathbf{r}(t)$ and $\mathbf{s}(t)$ represent the same line, $\mathcal{L}$? Explain.

The Parametric Equations of a Line

The vector form of a line, $\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{v}$ in Equation (9.5.2), describes a line as the set of terminal points of the vectors $\mathbf{r}(t)$. If we write this in terms of components letting

$$\mathbf{r}(t) = \langle x(t), y(t), z(t)\rangle, \quad \mathbf{r}_0 = \langle x_0, y_0, z_0\rangle, \quad \text{and} \quad \mathbf{v} = \langle a, b, c\rangle,$$

then we can equate the components on both sides of $\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{v}$ to obtain the equations

$$x(t) = x_0 + at, \quad y(t) = y_0 + bt, \quad \text{and} \quad z(t) = z_0 + ct,$$

which describe the coordinates of the points on the line. The variable t represents an arbitrary scalar and is called a *parameter*. In particular, we use the following language.

The parametric equations of a line

The *parametric equations* for a line through the point $P = (x_0, y_0, z_0)$ in the direction of the vector $\mathbf{v} = \langle a, b, c\rangle$ are

$$x(t) = x_0 + at, \qquad y(t) = y_0 + bt, \qquad z(t) = z_0 + ct.$$

Notice that there are many different parametric equations for the same line. For example, choosing another point P on the line or another direction vector $\mathbf{v}$ produces another set of parametric equations. It is sometimes useful to think of t as a time parameter and the parametric equations as telling us where we are on the line at each time. In this way, the parametric equations describe a particular walk taken along the line; there are, of course, many possible ways to walk along a line.

Activity 9.5.3

Let $P_1 = (1, 2, -1)$ and $P_2 = (-2, 1, -2)$, and let $\mathcal{L}$ be the line in $\mathbb{R}^3$ through P_1 and P_2, which is the same line as in Activity 9.5.2.

1. Find parametric equations of the line $\mathcal{L}$.
2. Does the point $(1, 2, 1)$ lie on $\mathcal{L}$? If so, what value of t results in this point?
3. Consider another line, $\mathcal{K}$, whose parametric equations are
$$x(s) = 11 + 4s, \quad y(s) = 1 - 3s, \quad z(s) = 3 + 2s.$$
What is the direction of the line $\mathcal{K}$?
4. Do the lines $\mathcal{L}$ and $\mathcal{K}$ intersect? If so, provide the point of intersection and the t and s values, respectively, that result in the point. If not, explain why.

Planes in Space

Now that we have a way of describing lines, we would like to develop a means of describing planes in three dimensions. We studied the coordinate planes and planes parallel to them in Section 9.1. Each of those planes had one of the variables x, y, or z equal to a constant. We can note that any vector in a plane with x constant is orthogonal to the vector $\langle 1, 0, 0\rangle$, any vector in a plane with y constant is orthogonal to the vector $\langle 0, 1, 0\rangle$, and any vector in a plane with z constant is orthogonal to the vector $\langle 0, 0, 1\rangle$. This idea works in general to define a plane.

Definition 9.5.6

A *plane* p in space is the set of all terminal points of vectors emanating from a given point P_0 perpendicular to a fixed vector $\mathbf{n}$, as shown in Figure 9.5.5.

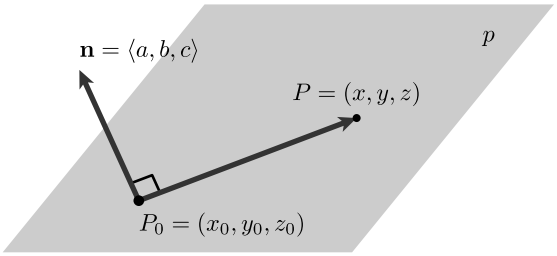


Figure 9.5.5. A point P_0 on a plane p with a normal vector $\mathbf{n}$

The definition allows us to find the equation of a plane. Assume that $\mathbf{n} = \langle a, b, c\rangle$, $P_0 = (x_0, y_0, z_0)$, and that $P = (x, y, z)$ is an arbitrary point on the plane. Since the vector $\overrightarrow{P_0P}$ lies in the plane, it must be perpendicular to $\mathbf{n}$. This means that

$$\begin{aligned} 0 &= \mathbf{n} \cdot \overrightarrow{P_0P} \\ &= \mathbf{n} \cdot \Big(\langle x, y, z \rangle - \langle x_0, y_0, z_0 \rangle \Big) \\ &= \mathbf{n} \cdot \langle x - x_0, y - y_0, z - z_0 \rangle \\ &= a(x - x_0) + b(y - y_0) + c(z - z_0). \end{aligned}$$

The fixed vector $\mathbf{n}$ perpendicular to the plane is frequently called a *normal vector* to the plane. We may now summarize as follows.

Equations of a plane

- The *scalar equation* of the plane with normal vector $\mathbf{n} = \langle a, b, c \rangle$ containing the point $P_0 = (x_0, y_0, z_0)$ is
$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0. \tag{9.5.3}$$
- The *vector equation* of the plane with normal vector $\mathbf{n} = \langle a, b, c \rangle$ containing the points $P_0 = (x_0, y_0, z_0)$ and $P = (x, y, z)$ is
$$\mathbf{n} \cdot \overrightarrow{P_0P} = 0. \tag{9.5.4}$$

We may take the scalar equation of a plane a little further and note that since

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0,$$

it equivalently follows that

$$ax + by + cz = ax_0 + by_0 + cz_0.$$

That is, we may write an equation of a plane as $ax + by + cz = d$ where $d = \mathbf{n} \cdot \langle x_0, y_0, z_0 \rangle$.

For instance, if we would like to describe the plane passing through the point $P_0 = (4, -2, 1)$ and perpendicular to the vector $\mathbf{n} = \langle 1, 2, 1 \rangle$, we have

$$\langle 1, 2, 1 \rangle \cdot \langle x, y, z \rangle = \langle 1, 2, 1 \rangle \cdot \langle 4, -2, 1 \rangle$$

or

$$x + 2y + z = 1.$$

Notice that the coefficients of x, y, and z in this description give a vector perpendicular to the plane. For instance, if we are presented with the plane

$$-2x + y - 3z = 4,$$

we know that $\mathbf{n} = \langle -2, 1, -3 \rangle$ is a vector perpendicular to the plane.

Activity 9.5.4

1. Write a scalar equation of the plane p_1 passing through the point $(0, 2, 4)$ and perpendicular to the vector $\mathbf{n} = \langle 2, -1, 1 \rangle$.
2. Is the point $(2, 0, 2)$ on the plane p_1?
3. Write a scalar equation of the plane p_2 that is parallel to p_1 and passing through the point $(3, 0, 4)$. (Hint: Compare normal vectors of the planes.)
4. Write a parametric description of the line l passing through the point $(2, 0, 2)$ and perpendicular to the plane p_3 described by the equation $x + 2y - 2z = 7$.
5. Find the point at which l intersects the plane p_3.

Just as two distinct points in space determine a line, three non-collinear points in space determine a plane. Consider three points P_0, P_1, and P_2 in space, not all lying on the same line as shown in Figure 9.5.6.

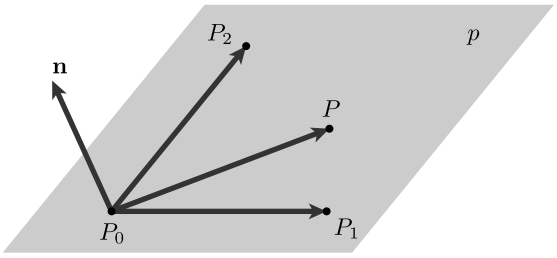


Figure 9.5.6. A plane determined by three points P_0, P_1, and P_2

Observe that the vectors $\overrightarrow{P_0P_1}$ and $\overrightarrow{P_0P_2}$ both lie in the plane p. If we form their cross-product

$$\mathbf{n} = \overrightarrow{P_0P_1} \times \overrightarrow{P_0P_2},$$

we obtain a normal vector to the plane p. Therefore, if P is any other point on p, it then follows that $\overrightarrow{P_0P}$ will be perpendicular to $\mathbf{n}$, and we have, as before, the equation

$$\mathbf{n} \cdot \overrightarrow{P_0P} = 0. \tag{9.5.5}$$

Activity 9.5.5

Let $P_0 = (1, 2, -1)$, $P_1 = (1, 0, -1)$, and $P_2 = (0, 1, 3)$ and let p be the plane containing P_0, P_1, and P_2.

1. Determine the components of the vectors $\overrightarrow{P_0P_1}$ and $\overrightarrow{P_0P_2}$.
2. Find a normal vector $\mathbf{n}$ to the plane p.
3. Find a scalar equation of the plane p.
4. Consider a second plane, q, with scalar equation $-3(x-1) + 4(y+3) + 2(z-5) = 0$. Find two different points on plane q, as well as a vector $\mathbf{m}$ that is normal to q.
5. The angle between two planes is the acute angle between their respective normal vectors. What is the angle between planes p and q?

Summary

- While lines in $\mathbb{R}^3$ do not have a slope, like lines in $\mathbb{R}^2$ they can be characterized by a point and a direction vector. Indeed, we define a line in space to be the set of terminal points of vectors emanating from a given point that are parallel to a fixed vector.
- Vectors play a critical role in representing the equation of a line. In particular, the terminal points of the vector $\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{v}$ define a linear function $\mathbf{r}$ in space through the terminal point of the vector $\mathbf{r}_0$ in the direction of the vector $\mathbf{v}$, tracing out a line in space.
- A plane in space is the set of all terminal points of vectors emanating from a given point perpendicular to a fixed vector.
- If P_1, P_2, and P_3 are non-collinear points in space, the vectors $\overrightarrow{P_1P_2}$ and and $\overrightarrow{P_1P_3}$ are vectors in the plane and the vector $\mathbf{n} = \overrightarrow{P_1P_2} \times \overrightarrow{P_1P_3}$ is a normal vector to the plane. So any point P in the plane satisfies the equation $\overrightarrow{PP_1} \cdot \mathbf{n} = 0$. If we let $P = (x, y, z)$, $\mathbf{n} = \langle a, b, c \rangle$ be the normal vector, and $P_1 = (x_0, y_0, z_0)$, we can also represent the plane with the equation
$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0.$$

This page titled 9.5: Lines and Planes in Space is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 9.5: Lines and Planes in Space by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

9.6: Vector-Valued Functions

Motivating Questions

- What is a vector-valued function? What do we mean by the graph of a vector-valued function?
- What is a parameterization of a curve in $\mathbb{R}^2$? In $\mathbb{R}^3$? What can the parameterization of a curve tell us?

So far, we have seen several different examples of curves in space, including traces and contours of functions of two variables, as well as lines in 3-space. Recall that for a line through a fixed point $\mathbf{r}_0$ in the direction of vector $\mathbf{v}$, we may express the line parametrically through the single vector equation

$$\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{v}.$$

From this perspective, the vector $\mathbf{r}(t)$ is a function that depends on the parameter t, and the terminal points of this vector trace out the line in space.

Like lines, other curves in space are one-dimensional objects, and thus we aspire to similarly express the coordinates of points on a given curve in terms of a single variable. Vectors are a perfect vehicle for doing so — we can use vectors based at the origin to identify points in space, and connect the terminal points of these vectors to draw a curve in space. This approach will allow us to draw an incredible variety of graphs in 2- and 3-space, as well as to identify and describe curves in n-space for any n. It will also allow us to represent traces and cross sections of surfaces in space.

Preview Activity 9.6.1

In this activity we consider how we might use vectors to define a curve in space.

1. On a single set of axes in $\mathbb{R}^2$, draw the vectors
   - $\langle \cos(0), \sin(0) \rangle$,
   - $\langle \cos\left(\frac{\pi}{2}\right), \sin\left(\frac{\pi}{2}\right) \rangle$,
   - $\langle \cos(\pi), \sin(\pi) \rangle$, and
   - $\left\langle \cos\left(\frac{3\pi}{2}\right), \sin\left(\frac{3\pi}{2}\right) \right\rangle$

   with their initial points at the origin.

2. On the same set of axes, draw the vectors
   - $\langle \cos\left(\frac{\pi}{4}\right), \sin\left(\frac{\pi}{4}\right) \rangle$,
   - $\langle \cos\left(\frac{3\pi}{4}\right), \sin\left(\frac{3\pi}{4}\right) \rangle$,
   - $\langle \cos\left(\frac{5\pi}{4}\right), \sin\left(\frac{5\pi}{4}\right) \rangle$, and
   - $\left\langle \cos\left(\frac{7\pi}{4}\right), \sin\left(\frac{7\pi}{4}\right) \right\rangle$

   with their initial points at the origin.

3. Based on the pictures from parts (a) and (b), sketch the set of *terminal* points of all of the vectors of the form $\langle \cos(t), \sin(t) \rangle$, where t assumes values from 0 to 2π. What is the resulting figure? Why?
4. Suppose we sketched the terminal points of all vectors of the form $\langle \cos(t), \sin(t) \rangle$, where t assumes values from 0 to π. How does the resulting picture differ from the one in part (c)? What about for t from 0 to 4π?

Vector-Valued Functions

Consider the curve shown in Figure 9.6.1. As in Preview Activity 9.6.1, we can think of a point on this curve as resulting from a vector from the origin to the point. As the point travels along the curve, the vector changes in order to terminate at the desired

point. A few still pictures of this motion are shown in Figure 9.6.1.

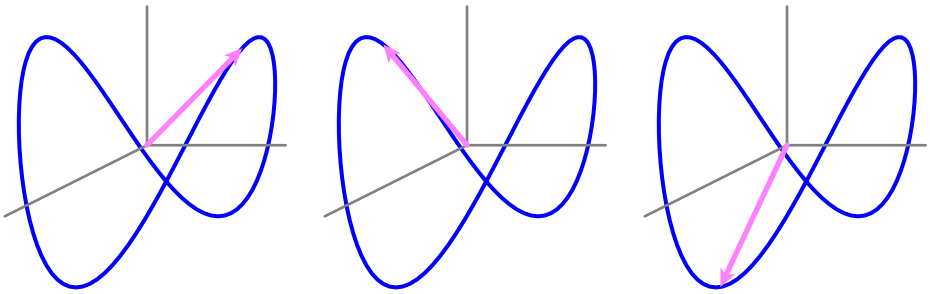


Figure 9.6.1. The graph of a curve in space.

Thus, we can think of the curve as a collection of terminal points of vectors emanating from the origin. We therefore view a point traveling along this curve as a function of time t, and define a function $\mathbf{r}$ whose input is the variable t and whose output is the vector from the origin to the point on the curve at time t. In so doing, we have introduced a new type of function, one whose input is a scalar and whose output is a vector.

The terminal points of the vector outputs of $\mathbf{r}$ then trace out the curve in space. From this perspective, the x, y, and z coordinates of the point are functions of time, t, say

$$x = x(t), \quad y = y(t), \quad \text{and} \quad z = z(t),$$

and thus we have three coordinate functions that enable us to represent the curve. The variable t is called a *parameter* and the equations $x = x(t)$, $y = y(t)$, and $z = z(t)$ are called *parametric equations* (or a *parameterization of the curve*). The function $\mathbf{r}$ whose output is the vector from the origin to a point on the curve is defined by

$$\mathbf{r}(t) = \langle x(t), y(t), z(t) \rangle.$$

Note that the input of $\mathbf{r}$ is the real-valued parameter t and the corresponding output is vector $\langle x(t), y(t), z(t) \rangle$. Such a function is called a *vector-valued function* because each real number input generates a vector output. More formally, we state the following definition.

Definition 9.6.1

A ***vector-valued function*** is a function whose input is a real parameter t and whose output is a vector that depends on t. The ***graph*** of a vector-valued function is the set of all terminal points of the output vectors with their initial points at the origin.

Parametric equations for a curve are equations of the form

$$x = x(t), \quad y = y(t), \quad \text{and} \quad z = z(t)$$

that describe the (x, y, z) coordinates of a point on a curve in $\mathbb{R}^3$.

Note particularly that every set of parametric equations determines a vector-valued function of the form

$$\mathbf{r}(t) = \langle x(t), y(t), z(t) \rangle,$$

and every vector-valued function defines a set of parametric equations for a curve. Moreover, we can consider vector-valued functions and parameterizations in $\mathbb{R}^2$, $\mathbb{R}^4$, or indeed a real space of any dimension. As a reminder, in Section 9.5, we determined the parametric equations of a line in space using a point and a direction vector. For a nonlinear example, the curve in Figure 9.6.1 has the parametric equations

$$x(t) = \cos(t), \quad y(t) = \sin(t), \quad \text{and} \quad z(t) = \cos(t)\sin(t).$$

Represented as a vector-valued function $\mathbf{r}$, the curve in Figure $9.6.1$ is the graph of

$$\mathbf{r}(t) = \langle \cos(t), \sin(t), \cos(t)\sin(t) \rangle.$$

Activity $9.6.2$

The same curve can be represented with different parameterizations. Use appropriate technology to plot the curves generated by the following vector-valued functions for values of t from 0 to 2π. Compare and contrast the graphs — explain how they are alike and how they are different.

1. $\mathbf{r}(t) = \langle \sin(t), \cos(t) \rangle$
2. $\mathbf{r}(t) = \langle \sin(2t), \cos(2t) \rangle$
3. $\mathbf{r}(t) = \langle \cos(t+\pi), \sin(t+\pi) \rangle$
4. $\mathbf{r}(t) = \langle \cos(t^2), \sin(t^2) \rangle$

The examples in Activity $9.6.2$ illustrate that a parameterization allows us to look not only at the graph, but at the direction and speed at which the graph is traversed as t changes. In the different parameterizations of the circle, we see that we can start at different points and move around the circle in either direction. The calculus of vector-valued functions — which we will begin to investigate in Section 9.7 — will enable us to precisely quantify the direction, speed, and acceleration of a particle moving along a curve in space. As such, describing curves parametrically will allow us to not only indicate the curve itself, but also to describe how motion occurs along the curve.

Using parametric equations to define vector-valued functions in two dimensions is much more versatile than just defining y as a function of x. In fact, if $y = f(x)$ is a function of x, then we can parameterize the graph of f by

$$\mathbf{r}(t) = \langle t, f(t) \rangle,$$

and thus every single-variable function may be described parametrically. In addition, as we saw in Preview Activity $9.6.1$ and Activity $9.6.2$, we can use vector-valued functions to represent curves in the plane that do not define y as a function of x (or x as a function of y). (As a side note: vector-valued functions make it easy to plot the inverse of a one-to-one function in two dimensions. To see how, if $y = f(x)$defines a one-to-one function, then we can parameterize this function by $\mathbf{r}(t) = \langle t, f(t) \rangle$. Since the inverse function just reverses the role of input and output, a parameterization for f^{-1} is $\langle f(t), t \rangle$.)

Activity $9.6.3$

Vector-valued functions can be used to generate many interesting curves. Graph each of the following using an appropriate technological tool, and then write one sentence for each function to describe the behavior of the resulting curve.

1. $\mathbf{r}(t) = \langle t\cos(t), t\sin(t) \rangle$
2. $\mathbf{r}(t) = \langle \sin(t)\cos(t), t\sin(t) \rangle$
3. $\mathbf{r}(t) = \langle \sin(5t), \sin(4t) \rangle$
4. $\mathbf{r}(t) = \langle t^2\sin(t)\cos(t), 0.9t\cos(t^2), \sin(t) \rangle$ (Note that this defines a curve in 3-space.)
5. Experiment with different formulas for $x(t)$ and $y(t)$ and ranges for t to see what other interesting curves you can generate. Share your best results with peers.

Recall from our earlier work that the traces and level curves of a function are themselves curves in space. Thus, we may determine parameterizations for them. For example, if $z = f(x, y) = \cos(x^2 + y^2)$, the $y = 1$ trace of the function is given by setting $y = 1$ and letting x be parameterized by the variable t; then, the trace is the curve whose parameterization is $\langle t, 1, \cos(t^2 + 1) \rangle$.

Activity $9.6.4$

Consider the paraboloid defined by $f(x, y) = x^2 + y^2$.

1. Find a parameterization for the $x = 2$ trace of f. What type of curve does this trace describe?
2. Find a parameterization for the $y = -1$ trace of f. What type of curve does this trace describe?

3. Find a parameterization for the level curve $f(x, y) = 25$. What type of curve does this trace describe?
4. How do your responses change to all three of the preceding questions if you instead consider the function g defined by $g(x, y) = x^2 - y^2$? (Hint for generating one of the parameterizations: $\sec^2(t) - \tan^2(t) = 1$.)

Summary

- A vector-valued function is a function whose input is a real parameter t and whose output is a vector that depends on t. The graph of a vector-valued function is the set of all terminal points of the output vectors with their initial points at the origin.
- Every vector-valued function provides a parameterization of a curve. In $\mathbb{R}^2$, a parameterization of a curve is a pair of equations $x = x(t)$ and $y = y(t)$ that describes the coordinates of a point (x, y) on the curve in terms of a parameter t. In $\mathbb{R}^3$, a parameterization of a curve is a set of three equations $x = x(t)$, $y = y(t)$, and $z = z(t)$ that describes the coordinates of a point (x, y, z) on the curve in terms of a parameter t.

This page titled 9.6: Vector-Valued Functions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 9.6: Vector-Valued Functions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

9.7: Derivatives and Integrals of Vector-Valued Functions

5Motivating Questions

- What do we mean by the derivative of a vector-valued function and how do we calculate it?
- What does the derivative of a vector-valued function measure?
- What do we mean by the integral of a vector-valued function and how do we compute it?
- How do we describe the motion of a projectile if the only force acting on the object is acceleration due to gravity?

A vector-valued function $\mathbf{r}$ determines a curve in space as the collection of terminal points of the vectors $\mathbf{r}(t)$. If the curve is smooth, it is natural to ask whether $\mathbf{r}(t)$ has a derivative. In the same way, our experiences with integrals in single-variable calculus prompt us to wonder what the integral of a vector-valued function might be and what it might tell us. We explore both of these questions in detail in this section.

For now, let's recall some important ideas from calculus I. Given a function s that measures the position of an object moving along an axis, its derivative, s', is defined by

$$s'(t) = \lim_{h \to 0} \frac{s(t+h) - s(t)}{h},$$

and measures the instantaneous rate of change of s with respect to time. In particular, for a fixed value $t = a$, $s'(a)$ measures the velocity of the moving object, as well as the slope of the tangent line to the curve $y = s(t)$ at the point $(a, s(a))$.

As we work with vector-valued functions, we will strive to update these ideas and perspectives into the context of curves in space and outputs that are vectors.

Preview Activity 9.7.1

Let $\mathbf{r}(t) = \cos(t)\mathbf{i} + \sin(2t)\mathbf{j}$ describe the path traveled by an object at time t.

1. Use appropriate technology to help you sketch the graph of the vector-valued function $\mathbf{r}$, and then locate and label the point on the graph when $t = \pi$.
2. Recall that for functions of a single variable, the derivative of a sum is the sum of the derivatives; that is, $\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$. With this idea in mind and viewing $\mathbf{i}$ and $\mathbf{j}$ as constant vectors, what do you expect the derivative of $\mathbf{r}$ to be? Write a proposed formula for $\mathbf{r}'(t)$.
3. Use your result from part (b) to compute $\mathbf{r}'(\pi)$. Sketch this vector $\mathbf{r}'(\pi)$ as emanating from the point on the graph of $\mathbf{r}$ when $t = \pi$, and explain what you think $\mathbf{r}'(\pi)$ tells us about the object's motion.

The Derivative

In single variable calculus, we define the derivative, f', of a given function f by

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

provided the limit exists. At a given value of a, $f'(a)$ measures the instantaneous rate of change of f, and also tells us the slope of the tangent line to the curve $y = f(x)$ at the point $(a, f(a))$. The definition of the derivative extends naturally to vector-valued functions and curves in space.

Definition 9.7.1: Derivative of a Vector-valued Function

The **derivative of a vector-valued function** $\mathbf{r}$ is defined to be

$$\mathbf{r}'(t) = \lim_{h \to 0} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{h}$$

for those values of t at which the limit exists. We also use the notation $\frac{d\mathbf{r}}{dt}$ and $\frac{d}{dt}[\mathbf{r}(t)]$ for $\mathbf{r}'(t)$.

Activity 9.7.2

Let's investigate how we can interpret the derivative $\mathbf{r}'(t)$. Let $\mathbf{r}$ be the vector-valued function whose graph is shown in Figure 9.7.2, and let h be a scalar that represents a small change in time. The vector $\mathbf{r}(t)$ is the blue vector in Figure 9.7.1 and $\mathbf{r}(t+h)$ is the green vector.

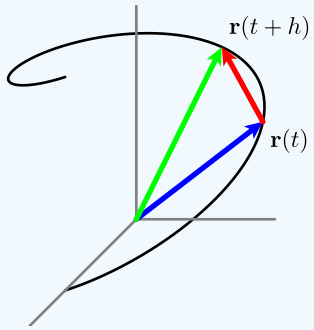


Figure 9.7.1. A single difference quotient.

1. Is the quantity $\mathbf{r}(t+h) - \mathbf{r}(t)$ a vector or a scalar? Identify this object in Figure 9.7.1.
2. Is $\frac{\mathbf{r}(t+h)-\mathbf{r}(t)}{h}$ a vector or a scalar? Sketch a representative vector $\frac{\mathbf{r}(t+h)-\mathbf{r}(t)}{h}$ with $h < 1$ in Figure 9.7.2.
3. Think of $\mathbf{r}(t)$ as providing the position of an object moving along the curve these vectors trace out. What do you think that the vector $\frac{\mathbf{r}(t+h)-\mathbf{r}(t)}{h}$ measures? Why? (Hint: You might think analogously about difference quotients such as $\frac{f(x+h)-f(x)}{h}$ or $\frac{s(t+h)-s(t)}{h}$ from calculus I.)
4. Figure 9.7.2 presents three snapshots of the vectors $\frac{\mathbf{r}(t+h)-\mathbf{r}(t)}{h}$ as we let $h \to 0$. Write 2-3 sentences to describe key attributes of the vector

$$\lim_{h \to 0} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{h}.$$

(Hint: Compare to limits such as $\lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$ or $\lim_{h \to 0} \frac{s(t+h)-s(t)}{h}$ from calculus I, keeping in mind that in three dimensions there is no general concept of slope.)

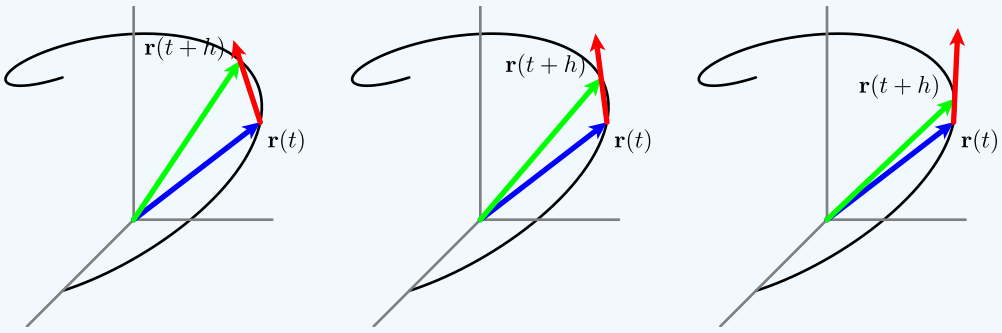


Figure 9.7.2. Snapshots of several difference quotients.

As Activity 9.7.2 indicates, if $\mathbf{r}(t)$ determines the position of an object at time t, then $\frac{\mathbf{r}(t+h)-\mathbf{r}(t)}{h}$ represents the average rate of

change in the position of the object over the interval $[t, t+h]$, which is also the average velocity of the object on this interval. Thus, the derivative

$$\mathbf{r}'(t) = \lim_{h\to 0} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{h}$$

is the instantaneous rate of change of $\mathbf{r}(t)$ at time t (for those values of t for which the limit exists), so $\mathbf{r}'(t) = \mathbf{v}(t)$ is the instantaneous velocity of the object at time t. Furthermore, we can interpret the derivative $\mathbf{r}'(t)$ as the direction vector of the line tangent to the graph of $\mathbf{r}$ at the value t.

Similarly,

$$\mathbf{v}'(t) = \mathbf{r}''(t) = \lim_{h\to 0} \frac{\mathbf{v}(t+h) - \mathbf{v}(t)}{h}$$

is the instantaneous rate of change of the velocity of the object at time t, for those values of t for which the limits exists, and thus $\mathbf{v}'(t) = \mathbf{a}(t)$ is the acceleration of the moving object.

Note well: Both the velocity and acceleration are *vector quantities*: they have magnitude and direction. By contrast, the magnitude of the velocity vector, $|\mathbf{v}(t)|$, which is the *speed* of the object at time t, is a scalar quantity.

Computing Derivatives

As we learned in single variable calculus, computing derivatives from the definition is often difficult. Fortunately, properties of the limit make it straightforward to calculate the derivative of a vector-valued function similar to how we developed shortcut differentiation rules in calculus I. To see why, recall that the limit of a sum is the sum of the limits, and that we can remove constant factors from limits. Thus, as we observed in a particular example in Preview Activity 9.7.1, if $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$, it follows that

$$\begin{aligned}
\mathbf{r}'(t) &= \lim_{h\to 0} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{h} \\
&= \lim_{h\to 0} \frac{[x(t+h) - x(t)]\mathbf{i} + [y(t+h) - y(t)]\mathbf{j} + [z(t+h) - z(t)]\mathbf{k}}{h} \\
&= \left(\lim_{h\to 0} \frac{x(t+h) - x(t)}{h}\right)\mathbf{i} + \left(\lim_{h\to 0} \frac{y(t+h) - y(t)}{h}\right)\mathbf{j} \\
&\quad + \left(\lim_{h\to 0} \frac{z(t+h) - z(t)}{h}\right)\mathbf{k} \\
&= x'(t)\mathbf{i} + y'(t)\mathbf{j} + z'(t)\mathbf{k}.
\end{aligned}$$

Thus, we can calculate the derivative of a vector-valued function by simply differentiating its components.

The derivative of a vector-valued function

If $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$, then

$$\frac{d}{dt}\mathbf{r}(t) = x'(t)\mathbf{i} + y'(t)\mathbf{j} + z'(t)\mathbf{k}$$

for those values of t at which x, y, and z are differentiable.

Activity 9.7.3

For each of the following vector-valued functions, find $\mathbf{r}'(t)$.

1. $\mathbf{r}(t) = \langle \cos(t), t\sin(t), \ln(t)\rangle$.

2. $\mathbf{r}(t) = \langle t^2 + 3t, e^{-2t}, \frac{t}{t^2+1} \rangle$.
3. $\mathbf{r}(t) = \langle \tan(t), \cos(t^2), te^{-t} \rangle$.
4. $\mathbf{r}(t) = \langle \sqrt{t^4+4}, \sin(3t), \cos(4t) \rangle$.

In first-semester calculus, we developed several important differentiation rules, including the constant multiple, product, quotient, and chain rules. For instance, recall that we formally state the product rule as

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

There are several analogous rules for vector-valued functions, including a product rule for scalar functions and vector-valued functions. These rules, which are easily verified, are summarized as follows.

Properties of derivatives of vector-valued functions

Let f be a differentiable real-valued function of a real variable t and let $\mathbf{r}$ and $\mathbf{s}$ be differentiable vector-valued functions of the real variable t. Then

1. $\frac{d}{dt}[\mathbf{r}(t) + \mathbf{s}(t)] = \mathbf{r}'(t) + \mathbf{s}'(t)$
2. $\frac{d}{dt}[f(t)\mathbf{r}(t)] = f(t)\mathbf{r}'(t) + f'(t)\mathbf{r}(t)$
3. $\frac{d}{dt}[\mathbf{r}(t) \cdot \mathbf{s}(t)] = \mathbf{r}'(t) \cdot \mathbf{s}(t) + \mathbf{r}(t) \cdot \mathbf{s}'(t)$
4. $\frac{d}{dt}[\mathbf{r}(t) \times \mathbf{s}(t)] = \mathbf{r}'(t) \times \mathbf{s}(t) + \mathbf{r}(t) \times \mathbf{s}'(t)$
5. $\frac{d}{dt}[\mathbf{r}(f(t))] = f'(t)\mathbf{r}'(f(t))$.

Note well. When applying these properties, use care to interpret the quantities involved as either scalars or vectors. For example, $\mathbf{r}(t) \cdot \mathbf{s}(t)$ defines a scalar function because we have taken the dot product of two vector-valued functions. However, $\mathbf{r}(t) \times \mathbf{s}(t)$ defines a vector-valued function since we have taken the cross product of two vector-valued functions.

Activity 9.7.4

The left side of Figure 9.7.3 shows the curve described by the vector-valued function $\mathbf{r}$ defined by

$$\mathbf{r}(t) = \left\langle 2t - \frac{1}{2}t^2 + 1, t - 1 \right\rangle.$$

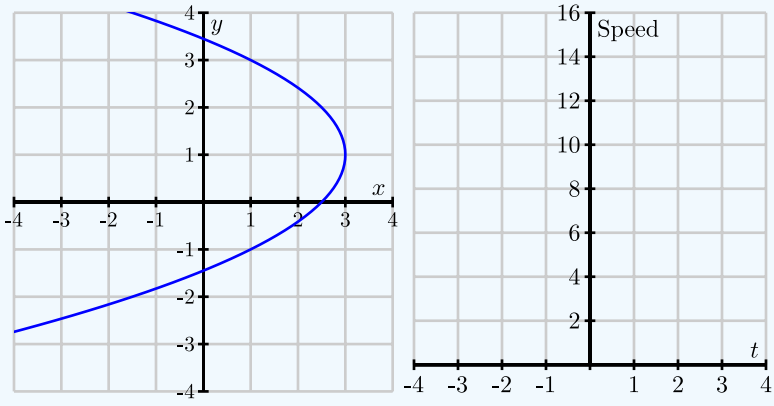


Figure 9.7.3. The curve $\mathbf{r}(t) = \left\langle 2t - \frac{1}{2}t^2 + 1, t - 1 \right\rangle$ and its speed.

1. Find the object's velocity $\mathbf{v}(t)$.
2. Find the object's acceleration $\mathbf{a}(t)$.

3. Indicate on the left of Figure 9.7.4 the object's position, velocity and acceleration at the times $t = 0, 2, 4$. Draw the velocity and acceleration vectors with their tails placed at the object's position.
4. Recall that the speed is $|\mathbf{v}| = \sqrt{\mathbf{v} \cdot \mathbf{v}}$. Find the object's speed and graph it as a function of time t on the right of Figure 9.7.4. When is the object's speed the slowest? When is the speed increasing? When is it decreasing?
5. What seems to be true about the angle between $\mathbf{v}$ and $\mathbf{a}$ when the speed is at a minimum? What is the angle between $\mathbf{v}$ and $\mathbf{a}$ when the speed is increasing? when the speed is decreasing?
6. Since the square root is an increasing function, we see that the speed increases precisely when $\mathbf{v} \cdot \mathbf{v}$ is increasing. Use the product rule for the dot product to express $\frac{d}{dt}(\mathbf{v} \cdot \mathbf{v})$ in terms of the velocity $\mathbf{v}$ and acceleration $\mathbf{a}$. Use this to explain why the speed is increasing when $\mathbf{v} \cdot \mathbf{a} > 0$ and decreasing when $\mathbf{v} \cdot \mathbf{a} < 0$. Compare this to part (d).
7. Show that the speed's rate of change is

$$\frac{d}{dt}|\mathbf{v}(t)| = comp_{\mathbf{v}}\mathbf{a}.$$

Tangent Lines

One of the most important ideas in first-semester calculus is that a differentiable function is *locally linear*: that is, when viewed up close, the curve generated by a differentiable function looks very much like a line. Indeed, when we zoom in sufficiently far on a particular point, the curve looks indistinguishable from its tangent line.

In the same way, we expect that a smooth curve in 3-space will be locally linear. In the following activity, we investigate how to find the tangent line to such a curve. Recall from our work in Section 9.5 that the vector equation of a line that passes through the point at the tip of the vector $\mathbf{L}_0 = \langle x_0, y_0, z_0 \rangle$ in the direction of the vector $\mathbf{u} = \langle a, b, c \rangle$ can be written as

$$\mathbf{L}(t) = \mathbf{L}_0 + t\mathbf{u}.$$

In parametric form, the line $\mathbf{L}$ is given by

$$x(t) = x_0 + at, \quad y(t) = y_0 + bt, \quad z(t) = z_0 + ct.$$

Activity 9.7.5

Let

$$\mathbf{r}(t) = \cos(t)\mathbf{i} - \sin(t)\mathbf{j} + t\mathbf{k}.$$

Sketch the curve using some appropriate tool.

1. Determine the coordinates of the point on the curve traced out by $\mathbf{r}(t)$ when $t = \pi$.
2. Find a direction vector for the line tangent to the graph of $\mathbf{r}$ at the point where $t = \pi$.
3. Find the parametric equations of the line tangent to the graph of $\mathbf{r}$ when $t = \pi$.
4. Sketch a plot of the curve $\mathbf{r}(t)$ and its tangent line near the point where $t = \pi$. In addition, include a sketch of $\mathbf{r}'(\pi)$. What is the important role of $\mathbf{r}'(\pi)$ in this activity?

We see that our work in Activity 9.7.5 can be generalized. Given a differentiable vector-valued function $\mathbf{r}$, the tangent line to the curve at the input value a is given by

$$\mathbf{L}(t) = \mathbf{r}(a) + t\mathbf{r}'(a). \tag{9.7.1}$$

Here we see that because the tangent line is determined entirely by a given point and direction, the point is provided by the function $\mathbf{r}$, evaluated at $t = a$, while the direction is provided by the derivative, $\mathbf{r}'$, again evaluated at $t = a$. Note how analogous the formula for $\mathbf{L}(t)$ is to the tangent line approximation from single-variable calculus: in that context, for a given function $y = f(x)$ at a value $x = a$, we found that the tangent line can be expressed by the linear function $y = L(x)$ whose formula is

$$L(x) = f(a) + f'(a)(x - a).$$

Equation (9.7.1) for the tangent line $\mathbf{L}(t)$ to the vector-valued function $\mathbf{r}(t)$ is nearly identical. Indeed, because there are multiple parameterizations for a single line, it is even possible to write the parameterization as

$$\mathbf{L}(t) = \mathbf{r}(a) + (t-a)\mathbf{r}'(a). \tag{9.7.2}$$

(For example, in Equation $(9.7.1)$, $\mathbf{L}(0) = \mathbf{r}(a)$, so the line's parameterization "starts" at $t = 0$. When we write the parameterization in the form of Equation $(9.7.2)$, $\mathbf{L}(a) = \mathbf{r}(a)$, so the line's parameterization "starts" at $t = a$.)

As we will learn more in Chapter 10, a smooth surface in 3-space is also locally linear. That means that the surface will look like a plane, which we call its *tangent plane*, as we zoom in on the graph. It is possible to use tangent lines to traces of the surface to generate a formula for the tangent plane; see Exercise $9.7.7.15$ at the end of this section for more details.

Integrating a Vector-Valued Function

Recall from single variable calculus that an antiderivative of a function f of the independent variable x is a function F that satisfies $F'(x) = f(x)$. We then defined the indefinite integral $\int f(x)\ dx$ to be the general antiderivative of f. Recall that the general antiderivative includes an added constant C in order to indicate that the general antiderivative is in fact an entire family of functions. We can do the similar work with vector-valued functions.

Definition $9.7.2$

An *antiderivative* of a vector-valued function $\mathbf{r}$ is a vector-valued function $\mathbf{R}$ such that

$$\mathbf{R}'(t) = \mathbf{r}(t).$$

The *indefinite integral* $\int \mathbf{r}(t)\ dt$ of a vector-valued function $\mathbf{r}$ is the general antiderivative of $\mathbf{r}$ and represents the collection of all antiderivatives of $\mathbf{r}$.

The same reasoning that allows us to differentiate a vector-valued function componentwise applies to integrating as well. Recall that the integral of a sum is the sum of the integrals and also that we can remove constant factors from integrals. So, given $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$, it follows that we can integrate componentwise. Expressed more formally,

Integrating a vector-valued function

If $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$, then

$$\int \mathbf{r}(t)\ dt = \left(\int x(t)\ dt\right)\mathbf{i} + \left(\int y(t)\ dt\right)\mathbf{j} + \left(\int z(t)\ dt\right)\mathbf{k}.$$

In light of being able to integrate and differentiate componentwise with vector-valued functions, we can solve many problems that are analogous to those we encountered in single-variable calculus. For instance, recall problems where we were given an object moving along an axis with velocity function v and an initial position $s(0)$. In that context, we were able to differentiate v in order to find acceleration, and integrate v and use the initial condition in order to find the position function s. In the following activity, we explore similar ideas with vector-valued functions.

Activity $9.7.6$

Suppose a moving object in space has its velocity given by

$$\mathbf{v}(t) = (-2\sin(2t))\mathbf{i} + (2\cos(t))\mathbf{j} + \left(1 - \frac{1}{1+t}\right)\mathbf{k}.$$

A graph of the position of the object for times t in $[-0.5, 3]$ is shown in Figure $9.7.6$. Suppose further that the object is at the point $(1.5, -1, 0)$ at time $t = 0$.

1. Determine $\mathbf{a}(t)$, the acceleration of the object at time t.
2. Determine $\mathbf{r}(t)$, position of the object at time t.
3. Compute and sketch the position, velocity, and acceleration vectors of the object at time $t = 1$, using Figure $9.7.4$.

4. Finally, determine the vector equation for the tangent line, $\mathbf{L}(t)$, that is tangent to the position curve at $t = 1$.

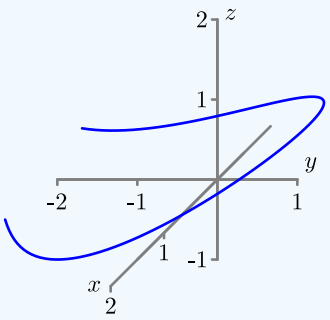


Figure 9.7.4. The position graph for the function in Activity 9.7.6.

Projectile Motion

Any time that an object is launched into the air with a given velocity and launch angle, the path the object travels is determined almost exclusively by the force of gravity. Whether in sports such as archery or shotput, in military applications with artillery, or in important fields like firefighting, it is important to be able to know when and where a launched projectile will land. We can use our knowledge of vector-valued functions in order to completely determine the path traveled by an object that is launched from a given position at a given angle from the horizontal with a given initial velocity.



Figure 9.7.5. Projectile motion.

Assume we fire a projectile from a launcher and the only force acting on the fired object is the force of gravity pulling down on the object. That is, we assume no effect due to spin, wind, or air resistance. With these assumptions, the motion of the object will be planar, so we can also assume that the motion occurs in two-dimensional space. Suppose we launch the object from an initial position (x_0, y_0) at an angle θ with the positive x-axis as illustrated in Figure 9.7.5, and that we fire the object with an initial speed of $v_0 = |\mathbf{v}(0)|$, where $\mathbf{v}(t)$ is the velocity vector of the object at time t. Assume g is the positive constant acceleration force due to gravity, which acts to pull the fired object toward the ground (in the negative y direction). Note particularly that there is no external force acting on the object to move it in the x direction.

We first observe that since gravity only acts in the downward direction and that the acceleration due to gravity is constant, the acceleration vector is $\langle 0, -g \rangle$. That is, $\mathbf{a}(t) = \langle 0, -g \rangle$. We may use this fact about acceleration, together with the initial position and initial velocity in order to fully determine the position $\mathbf{r}(t)$ of the object at time t. In Exercise 9.7.5, you can work through the details to show that the following general formula holds.

The motion of a projectile

If an object is launched from a point (x_0, y_0) with initial velocity v_0 at an angle θ with the horizontal, then the position of the object at time t is given by

$$\mathbf{r}(t) = \left\langle v_0 \cos(\theta) t + x_0, -\frac{g}{2} t^2 + v_0 \sin(\theta) t + y_0 \right\rangle.$$

This assumes that the only force acting on the object is the acceleration g due to gravity.

Summary

- If $\mathbf{r}$ is a vector-valued function, then the derivative of $\mathbf{r}$ is defined by

$$\mathbf{r}'(t) = \lim_{h \to 0} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{h}$$

for those values of t at which the limit exists, and is computed componentwise by the formula

$$\mathbf{r}'(t) = x'(t)\mathbf{i} + y'(t)\mathbf{j} + z'(t)\mathbf{k}$$

for those values of t at which x, y, and z are differentiable, where $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$.

- The derivative $\mathbf{r}'(t)$ of the vector-valued function $\mathbf{r}$ tells us the instantaneous rate of change of $\mathbf{r}$ with respect to time, t, which can be interpreted as a direction vector for the line tangent to the graph of $\mathbf{r}$ at the point $\mathbf{r}(t)$, or also as the instantaneous velocity of an object traveling along the graph defined by $\mathbf{r}(t)$ at time t.
- An antiderivative of $\mathbf{r}$ is a vector-valued function $\mathbf{R}$ such that $\mathbf{R}'(t) = \mathbf{r}(t)$. The indefinite integral $\int \mathbf{r}(t)\ dt$ of a vector-valued function $\mathbf{r}$ is the general antiderivative of $\mathbf{r}$ (which is a collection of all of the antiderivatives of $\mathbf{r}$, with any two antiderivatives differing by at most a constant vector). Moreover, if $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$, then

$$\int \mathbf{r}(t)\ dt = \left(\int x(t)\ dt\right)\mathbf{i} + \left(\int y(t)\ dt\right)\mathbf{j} + \left(\int z(t)\ dt\right)\mathbf{k}.$$

- If an object is launched from a point (x_0, y_0) with initial velocity v_0 at an angle θ with the horizontal, then the position of the object at time t is given by

$$\mathbf{r}(t) = \left\langle v_0 \cos(\theta)t + x_0, -\frac{g}{2}t^2 + v_0 \sin(\theta)t + y_0 \right\rangle,$$

provided that that the only force acting on the object is the acceleration g due to gravity.

This page titled 9.7: Derivatives and Integrals of Vector-Valued Functions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 9.7: Derivatives and Integrals of Vector-Valued Functions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

9.8: Arc Length and Curvature

Motivating Questions

- How can a definite integral be used to measure the length of a curve in 2- or 3-space?
- Why is arc length useful as a parameter?
- What is the curvature of a curve?

Given a space curve, there are two natural geometric questions one might ask: how long is the curve and how much does it bend? In this section, we answer both questions by developing techniques for measuring the length of a space curve as well as its curvature.

Preview Activity 9.8.1

In earlier investigations, we have used integration to calculate quantities such as area, volume, mass, and work. We are now interested in determining the length of a space curve.

Consider the smooth curve in 3-space defined by the vector-valued function $\mathbf{r}$, where

$$\mathbf{r}(t) = \langle x(t), y(t), z(t) \rangle = \langle \cos(t), \sin(t), t \rangle$$

for t in the interval $[0, 2\pi]$. Pictures of the graph of $\mathbf{r}$ are shown in Figure 9.8.1. We will use the integration process to calculate the length of this curve. In this situation we partition the interval $[0, 2\pi]$ into n subintervals of equal length and let $0 = t_0 < t_1 < t_2 < \cdots < t_n = b$ be the endpoints of the subintervals. We then approximate the length of the curve on each subinterval with some related quantity that we can compute. In this case, we approximate the length of the curve on each subinterval with the length of the segment connecting the endpoints. Figure 9.8.1 illustrates the process in three different instances using increasing values of n.

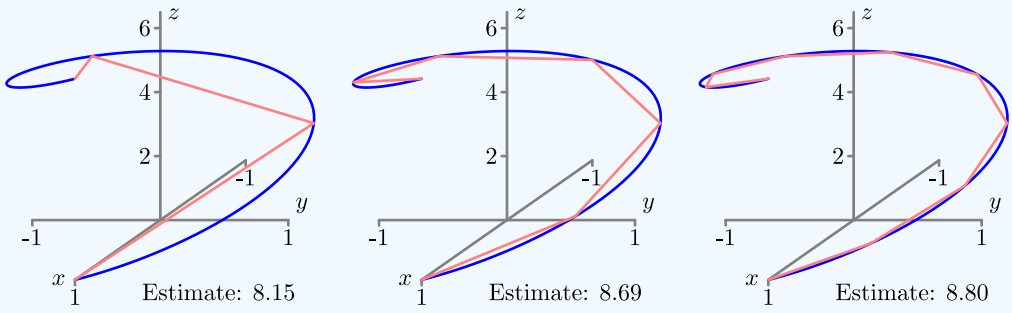


Figure 9.8.1. Approximating the length of the curve with $n = 3$, $n = 6$, and $n = 9$.

1. Write a formula for the length of the line segment that connects the endpoints of the curve on the ith subinterval $[t_{i-1}, t_i]$. (This length is our approximation of the length of the curve on this interval.)
2. Use your formula in part (a) to write a sum that adds all of the approximations to the lengths on each subinterval.
3. What do we need to do with the sum in part (b) in order to obtain the exact value of the length of the graph of $\mathbf{r}(t)$ on the interval $[0, 2\pi]$?

9.8.1 Arc Length

Consider a smooth curve in 3-space that is parametrically described by the vector-valued function $\mathbf{r}$ defined by $\mathbf{r}(t) = \langle x(t), y(t), z(t) \rangle$. Preview Activity 9.8.1 shows that to approximate the length of the curve defined by $\mathbf{r}(t)$ as the values of t run over an interval $[a, b]$, we partition the interval $[a, b]$ into n subintervals of equal length Δt, with $a = t_0 < t_1 < \cdots < t_n = b$

as the endpoints of the subintervals. On each subinterval, we approximate the length of the curve by the length of the line segment connecting the endpoints. The points on the curve corresponding to $t = t_{i-1}$ and $t = t_i$ are $(x(t_{i-1}), y(t_{i-1}), z(t_{i-1}))$ and $(x(t_i), y(t_i), z(t_i))$, respectively, so the length of the line segment connecting these points is

$$\sqrt{(x(t_i) - x(t_{i-1}))^2 + (y(t_i) - y(t_{i-1}))^2 + (z(t_i) - z(t_{i-1}))^2}.$$

Now we add all of these approximations together to obtain an approximation to the length L of the curve:

$$L \approx \sum_{i=1}^{n} \sqrt{(x(t_i) - x(t_{i-1}))^2 + (y(t_i) - y(t_{i-1}))^2 + (z(t_i) - z(t_{i-1}))^2}.$$

We now want to take the limit of this sum as n goes to infinity, but in its present form it might be difficult to see how. We first introduce Δt by multiplying by $\frac{\Delta t}{\Delta t}$, and see that

$$\begin{aligned} L &\approx \sum_{i=1}^{n} \sqrt{(x(t_i) - x(t_{i-1}))^2 + (y(t_i) - y(t_{i-1}))^2 + (z(t_i) - z(t_{i-1}))^2} \\ &= \sum_{i=1}^{n} \sqrt{(x(t_i) - x(t_{i-1}))^2 + (y(t_i) - y(t_{i-1}))^2 + (z(t_i) - z(t_{i-1}))^2} \frac{\Delta t}{\Delta t} \\ &= \sum_{i=1}^{n} \sqrt{(x(t_i) - x(t_{i-1}))^2 + (y(t_i) - y(t_{i-1}))^2 + (z(t_i) - z(t_{i-1}))^2} \frac{\Delta t}{\sqrt{(\Delta t)^2}} \end{aligned}$$

To get the difference quotients under the radical, we use properties of the square root function to see further that

$$\begin{aligned} L &\approx \sum_{i=1}^{n} \sqrt{[(x(t_i) - x(t_{i-1}))^2 + (y(t_i) - y(t_{i-1}))^2 + (z(t_i) - z(t_{i-1})^2] \frac{1}{(\Delta t)^2}} \Delta t \\ &= \sum_{i=1}^{n} \sqrt{\left(\frac{x(t_i) - x(t_{i-1})}{\Delta t}\right)^2 + \left(\frac{y(t_i) - y(t_{i-1})}{\Delta t}\right)^2 + \left(\frac{z(t_i) - z(t_{i-1})}{\Delta t}\right)^2} \Delta t. \end{aligned}$$

Recall that as $n \to \infty$ we also have $\Delta t \to 0$. Since

$$\begin{aligned} x'(t) &= \lim_{\Delta t \to 0} \frac{x(t_i) - x(t_{i-1})}{\Delta t}, \\ y'(t) &= \lim_{\Delta t \to 0} \frac{y(t_i) - y(t_{i-1})}{\Delta t}, \quad \text{and} \\ z'(t) &\lim_{\Delta t \to 0} \frac{z(t_i) - z(t_{i-1})}{\Delta t}, \end{aligned}$$

we see that

$$\lim_{n \to \infty} \sum_{i=1}^{n} \sqrt{\left(\frac{x(t_i) - x(t_{i-1})}{\Delta t}\right)^2 + \left(\frac{y(t_i) - y(t_{i-1})}{\Delta t}\right)^2 + \left(\frac{z(t_i) - z(t_{i-1})}{\Delta t}\right)^2} \Delta t$$

is equal to

$$\int_a^b \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2}\, dt.$$

Noting further that

$$|\mathbf{r}'(t)| = \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2},$$

we can rewrite our arclength formula in a more succinct form as follows.

The length of a curve

If $\mathbf{r}(t)$ defines a smooth curve C on an interval $[a, b]$, then the *length* L of C is given by

$$L = \int_a^b |\mathbf{r}'(t)|\, dt. \tag{9.8.1}$$

Note that formula (9.8.1) applies to curves in any dimensional space. Moreover, this formula has a natural interpretation: if $\mathbf{r}(t)$ records the position of a moving object, then $\mathbf{r}'(t)$ is the object's velocity and $|\mathbf{r}'(t)|$ its speed. Formula (9.8.1) says that we simply integrate the speed of an object traveling over the curve to find the distance traveled by the object, which is the same as the length of the curve, just as in one-variable calculus.

Activity 9.8.2

Here we calculate the arc length of two familiar curves.

1. Use Equation (9.8.1) to calculate the circumference of a circle of radius r.
2. Find the exact length of the spiral defined by $\mathbf{r}(t) = \langle \cos(t), \sin(t), t \rangle$ on the interval $[0, 2\pi]$.

We can adapt the arc length formula to curves in 2-space that define y as a function of x as the following activity shows.

Activity 9.8.3

Let $y = f(x)$ define a smooth curve in 2-space. Parameterize this curve and use Equation (9.8.1) to show that the length of the curve defined by f on an interval $[a, b]$ is

$$\int_a^b \sqrt{1 + [f'(t)]^2}\, dt.$$

9.8.2 Parameterizing With Respect To Arc Length

In addition to helping us to find the length of space curves, the expression for the length of a curve enables us to find a natural parametrization of space curves in terms of arc length, as we now explain.

Shown below in Figure 9.8.2 is a portion of the parabola $y = x^2/2$. Of course, this space curve may be parametrized by the vector-valued function $\mathbf{r}$ defined by $\mathbf{r}(t) = \langle t, t^2/2 \rangle$ as shown on the left, where we see the location at a few different times t. Notice that the points are not equally spaced on the curve.

A more natural parameter describing the points along the space curve is the distance traveled s as we move along the parabola starting at the origin. For instance, the right side of Figure 9.8.2 shows the points corresponding to various values of s. We call this an *arc length parametrization*.

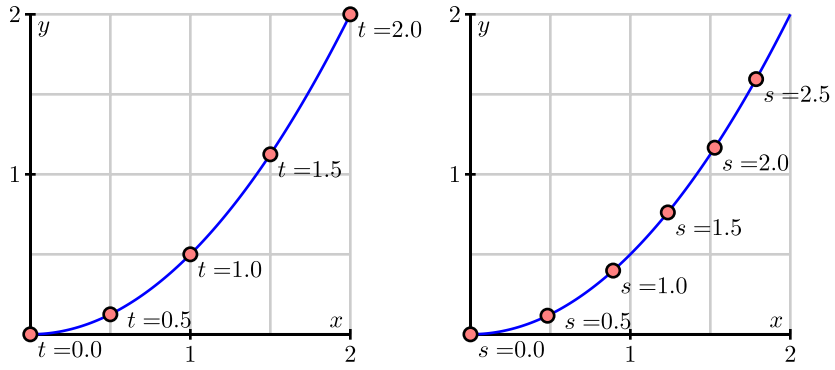


Figure 9.8.2. The parametrization $\mathbf{r}(t)$ (left) and a reparametrization by arc length.

To see that this is a more natural parametrization, consider an interstate highway cutting across a state. One way to parametrize the curve defined by the highway is to drive along the highway and record our position at every time, thus creating a function $\mathbf{r}$. If we encounter an accident or road construction, however, this parametrization might not be at all relevant to another person driving the same highway. An arc length parametrization, however, is like using the mile markers on the side of road to specify our position on the highway. If we know how far we've traveled along the highway, we know exactly where we are.

If we begin with a parametrization of a space curve, we can modify it to find an arc length parametrization, as we now describe. Suppose that the curve is parametrized by the vector-valued function $\mathbf{r} = \mathbf{r}(t)$ where t is in the interval $[a, b]$. We define the parameter s through the function

$$s = L(t) = \int_a^t \sqrt{(x'(w))^2 + (y'(w))^2 + (z'(w))^2}\, dw,$$

which measures the length along the curve from $\mathbf{r}(a)$ to $\mathbf{r}(t)$.

The Fundamental Theorem of Calculus shows us that

$$\frac{ds}{dt} = L'(t) = \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2} = |\mathbf{r}'(t)| \tag{9.8.2}$$

and so

$$L(t) = \int_a^t \left| \frac{d}{dw}\mathbf{r}(w) \right| dw.$$

If we assume that $\mathbf{r}'(t)$ is never $\mathbf{0}$, then $L'(t) > 0$ for all t and $s = L(t)$ is always increasing. This should seem reasonable: unless we stop, the distance traveled along the curve increases as we move along the curve.

Since $s = L(t)$ is an increasing function, it is invertible, which means we may view the time t as a function of the distance traveled; that is, we have the relationship $t = L^{-1}(s)$. We then obtain the arc length parametrization by composing $\mathbf{r}(t)$ with $t = L^{-1}(s)$ to obtain $\mathbf{r}(s)$. Let's illustrate this with an example.

Example 9.8.3

Consider a circle of radius 5 in 2-space centered at the origin. We know that we can parameterize this circle as

$$\mathbf{r}(t) = \langle 5\cos(t), 5\sin(t) \rangle,$$

where t runs from 0 to 2π. We see that $\mathbf{r}'(t) = \langle -5\sin(t), 5\cos(t) \rangle$, and hence $|\mathbf{r}'(t)| = 5$. It then follows that

$$s = L(t) = \int_0^t |\mathbf{r}'(w)|\, dw = \int_0^t 5\, dw = 5t.$$

Since $s = L(t) = 5t$, we may solve for t in terms of s to obtain $t(s) = L^{-1}(s) = s/5$. We then find the arc length parametrization by composing

$$\mathbf{r}(t(s)) = \mathbf{r}(L^{-1}(s)) = \left\langle 5\cos\left(\frac{s}{5}\right), 5\sin\left(\frac{s}{5}\right)\right\rangle.$$

More generally, for a circle of radius a centered at the origin, a similar computation shows that

$$\left\langle a\cos\left(\frac{s}{a}\right), a\sin\left(\frac{s}{a}\right)\right\rangle \tag{9.8.3}$$

is an arc length parametrization.

Notice that equation (9.8.2) shows that

$$\frac{d\mathbf{r}}{dt} = \frac{d\mathbf{r}}{ds}\frac{ds}{dt} = \frac{d\mathbf{r}}{ds}|\mathbf{r}'(t)|,$$

so

$$\left|\frac{d\mathbf{r}}{ds}\right| = \left|\frac{1}{|\mathbf{r}'(t)|}\frac{d\mathbf{r}}{dt}\right| = 1,$$

which means that we move along the curve with unit speed when we parameterize by arc length. This is clearly seen in Example 9.8.3 where $|\mathbf{r}'(s)| = 1$. It follows that the parameter s is the distance traveled along the curve, as shown by:

$$L(s) = \int_0^s \left|\frac{d}{ds}\mathbf{r}(w)\right|\, dw = \int_0^s 1\, dw = s.$$

Activity 9.8.4

In this activity we parameterize a line in 2-space in terms of arc length. Consider the line with parametric equations

$$x(t) = x_0 + at \quad \text{and} \quad y(t) = y_0 + bt.$$

1. To write t in terms of s, evaluate the integral
$$s = L(t) = \int_0^t \sqrt{(x'(w))^2 + (y'(w))^2}\, dw$$
to determine the length of the line from time 0 to time t.
2. Use the formula from (a) for s in terms of t to write t in terms of s. Then explain why a parameterization of the line in terms of arc length is
$$x(s) = x_0 + \frac{a}{\sqrt{a^2+b^2}}s \quad \text{and} \quad y(s) = y_0 + \frac{b}{\sqrt{a^2+b^2}}s. \tag{9.8.4}$$

A little more complicated example is the following.

Example 9.8.4

Let us parameterize the curve defined by

$$\mathbf{r}(t) = \left\langle t^2, \frac{8}{3}t^{3/2}, 4t\right\rangle$$

for $t \geq 0$ in terms of arc length. To write t in terms of s we find s in terms of t:

$$\begin{aligned} s(t) &= \int_0^t \sqrt{(x'(w))^2 + (y'(w))^2 + (z'(w))^2}\, dw \\ &= \int_0^t \sqrt{(2w)^2 + (4w^{1/2})^2 + (4)^2}\, dw \\ &= \int_0^t \sqrt{4w^2 + 16w + 16}\, dw \\ &= 2\int_0^t \sqrt{(w+2)^2}\, dw \\ &= 2\int_0^t w + 2\, dw \\ &= \left(w^2 + 4w\right)\Big|_0^t \\ &= t^2 + 4t. \end{aligned}$$

Since $t \geq 0$, we can solve the equation $s = t^2 + 4t$ (or $t^2 + 4t - s = 0$) for t to obtain $t = \frac{-4+\sqrt{16+4s}}{2} = -2 + \sqrt{4+s}$. So we can parameterize our curve in terms of arc length by

$$\mathbf{r}(s) = \left\langle \left(-2+\sqrt{4+s}\right)^2, \frac{8}{3}\left(-2+\sqrt{4+s}\right)^{3/2}, 4\left(-2+\sqrt{4+s}\right) \right\rangle.$$

These examples illustrate a general method. Of course, evaluating an arc length integral and finding a formula for the inverse of a function can be difficult, so while this process is theoretically possible, it is not always practical to parameterize a curve in terms of arc length. However, we can guarantee that such a parameterization exists, and this observation plays an important role in the next section.

9.8.3 Curvature

For a smooth space curve, the *curvature* measures how fast the curve is bending or changing direction at a given point. For example, we expect that a line should have zero curvature everywhere, while a circle (which is bending the same at every point) should have constant curvature. Circles with larger radii should have smaller curvatures.

To measure the curvature, we first need to describe the direction of the curve at a point. We may do this using a continuously varying tangent vector to the curve, as shown at left in Figure 9.8.5. The direction of the curve is then determined by the angle ϕ each tangent vector makes with a horizontal vector, as shown at right in Figure 9.8.5.

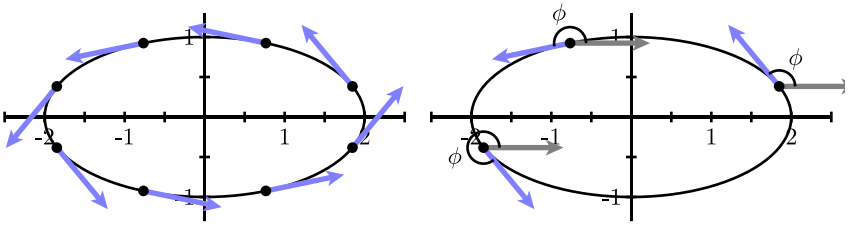


Figure 9.8.5. Left: Tangent vectors to an ellipse. Right: Angles of tangent vectors.

Informally speaking, the curvature will be the rate at which the angle ϕ is changing as we move along the curve. Of course, this rate of change will depend on how we move along the curve; if we move with a greater speed along the curve, then ϕ will change more rapidly. This is why the speed limit is sometimes lowered when we enter a curve on a highway. In other words, the rate of change of ϕ will depend on the parametrization we use to describe the space curve. To eliminate this dependence on the parametrization, we choose to work with an arc length parametrization $\mathbf{r}(s)$, which means we move along the curve with unit speed.

Using an arc length parametrization $\mathbf{r}(s)$, we define the tangent vector $\mathbf{T}(s) = \mathbf{r}'(s)$, and note that $|\mathbf{T}(s)| = 1$; that is, $\mathbf{T}(s)$ is a unit tangent vector. We then have $\mathbf{T}(s) = \langle \cos(\phi(s)), \sin(\phi(s)) \rangle$, which means that

$$\frac{d\mathbf{T}}{ds} = \left\langle -\sin(\phi(s))\frac{d\phi}{ds},\ \cos(\phi(s))\frac{d\phi}{ds} \right\rangle = \langle -\sin(\phi(s)),\ \cos(\phi(s))\rangle \frac{d\phi}{ds}.$$

Therefore

$$\left|\frac{d\mathbf{T}}{ds}\right| = |\langle -\sin(\phi(s)),\ \cos(\phi(s))\rangle| \ \left|\frac{d\phi}{ds}\right| = \left|\frac{d\phi}{ds}\right|$$

This observation leads us to adopt the following definition.

Definition 9.8.6

If C is a smooth space curve and s is an arc length parameter for C, then the *curvature*, κ, of C is

$$\kappa = \kappa(s) = \left|\frac{d\mathbf{T}}{ds}\right|.$$

Note that κ is the Greek lowercase letter "kappa".

Activity 9.8.5

1. We should expect that the curvature of a line is 0 everywhere. To show that our definition of curvature measures this correctly in 2-space, recall that (9.8.4) gives us the arc length parameterization

$$x(s) = x_0 + \frac{a}{\sqrt{a^2+b^2}}s \qquad \text{and} \qquad y(s) = y_0 + \frac{b}{\sqrt{a^2+b^2}}s$$

of a line. Use this information to explain why the curvature of a line is 0 everywhere.

2. Recall that an arc length parameterization of a circle in 2-space of radius a centered at the origin is, from (9.8.3),

$$\mathbf{r}(s) = \left\langle a\cos\left(\frac{s}{a}\right),\ a\sin\left(\frac{s}{a}\right)\right\rangle.$$

Show that the curvature of this circle is the constant $\frac{1}{a}$. What can you say about the relationship between the size of the radius of a circle and the value of its curvature? Why does this make sense?

The definition of curvature relies on our ability to parameterize curves in terms of arc length. Since we have seen that finding an arc length parametrization can be difficult, we would like to be able to express the curvature in terms of a more general parametrization $\mathbf{r}(t)$.

To begin, we need to describe the vector $\mathbf{T}$, which is a vector tangent to the curve having unit length. Of course, the velocity vector $\mathbf{r}'(t)$ is tangent to the curve; we simply need to normalize its length to be one. This means that we may take

$$\mathbf{T}(t) = \frac{\mathbf{r}'(t)}{|\mathbf{r}'(t)|}. \tag{9.8.5}$$

Then the curvature of the curve defined by $\mathbf{r}$ is

$$\begin{aligned}\kappa &= \left|\frac{d\mathbf{T}}{ds}\right| \\ &= \left|\frac{d\mathbf{T}}{dt}\frac{dt}{ds}\right| \\ &= \frac{\left|\frac{d\mathbf{T}}{dt}\right|}{\left|\frac{ds}{dt}\right|} \\ &= \frac{|\mathbf{T}'(t)|}{|\mathbf{r}'(t)|}.\end{aligned}$$

This last formula allows us to use any parameterization of a curve to calculate its curvature. There is another useful formula, given below, whose derivation is left for the exercises.

Formulas for curvature

If $\mathbf{r}$ is a vector-valued function defining a smooth space curve C, and if $\mathbf{r}'(t)$ is not zero and if $\mathbf{r}''(t)$ exists, then the curvature κ of C satisfies

- $\kappa = \kappa(t) = \dfrac{|\mathbf{T}'(t)|}{|\mathbf{r}'(t)|}$
- $\kappa = \frac{|\mathbf{r}'(t)\times\mathbf{r}''(t)|}{|\mathbf{r}'(t)|^3}$.

Activity 9.8.6

Use one of the two formulas for κ in terms of t to help you answer the following questions.

1. The ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ has parameterization
$$\mathbf{r}(t) = \langle a\cos(t), b\sin(t)\rangle.$$
Find the curvature of the ellipse. Assuming $0 < b < a$, at what points is the curvature the greatest and at what points is the curvature the smallest? Does this agree with your intuition?

2. The standard helix has parameterization $\mathbf{r}(t) = \cos(t)\mathbf{i} + \sin(t)\mathbf{j} + t\mathbf{k}$. Find the curvature of the helix. Does the result agree with your intuition?

The curvature has another interpretation. Recall that the tangent line to a curve at a point is the line that best approximates the curve at that point. The curvature at a point on a curve describes the *circle* that best approximates the curve at that point. Remembering that a circle of radius a has curvature $1/a$, then the circle that best approximates the curve near a point on a curve whose curvature is κ has radius $1/\kappa$ and will be tangent to the tangent line at that point and has its center on the concave side of the curve. This circle, called the *osculating circle* of the curve at the point, is shown in Figure 9.8.7 for a portion of a parabola.

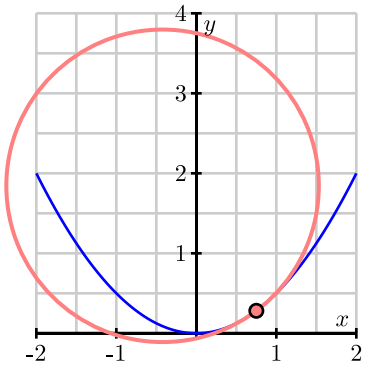


Figure 9.8.7. The osculating circle

9.8.4 Summary

- The integration process shows that the length L of a smooth curve defined by $\mathbf{r}(t)$ on an interval $[a, b]$ is
$$L = \int_a^b |\mathbf{r}'(t)|\, dt.$$
- Arc length is useful as a parameter because when we parameterize with respect to arc length, we eliminate the role of speed in our calculation of curvature and the result is a measure that depends only on the geometry of the curve and not on the

parameterization of the curve.

- We define the curvature κ of a curve in 2- or 3-space to be the rate of change of the magnitude of the unit tangent vector with respect to arc length, or

$$\kappa = \left| \frac{d\mathbf{T}}{ds} \right|.$$

This page titled 9.8: Arc Length and Curvature is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 9.8: Arc Length and Curvature by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

10: DERIVATIVES OF MULTIVARIABLE FUNCTIONS

http://scholarworks.gvsu.edu/books/14/

10.1: Limits

10.2: First-Order Partial Derivatives

10.3: Second-Order Partial Derivatives

10.4: Linearization- Tangent Planes and Differentials

10.5: The Chain Rule

10.6: Directional Derivatives and the Gradient

10.7: Optimization

10.8: Constrained Optimization - Lagrange Multipliers

This page titled 10: Derivatives of Multivariable Functions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 10: Derivatives of Multivariable Functions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

10.1: Limits

Motivating Questions

- What do we mean by the limit of a function f of two variables at a point (a, b)?
- What techniques can we use to show that a function of two variables does not have a limit at a point (a, b)?
- What does it mean for a function f of two variables to be continuous at a point (a, b)?

In this section, we will study limits of functions of several variables, with a focus on limits of functions of two variables. In single variable calculus, we studied the notion of limit, which turned out to be a critical concept that formed the basis for the derivative and the definite integral. In this section we will begin to understand how the concept of limit for functions of two variables is similar to what we encountered for functions of a single variable. The limit will again be the fundamental idea in multivariable calculus, and we will use this notion of the limit of a function of several variables to define the important concept of differentiability later in this chapter. We have already seen its use in the derivatives of vector-valued functions in Section 9.7.

Let's begin by reviewing what we mean by the limit of a function of one variable. We say that a function f has a limit L as x approaches a provided that we can make the values $f(x)$ as close to L as we like by taking x sufficiently close (but not equal) to a. We denote this behavior by writing

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

Preview Activity 10.1.1

We investigate the limits of several different functions by working with tables and graphs.

1. Consider the function f defined by

$$f(x) = 3 - x.$$

Complete Table 10.1.1.

Table 10.1.1. Values of $f(x) = 3 - x$.

x	-0.2	-0.1	0.0	0.1	0.2
$f(x)$					

What does the table suggest regarding $\lim_{x \to 0} f(x)$?

- Explain how your results in (a) are reflected in Figure 10.1.1.

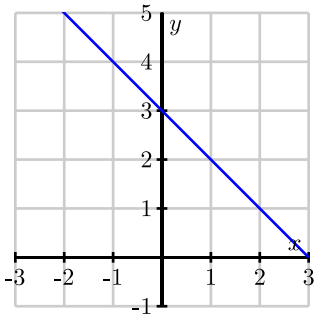


Figure 10.1.1. The graph of $f(x) = 3 - x$.

- Next, consider

$$g(x) = \frac{x}{|x|}.$$

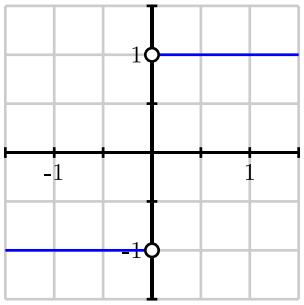
Complete Table 10.1.2 with values near $x = 0$, the point at which g is not defined.

Table 10.1.2. Values of $g(x) = \frac{x}{|x|}$.

x	-0.1	-0.01	-0.001	0.001	0.01	0.1
$g(x)$						

What does this suggest about $\lim_{x \to 0} g(x)$?

- Explain how your results in (c) are reflected in Figure 10.1.2.



- Figure 10.1.2. The graph of $g(x) = \frac{x}{|x|}$.

- Now, let's examine a function of two variables. Let

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

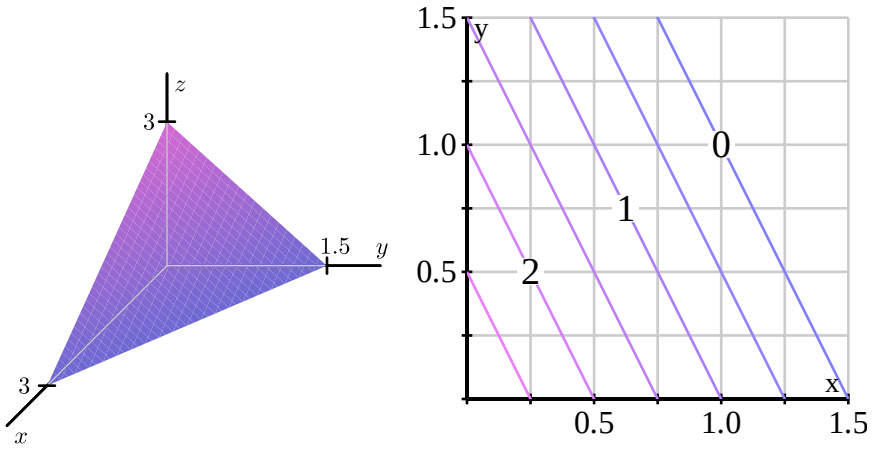
Complete Table 10.1.3.

Table 10.1.3. Values of $f(x, y) = 3 - x - 2y$.

$x \backslash y$	-1.0	-0.1	0.0	0.1	1.0
-1.0		4.2			
-0.1					1.1
0.0				2.8	
0.1	4.9				
1.0			2.0		

What does the table suggest about $\lim_{(x,y) \to (0,0)} f(x, y)$?

- Explain how your results in (e) are reflected in Figure 10.1.3. Compare this limit to the limit in part (a). How are the limits similar and how are they different?



- Figure 10.1.3. Left: The graph of $f(x, y) = 3 - x - 2y$. Right: A contour plot.

- Finally, consider

$$g(x, y) = \frac{2xy}{x^2 + y^2},$$

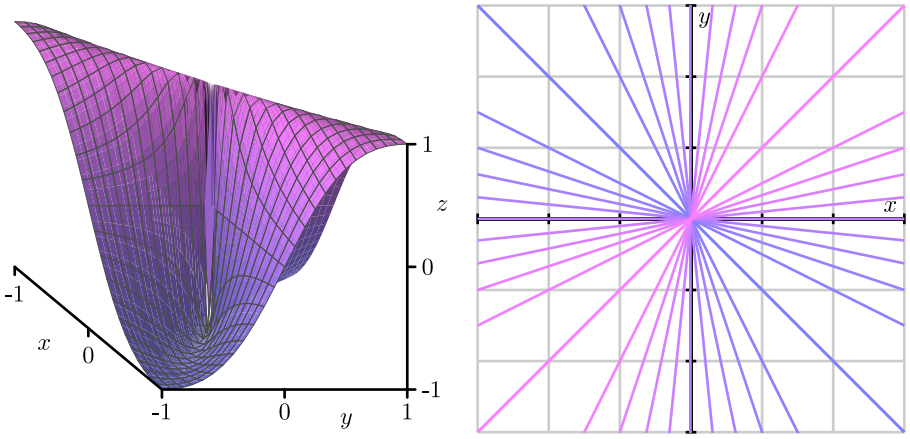
which is not defined at $(0, 0)$. Complete Table 10.1.4. Round to three decimal places.

Table 10.1.4. Values of $g(x, y) = \frac{2xy}{x^2+y^2}$.

$x \backslash y$	-1.0	-0.1	0.0	0.1	1.0
-1.0		0.198			
-0.1					-0.198
0.0			—	0.000	
0.1	-0.198				
1.0			0.000		

What does this suggest about $\lim_{(x,y)\to(0,0)} g(x, y)$?

- Explain how your results are reflected in Figure 10.1.4. Compare this limit to the limit in part (c). How are the results similar and how are they different?



- Figure 10.1.4. Left: The graph of $g(x, y) = \frac{2xy}{x^2+y^2}$. Right: A contour plot.

Limits of Functions of Two Variables

In Preview Activity 10.1.1, we recalled the notion of limit from single variable calculus and saw that a similar concept applies to functions of two variables. Though we will focus on functions of two variables, for the sake of discussion, all the ideas we establish here are valid for functions of any number of variables. In a natural followup to our work in Preview Activity 10.1.1, we now formally define what it means for a function of two variables to have a limit at a point.

Definition 10.1.1

Given a function $f = f(x, y)$, we say that f *has limit* L *as* (x, y) *approaches* (a, b) provided that we can make $f(x, y)$ as close to L as we like by taking (x, y) sufficiently close (but not equal) to (a, b). We write

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

To investigate the limit of a single variable function, $\lim_{x\to a} f(x)$, we often consider the behavior of f as x approaches a from the right and from the left. Similarly, we may investigate limits of two-variable functions, $\lim_{(x,y)\to(a,b)} f(x, y)$ by considering the behavior of f as (x, y) approaches (a, b) from various directions. This situation is more complicated because there are infinitely many ways in which (x, y) may approach (a, b). In the next activity, we see how it is important to consider a variety of those paths in investigating whether or not a limit exists.

Activity 10.1.2

Consider the function f, defined by

$$f(x, y) = \frac{y}{\sqrt{x^2 + y^2}},$$

whose graph is shown below in Figure 10.1.5

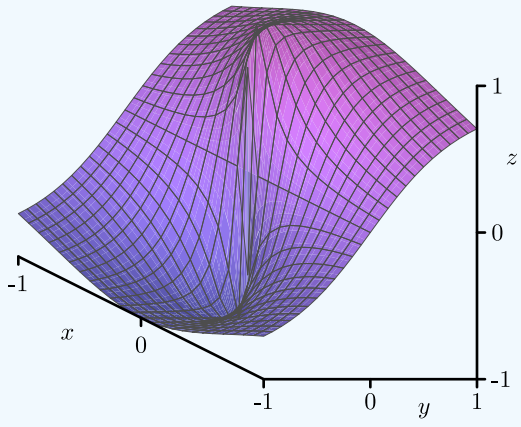


Figure 10.1.5. The graph of $f(x, y) = \frac{y}{\sqrt{x^2+y^2}}$.

1. Is f defined at the point $(0, 0)$? What, if anything, does this say about whether f has a limit at the point $(0, 0)$?
2. Values of f (to three decimal places) at several points close to $(0, 0)$ are shown in Table 10.1.4.

   Table 10.1.4. Values of a function f.

$x\backslash y$	-1.000	-0.100	0.000	0.100	1.000
-1.000	-0.707	—	0.000	—	0.707

$x\backslash y$	-1.000	-0.100	0.000	0.100	1.000
-0.100	—	-0.707	0.000	0.707	—
0.000	-1.000	-1.000	—	1.000	1.000
0.100	—	-0.707	0.000	0.707	—
1.000	-0.707	—	0.000	—	0.707

Based on these calculations, state whether f has a limit at $(0,0)$ and give an argument supporting your statement. (Hint: The blank spaces in the table are there to help you see the patterns.)

3. Now we formalize the conjecture from the previous part by considering what happens if we restrict our attention to different paths. First, we look at f for points in the domain along the x-axis; that is, we consider what happens when $y = 0$. What is the behavior of $f(x,0)$ as $x \to 0$? If we approach $(0,0)$ by moving along the x-axis, what value do we find as the limit?
4. What is the behavior of f along the line $y = x$ when $x > 0$; that is, what is the value of $f(x,x)$ when $x > 0$? If we approach $(0,0)$ by moving along the line $y = x$ in the first quadrant (thus considering $f(x,x)$ as $x \to 0^+$), what value do we find as the limit?
5. In general, if $\lim_{(x,y)\to(0,0)} f(x,y) = L$, then $f(x,y)$ approaches L as (x,y) approaches $(0,0)$, regardless of the path we take in letting $(x,y) \to (0,0)$. Explain what the last two parts of this activity imply about the existence of $\lim_{(x,y)\to(0,0)} f(x,y)$.
6. Shown below in Figure 10.1.6 is a set of contour lines of the function f. What is the behavior of $f(x,y)$ as (x,y) approaches $(0,0)$ along any straight line? How does this observation reinforce your conclusion about the existence of $\lim_{(x,y)\to(0,0)} f(x,y)$ from the previous part of this activity? (Hint: Use the fact that a non-vertical line has equation $y = mx$ for some constant m.)

1.

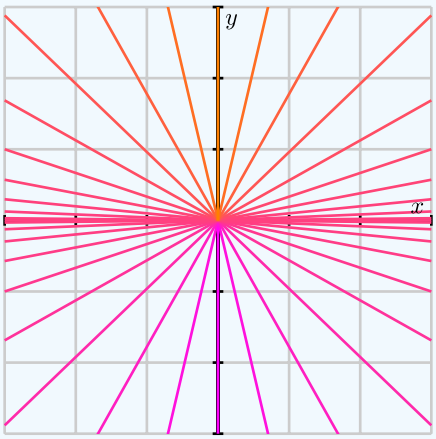


Figure 10.1.6. Contour lines of $f(x,y) = \frac{y}{\sqrt{x^2+y^2}}$.

As we have seen in Activity 10.1.2, if (x,y) approaches (a,b) along two different paths and we find that $f(x,y)$ has two different limits, we can conclude that $\lim_{(x,y)\to(a,b)} f(x,y)$ does not exist. This is similar to the one-variable example $g(x) = x/|x|$ as shown in Figure 10.1.7; $\lim_{x\to 0} g(x)$ does not exist because we see different limits as x approaches 0 from the left and the right.

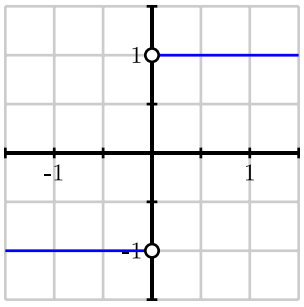


Figure 10.1.7. The graph of $g(x) = \frac{x}{|x|}$.

As a general rule, we have

Limits along different paths

If $f(x, y)$ has two different limits as (x, y) approaches (a, b) along two different paths, then $\lim_{(x,y)\to(a,b)} f(x, y)$ does not exist.

As the next activity shows, studying the limit of a two-variable function f by considering the behavior of f along various paths can require subtle insights.

Activity 10.1.3

Let's consider the function g defined by

$$g(x, y) = \frac{x^2 y}{x^4 + y^2}$$

and investigate the limit $\lim_{(x,y)\to(0,0)} g(x, y)$.

1. What is the behavior of g on the x-axis? That is, what is $g(x, 0)$ and what is the limit of g as (x, y) approaches $(0, 0)$ along the x-axis?
2. What is the behavior of g on the y-axis? That is, what is $g(0, y)$ and what is the limit of g as (x, y) approaches $(0, 0)$ along the y-axis?
3. What is the behavior of g on the line $y = mx$? That is, what is $g(x, mx)$ and what is the limit of g as (x, y) approaches $(0, 0)$ along the line $y = mx$?
4. Based on what you have seen so far, do you think $\lim_{(x,y)\to(0,0)} g(x, y)$ exists? If so, what do you think its value is?
5. Now consider the behavior of g on the parabola $y = x^2$? What is $g(x, x^2)$ and what is the limit of g as (x, y) approaches $(0, 0)$ along this parabola?
6. State whether the limit $\lim_{(x,y)\to(0,0)} g(x, y)$ exists or not and provide a justification of your statement.

This activity shows that we need to be careful when studying the limit of a two-variable functions by considering its behavior along different paths. If we find two different paths that result in two different limits, then we may conclude that the limit does not exist. However, we can never conclude that the limit of a function exists only by considering its behavior along different paths.

Generally speaking, concluding that a limit $\lim_{(x,y)\to(a,b)} f(x, y)$ exists requires a more careful argument.

Example 10.1.1

Consider the function f defined by

$$f(x, y) = \frac{x^2 y^2}{x^2 + y^2}.$$

We want to know whether $\lim_{(x,y)\to(0,0)} f(x, y)$ exists.

Note that if either x or y is 0, then $f(x, y) = 0$. Therefore, if f has a limit at $(0, 0)$, it must be 0. We will therefore argue that

$$\lim_{(x,y)\to(0,0)} f(x, y) = 0,$$

by showing that we can make $f(x, y)$ as close to 0 as we wish by taking (x, y) sufficiently close (but not equal) to $(0, 0)$. In what follows, we view x and y as being real numbers that are close, but not equal, to 0.

Since $0 \leq x^2$, we have

$$y^2 \leq x^2 + y^2,$$

which implies that

$$\frac{y^2}{x^2+y^2} \leq 1.$$

Multiplying both sides by x^2 and observing that $f(x,y) \geq 0$ for all (x,y) gives

$$0 \leq f(x,y) = \frac{x^2y^2}{x^2+y^2} = x^2\left(\frac{y^2}{x^2+y^2}\right) \leq x^2.$$

Thus, $0 \leq f(x,y) \leq x^2$. Since $x^2 \to 0$ as $x \to 0$, we can make $f(x,y)$ as close to 0 as we like by taking x sufficiently close to 0 (for this example, it turns out that we don't even need to worry about making y close to 0). Therefore,

$$\lim_{(x,y)\to(0,0)} \frac{x^2y^2}{x^2+y^2} = 0.$$

In spite of the fact that these two most recent examples illustrate some of the complications that arise when studying limits of two-variable functions, many of the properties that are familiar from our study of single variable functions hold in precisely the same way.

Properties of Limits

Let $f = f(x,y)$ and $g = g(x,y)$ be functions so that $\lim_{(x,y)\to(a,b)} f(x,y)$ and $\lim_{(x,y)\to(a,b)} g(x,y)$ both exist. Then

1. $\lim\limits_{(x,y)\to(a,b)} x = a$ and $\lim\limits_{(x,y)\to(a,b)} y = b$
2. $\lim\limits_{(x,y)\to(a,b)} cf(x,y) = c\left(\lim\limits_{(x,y)\to(a,b)} f(x,y)\right)$ for any scalar c
3. $\lim\limits_{(x,y)\to(a,b)} [f(x,y) \pm g(x,y)] = \lim\limits_{(x,y)\to(a,b)} f(x,y) \pm \lim\limits_{(x,y)\to(a,b)} g(x,y)$
4. $\lim\limits_{(x,y)\to(a,b)} [f(x,y)g(x,y)] = \left(\lim\limits_{(x,y)\to(a,b)} f(x,y)\right)\left(\lim\limits_{(x,y)\to(a,b)} g(x,y)\right)$
5. $\lim\limits_{(x,y)\to(a,b)} \dfrac{f(x,y)}{g(x,y)} = \dfrac{\lim\limits_{(x,y)\to(a,b)} f(x,y)}{\lim\limits_{(x,y)\to(a,b)} g(x,y)}$ if $\lim\limits_{(x,y)\to(a,b)} g(x,y) \neq 0.$

We can use these properties and results from single variable calculus to verify that many limits exist. For example, these properties show that the function f defined by

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

has a limit at every point (a,b) and, moreover,

$$\lim_{(x,y)\to(a,b)} f(x,y) = f(a,b).$$

The reason for this is that polynomial functions of a single variable have limits at every point.

Continuity

Recall that a function f of a single variable x is said to be continuous at $x = a$ provided that the following three conditions are satisfied:

1. $f(a)$ exists,
2. $\lim_{x\to a} f(x)$ exists, and
3. $\lim_{x\to a} f(x) = f(a)$.

Using our understanding of limits of multivariable functions, we can define continuity in the same way.

Definition 10.1.2

A function $f = f(x, y)$ is *continuous* at the point (a, b) provided that

1. f is defined at the point (a, b),
2. $\lim_{(x,y)\to(a,b)} f(x, y)$ exists, and
3. $\lim_{(x,y)\to(a,b)} f(x, y) = f(a, b)$.

For instance, we have seen that the function f defined by $f(x, y) = 3x^2y^3 + 2xy^2 - 3x + 1$ is continous at every point. And just as with single variable functions, continuity has certain properties that are based on the properties of limits.

Properties of continuity

Let f and g be functions of two variables that are continuous at the point (a, b). Then

1. cf is continuous at (a, b) for any scalar c
2. $f + g$ is continuous at (a, b)
3. $f - g$ is continuous at (a, b)
4. fg is continuous at (a, b)
5. $\frac{f}{g}$ is continuous at (a, b) if $g(a, b) \neq 0$

Using these properties, we can apply results from single variable calculus to decide about continuity of multivariable functions. For example, the coordinate functions f and g defined by $f(x, y) = x$ and $g(x, y) = y$ are continuous at every point. We can then use properties of continuity listed to conclude that every polynomial function in x and y is continuous at every point. For example, $g(x, y) = x^2$ and $h(x, y) = y^3$ are continuous functions, so their product $f(x, y) = x^2y^3$ is a continuous multivariable function.

Summary

- A function $f = f(x, y)$ has a limit L at a point (a, b) provided that we can make $f(x, y)$ as close to L as we like by taking (x, y) sufficiently close (but not equal) to (a, b).
- If (x, y) has two different limits as (x, y) approaches (a, b) along two different paths, we can conclude that $\lim_{(x,y)\to(a,b)} f(x, y)$ does not exist.
- Properties similar to those for one-variable functions allow us to conclude that many limits exist and to evaluate them.
- A function $f = f(x, y)$ is continuous at a point (a, b) in its domain if f has a limit at (a, b) and

$$f(a, b) = \lim_{(x,y)\to(a,b)} f(x, y).$$

This page titled 10.1: Limits is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.1: Limits by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

10.2: First-Order Partial Derivatives

Motivating Questions

- How are the first-order partial derivatives of a function f of the independent variables x and y defined?
- Given a function f of the independent variables x and y, what do the first-order partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ tell us about f?

The derivative plays a central role in first semester calculus because it provides important information about a function. Thinking graphically, for instance, the derivative at a point tells us the slope of the tangent line to the graph at that point. In addition, the derivative at a point also provides the instantaneous rate of change of the function with respect to changes in the independent variable.

Now that we are investigating functions of two or more variables, we can still ask how fast the function is changing, though we have to be careful about what we mean. Thinking graphically again, we can try to measure how steep the graph of the function is in a particular direction. Alternatively, we may want to know how fast a function's output changes in response to a change in one of the inputs. Over the next few sections, we will develop tools for addressing issues such as these. Preview Activity $10.2.1$ explores some issues with what we will come to call *partial derivatives*.

Preview Activity $10.2.1$

Suppose we take out an \$18,000 car loan at interest rate r and we agree to pay off the loan in t years. The monthly payment, in dollars, is

$$M(r,t) = \frac{1500r}{1-\left(1+\frac{r}{12}\right)^{-12t}}.$$

a. What is the monthly payment if the interest rate is 3% so that $r = 0.03$, and we pay the loan off in $t = 4$ years?
b. Suppose the interest rate is fixed at 3%. Express M as a function f of t alone using $r = 0.03$. That is, let $f(t) = M(0.03, t)$. Sketch the graph of f on the left of Figure $10.2.1$. Explain the meaning of the function f.

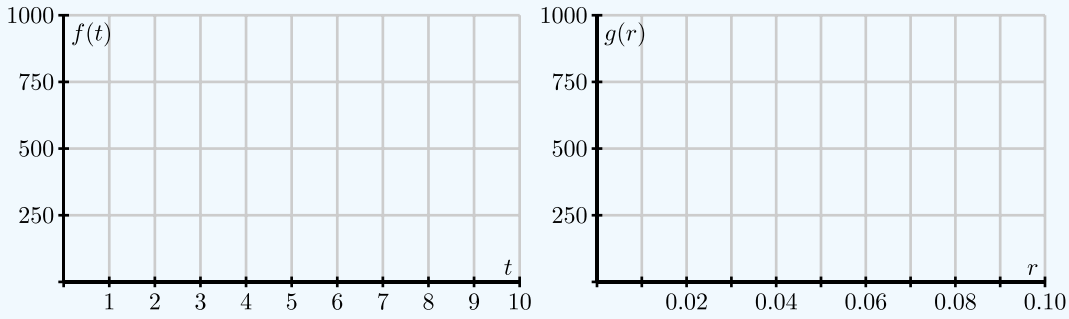


Figure $10.2.1$. Left: Graphs $f(t) = M(0.03, t)$. Right: Graph $g(r) = M(r, 4)$.

c. Find the instantaneous rate of change $f'(4)$ and state the units on this quantity. What information does $f'(4)$ tell us about our car loan? What information does $f'(4)$ tell us about the graph you sketched in (b)?
d. Express M as a function of r alone, using a fixed time of $t = 4$. That is, let $g(r) = M(r, 4)$. Sketch the graph of g on the right of Figure $10.2.1$. Explain the meaning of the function g.
e. Find the instantaneous rate of change $g'(0.03)$ and state the units on this quantity. What information does $g'(0.03)$ tell us about our car loan? What information does $g'(0.03)$ tell us about the graph you sketched in (d)?

First-Order Partial Derivatives

In Section 9.1, we studied the behavior of a function of two or more variables by considering the *traces* of the function. Recall that in one example, we considered the function f defined by

$$f(x,y) = \frac{x^2 \sin(2y)}{32},$$

which measures the range, or horizontal distance, in feet, traveled by a projectile launched with an initial speed of x feet per second at an angle y radians to the horizontal. The graph of this function is given again on the left in Figure 10.2.2. Moreover, if we fix the angle $y = 0.6$, we may view the trace $f(x, 0.6)$ as a function of x alone, as seen at right in Figure 10.2.2.

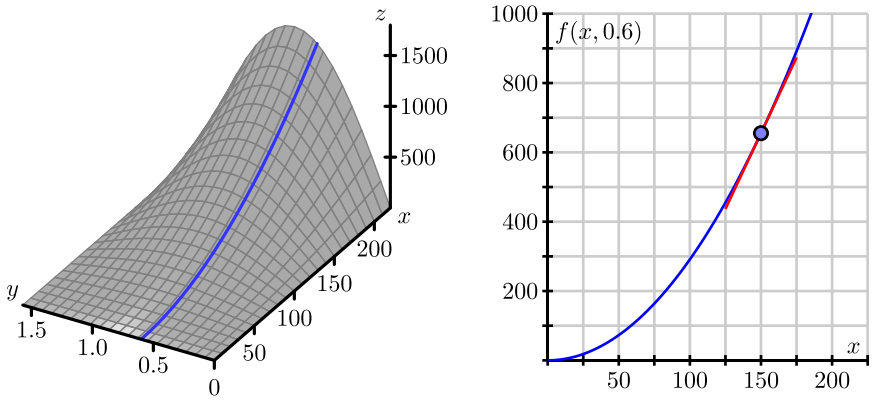


Figure 10.2.2. Left: The trace of $z = \frac{x^2 \sin(2y)}{32}$ with $y = 0.6$.

Since the trace is a one-variable function, we may consider its derivative just as we did in the first semester of calculus. With $y = 0.6$, we have

$$f(x, 0.6) = \frac{\sin(1.2)}{32} x^2,$$

and therefore

$$\frac{d}{dx}[f(x, 0.6)] = \frac{\sin(1.2)}{16} x.$$

When $x = 150$, this gives

$$\frac{d}{dx}[f(x, 0.6)]|_{x=150} = \frac{\sin(1.2)}{16} 150 \approx 8.74 \text{ feet per feet per second},$$

which gives the slope of the tangent line shown on the right of Figure 10.2.2. Thinking of this derivative as an instantaneous rate of change implies that if we increase the initial speed of the projectile by one foot per second, we expect the horizontal distance traveled to increase by approximately 8.74 feet if we hold the launch angle constant at 0.6 radians.

By holding y fixed and differentiating with respect to x, we obtain the first-order *partial derivative of f with respect to x*. Denoting this partial derivative as f_x, we have seen that

$$f_x(150, 0.6) = \frac{d}{dx} f(x, 0.6)|_{x=150} = \lim_{h \to 0} \frac{f(150 + h, 0.6) - f(150, 0.6)}{h}.$$

More generally, we have

$$f_x(a, b) = \lim_{h \to 0} \frac{f(a + h, b) - f(a, b)}{h},$$

provided this limit exists.

In the same way, we may obtain a trace by setting, say, $x = 150$ as shown in Figure 10.2.3.

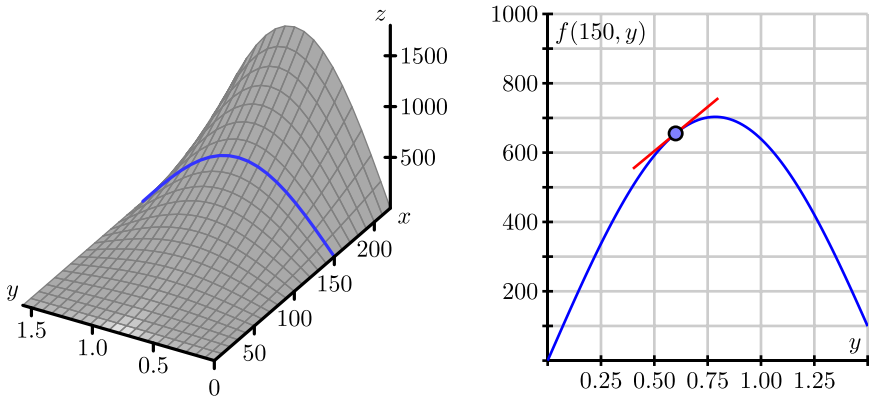


Figure 10.2.3. The trace of $z = \frac{x^2 \sin(2y)}{32}$ with $x = 150$.

This gives

$$f(150, y) = \frac{150^2}{32}\sin(2y),$$

and therefore

$$\frac{d}{dy}[f(150, y)] = \frac{150^2}{16}\cos(2y).$$

If we evaluate this quantity at $y = 0.6$, we have

$$\frac{d}{dy}[f(150, y)]|_{y=0.6} = \frac{150^2}{16}\cos(1.2) \approx 509.5 \text{ feet per radian.}$$

Once again, the derivative gives the slope of the tangent line shown on the right in Figure 10.2.3. Thinking of the derivative as an instantaneous rate of change, we expect that the range of the projectile increases by 509.5 feet for every radian we increase the launch angle y if we keep the initial speed of the projectile constant at 150 feet per second.

By holding x fixed and differentiating with respect to y, we obtain the first-order *partial derivative of f with respect to y*. As before, we denote this partial derivative as f_y and write

$$f_y(150, 0.6) = \frac{d}{dy}f(150, y)|_{y=0.6} = \lim_{h\to 0}\frac{f(150, 0.6 + h) - f(150, 0.6)}{h}.$$

As with the partial derivative with respect to x, we may express this quantity more generally at an arbitrary point (a, b). To recap, we have now arrived at the formal definition of the first-order partial derivatives of a function of two variables.

Definition 10.2.1

The first-order *partial derivatives of f with respect to x and y* at a point (a, b) are, respectively,

$$f_x(a, b) = \lim_{h\to 0}\frac{f(a + h, b) - f(a, b)}{h}, \text{ and}$$
$$f_y(a, b) = \lim_{h\to 0}\frac{f(a, b + h) - f(a, b)}{h},$$

provided the limits exist.

Activity 10.2.2

Consider the function f defined by

$$f(x,y) = \frac{xy^2}{x+1}$$

at the point $(1,2)$.

a. Write the trace $f(x,2)$ at the fixed value $y=2$. On the left side of Figure 10.2.4, draw the graph of the trace with $y=2$ around the point where $x=1$, indicating the scale and labels on the axes. Also, sketch the tangent line at the point $x=1$.

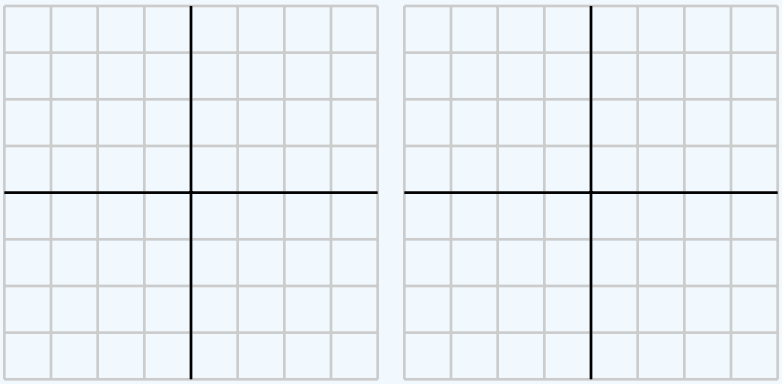


Figure 10.2.4. Traces of $f(x,y) = \frac{xy^2}{x+1}$.

b. Find the partial derivative $f_x(1,2)$ and relate its value to the sketch you just made.
c. Write the trace $f(1,y)$ at the fixed value $x=1$. On the right side of Figure 10.2.5, draw the graph of the trace with $x=1$ indicating the scale and labels on the axes. Also, sketch the tangent line at the point $y=2$.
d. Find the partial derivative $f_y(1,2)$ and relate its value to the sketch you just made.

As these examples show, each partial derivative at a point arises as the derivative of a one-variable function defined by fixing one of the coordinates. In addition, we may consider each partial derivative as defining a new function of the point (x,y), just as the derivative $f'(x)$ defines a new function of x in single-variable calculus. Due to the connection between one-variable derivatives and partial derivatives, we will often use Leibniz-style notation to denote partial derivatives by writing

$$\frac{\partial f}{\partial x}(a,b) = f_x(a,b), \text{ and } \frac{\partial f}{\partial y}(a,b) = f_y(a,b).$$

To calculate the partial derivative f_x, we hold y fixed and thus we treat y as a constant. In Leibniz notation, observe that

$$\frac{\partial}{\partial x}(x) = 1 \text{ and } \frac{\partial}{\partial x}(y) = 0.$$

To see the contrast between how we calculate single variable derivatives and partial derivatives, and the difference between the notations $\frac{d}{dx}[\,]$ and $\frac{\partial}{\partial x}[\,]$, observe that

$$\frac{d}{dx}[3x^2 - 2x + 3] = 3\frac{d}{dx}[x^2] - 2\frac{d}{dx}[x] + \frac{d}{dx}[3] = 3 \cdot 2x - 2,$$
$$\text{and } \frac{\partial}{\partial x}[x^2y - xy + 2y] = y\frac{\partial}{\partial x}[x^2] - y\frac{\partial}{\partial x}[x] + \frac{\partial}{\partial x}[2y] = y \cdot 2x - y$$

Thus, computing partial derivatives is straightforward: we use the standard rules of single variable calculus, but do so while holding one (or more) of the variables constant.

Activity 10.2.3

1. If $f(x,y) = 3x^3 - 2x^2y^5$, find the partial derivatives f_x and f_y.
2. If $f(x,y) = \dfrac{xy^2}{x+1}$, find the partial derivatives f_x and f_y.
3. If $g(r,s) = rs\cos(r)$, find the partial derivatives g_r and g_s.
4. Assuming $f(w,x,y) = (6w+1)\cos(3x^2+4xy^3+y)$, find the partial derivatives f_w, f_x, and f_y.
5. Find all possible first-order partial derivatives of $q(x,t,z) = \dfrac{x2^t z^3}{1+x^2}$.

Interpretations of First-Order Partial Derivatives

Recall that the derivative of a single variable function has a geometric interpretation as the slope of the line tangent to the graph at a given point. Similarly, we have seen that the partial derivatives measure the slope of a line tangent to a trace of a function of two variables as shown in Figure 10.2.5.

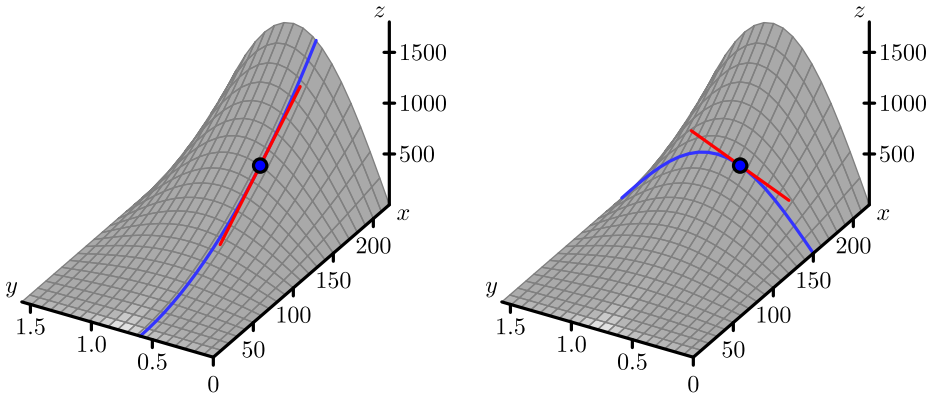


Figure 10.2.5. Tangent lines to two traces of the distance function.

Now we consider the first-order partial derivatives in context. Recall that the difference quotient $\frac{f(a+h)-f(a)}{h}$ for a function f of a single variable x at a point where $x = a$ tells us the average rate of change of f over the interval $[a, a+h]$, while the derivative $f'(a)$ tells us the instantaneous rate of change of f at $x = a$. We can use these same concepts to explain the meanings of the partial derivatives in context.

Activity 10.2.4

The speed of sound C traveling through ocean water is a function of temperature, salinity and depth. It may be modeled by the function

$$C = 1449.2 + 4.6T - 0.055T^2 + 0.00029T^3 + (1.34 - 0.01T)(S-35) + 0.016D.$$

Here C is the speed of sound in meters/second, T is the temperature in degrees Celsius, S is the salinity in grams/liter of water, and D is the depth below the ocean surface in meters.

a. State the units in which each of the partial derivatives, C_T, C_S and C_D, are expressed and explain the physical meaning of each.
b. Find the partial derivatives C_T, C_S and C_D.
c. Evaluate each of the three partial derivatives at the point where $T = 10$, $S = 35$ and $D = 100$. What does the sign of each partial derivatives tell us about the behavior of the function C at the point $(10, 35, 100)$?

Using tables and contours to estimate partial derivatives

Remember that functions of two variables are often represented as either a table of data or a contour plot. In single variable calculus, we saw how we can use the difference quotient to approximate derivatives if, instead of an algebraic formula, we only know the value of the function at a few points. The same idea applies to partial derivatives.

Activity 10.2.5

The wind chill, as frequently reported, is a measure of how cold it feels outside when the wind is blowing. In Table 10.2.1, the wind chill w, measured in degrees Fahrenheit, is a function of the wind speed v, measured in miles per hour, and the ambient air temperature T, also measured in degrees Fahrenheit. We thus view w as being of the form $w = w(v, T)$.

Table 10.2.1. Wind chill as a function of wind speed and temperature.

$v \backslash T$	-30	-25	-20	-15	-10	-5	0	5	10	15	20
5	-46	-40	-34	-28	-22	-16	-11	-5	1	7	13
10	-53	-47	-41	-35	-28	-22	-16	-10	-4	3	9
15	-58	-51	-45	-39	-32	-26	-19	-13	-7	0	6
20	-61	-55	-48	-42	-35	-29	-22	-15	-9	-2	4
25	-64	-58	-51	-44	-37	-31	-24	-17	-11	-4	3
30	-67	-60	-53	-46	-39	-33	-26	-19	-12	-5	1
35	-69	-62	-55	-48	-41	-34	-27	-21	-14	-7	0
40	-71	-64	-57	-50	-43	-36	-29	-22	-15	-8	-1

1. Estimate the partial derivative $w_v(20, -10)$. What are the units on this quantity and what does it mean? (Recall that we can estimate a partial derivative of a single variable function f using the symmetric difference quotient $\frac{f(x+h)-f(x-h)}{2h}$ for small values of h. A partial derivative is a derivative of an appropriate trace.)
2. Estimate the partial derivative $w_T(20, -10)$. What are the units on this quantity and what does it mean?
3. Use your results to estimate the wind chill $w(18, -10)$. (Recall from single variable calculus that for a function f of x, $f(x + h) \approx f(x) + hf'(x)$.)
4. Use your results to estimate the wind chill $w(20, -12)$.
5. Consider how you might combine your previous results to estimate the wind chill $w(18, -12)$. Explain your process.

Activity 10.2.6

Shown below in Figure 10.2.6 is a contour plot of a function f. The values of the function on a few of the contours are indicated to the left of the figure.

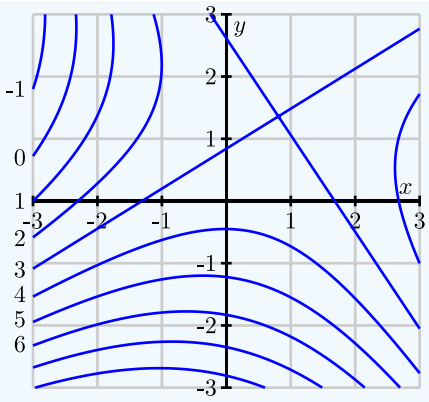


Figure 10.2.6. A contour plot of f.

a. Estimate the partial derivative $f_x(-2,-1)$. (Hint: How can you find values of f that are of the form $f(-2+h)$ and $f(-2-h)$ so that you can use a symmetric difference quotient?)
b. Estimate the partial derivative $f_y(-2,-1)$.
c. Estimate the partial derivatives $f_x(-1,2)$ and $f_y(-1,2)$.
d. Locate, if possible, one point (x,y) where $f_x(x,y)=0$.
e. Locate, if possible, one point (x,y) where $f_x(x,y)<0$.
f. Locate, if possible, one point (x,y) where $f_y(x,y)>0$.
g. Suppose you have a different function g, and you know that $g(2,2)=4$, $g_x(2,2)>0$, and $g_y(2,2)>0$. Using this information, sketch a possibility for the contour $g(x,y)=4$ passing through $(2,2)$ on the left side of Figure 10.2.7. Then include possible contours $g(x,y)=3$ and $g(x,y)=5$.

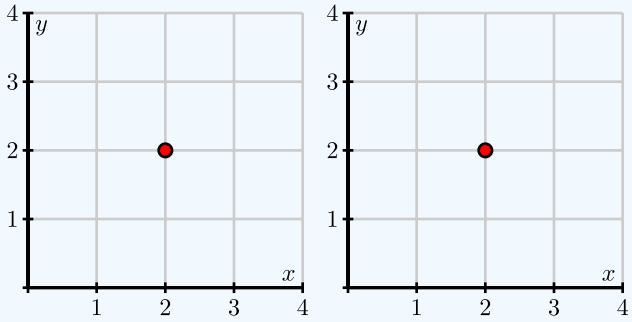


Figure 10.2.7. Plots for contours of g and h.

h. Suppose you have yet another function h, and you know that $h(2,2)=4$, $h_x(2,2)<0$, and $h_y(2,2)>0$. Using this information, sketch a possible contour $h(x,y)=4$ passing through $(2,2)$ on the right side of Figure 10.2.9. Then include possible contours $h(x,y)=3$ and $h(x,y)=5$.

Summary

- If $f=f(x,y)$ is a function of two variables, there are two first order partial derivatives of f: the partial derivative of f with respect to x,

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

and the partial derivative of f with respect to y,

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

where each partial derivative exists only at those points (x,y) for which the limit exists.

- The partial derivative $f_x(a,b)$ tells us the instantaneous rate of change of f with respect to x at $(x,y)=(a,b)$ when y is fixed at b. Geometrically, the partial derivative $f_x(a,b)$ tells us the slope of the line tangent to the $y=b$ trace of the function f at the point $(a,b,f(a,b))$.
- The partial derivative $f_y(a,b)$ tells us the instantaneous rate of change of f with respect to y at $(x,y)=(a,b)$ when x is fixed at a. Geometrically, the partial derivative $f_y(a,b)$ tells us the slope of the line tangent to the $x=a$ trace of the function f at the point $(a,b,f(a,b))$.

This page titled 10.2: First-Order Partial Derivatives is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.2: First-Order Partial Derivatives by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

10.3: Second-Order Partial Derivatives

Motivating Questions

- Given a function f of two independent variables x and y, how are the second-order partial derivatives of f defined?
- What do the second-order partial derivatives f_{xx}, f_{yy}, f_{xy}, and f_{yx} of a function f tell us about the function's behavior?

Recall that for a single-variable function f, the second derivative of f is defined to be the derivative of the first derivative. That is, $f''(x) = \frac{d}{dx}[f'(x)]$, which can be stated in terms of the limit definition of the derivative by writing

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

In what follows, we begin exploring the four different second-order partial derivatives of a function of two variables and seek to understand what these various derivatives tell us about the function's behavior.

Preview Activity 10.3.1

Once again, let's consider the function f defined by $f(x, y) = \frac{x^2 \sin(2y)}{32}$ that measures a projectile's range as a function of its initial speed x and launch angle y. The graph of this function, including traces with $x = 150$ and $y = 0.6$, is shown in Figure 10.3.1.

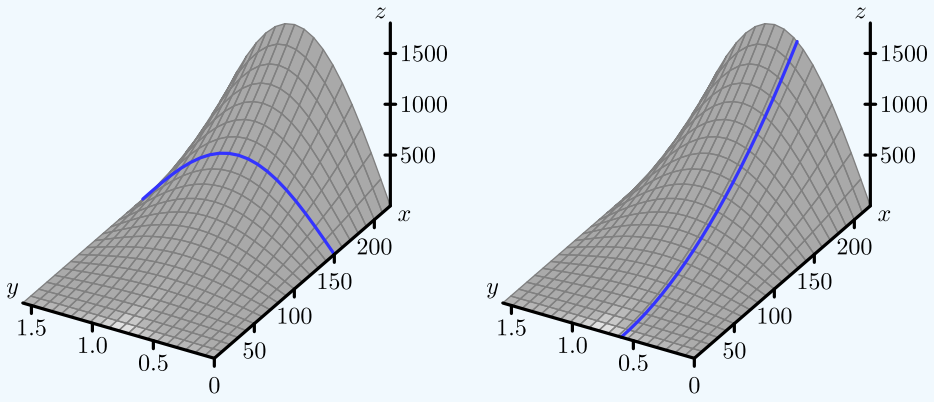


Figure 10.3.1. The distance function with traces $x = 150$ and $y = 0.6$.

a. Compute the partial derivative f_x. Notice that f_x itself is a new function of x and y, so we may now compute the partial derivatives of f_x. Find the partial derivative $f_{xx} = (f_x)_x$ and show that $f_{xx}(150, 0.6) \approx 0.058$.

b. Figure 10.3.2 shows the trace of f with $y = 0.6$ with three tangent lines included. Explain how your result from part (a) of this preview activity is reflected in this figure.

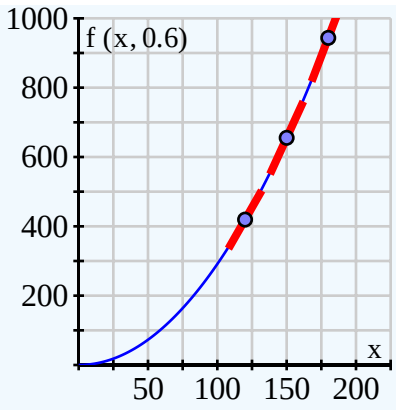


Figure 10.3.2. The trace with $y = 0.6$.

c. Determine the partial derivative f_y, and then find the partial derivative $f_{yy} = (f_y)_y$. Evaluate $f_{yy}(150, 0.6)$.

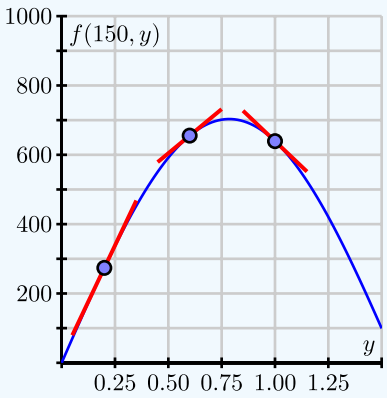


Figure 10.3.3. More traces of the range function.

d. Figure 10.3.3 shows the trace $f(150, y)$ and includes three tangent lines. Explain how the value of $f_{yy}(150, 0.6)$ is reflected in this figure.

e. Because f_x and f_y are each functions of both x and y, they each have two partial derivatives. Not only can we compute $f_{xx} = (f_x)_x$, but also $f_{xy} = (f_x)_y$; likewise, in addition to $f_{yy} = (f_y)_y$, but also $f_{yx} = (f_y)_x$. For the range function $f(x, y) = \frac{x^2 \sin(2y)}{32}$, use your earlier computations of f_x and f_y to now determine f_{xy} and f_{yx}. Write one sentence to explain how you calculated these "mixed" partial derivatives.

Second-Order Partial Derivatives

A function f of two independent variables x and y has two first order partial derivatives, f_x and f_y. As we saw in Preview Activity 10.3.1, each of these first-order partial derivatives has two partial derivatives, giving a total of four *second-order* partial derivatives:

- $f_{xx} = (f_x)_x = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2}$,
- $f_{yy} = (f_y)_y = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y^2}$,
- $f_{xy} = (f_x)_y = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y \partial x}$,
- $f_{yx} = (f_y)_x = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial x \partial y}$.

The first two are called *unmixed* second-order partial derivatives while the last two are called the *mixed* second-order partial derivatives.

One aspect of this notation can be a little confusing. The notation

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)$$

means that we first differentiate with respect to x and then with respect to y; this can be expressed in the alternate notation $f_{xy} = (f_x)_y$. However, to find the second partial derivative

$$f_{yx} = (f_y)_x$$

we first differentiate with respect to y and then x. This means that

$$\frac{\partial^2 f}{\partial y \partial x} = f_{xy}, \text{ and } \frac{\partial^2 f}{\partial x \partial y} = f_{yx}.$$

Be sure to note carefully the difference between Leibniz notation and subscript notation and the order in which x and y appear in each. In addition, remember that anytime we compute a partial derivative, we hold constant the variable(s) other than the one we are differentiating with respect to.

Activity 10.3.2

Find all second order partial derivatives of the following functions. For each partial derivative you calculate, state explicitly which variable is being held constant.

1. $f(x, y) = x^2 y^3$
2. $f(x, y) = y\cos(x)$
3. $g(s, t) = st^3 + s^4$
4. How many second order partial derivatives does the function h defined by $h(x, y, z) = 9x^9 z - xyz^9 + 9$ have? Find h_{xz} and h_{zx} (you do not need to find the other second order partial derivatives).

In Preview Activity 10.3.1 and Activity 10.3.2, you may have noticed that the mixed second-order partial derivatives are equal. This observation holds generally and is known as Clairaut's Theorem.

Clairaut's Theorem

Let f be a function of several variables for which the partial derivatives f_{xy} and f_{yx} are continuous near the point (a, b). Then

$$f_{xy}(a, b) = f_{yx}(a, b).$$

Interpreting the Second-Order Partial Derivatives

Recall from single variable calculus that the second derivative measures the instantaneous rate of change of the derivative. This observation is the key to understanding the meaning of the second-order partial derivatives.

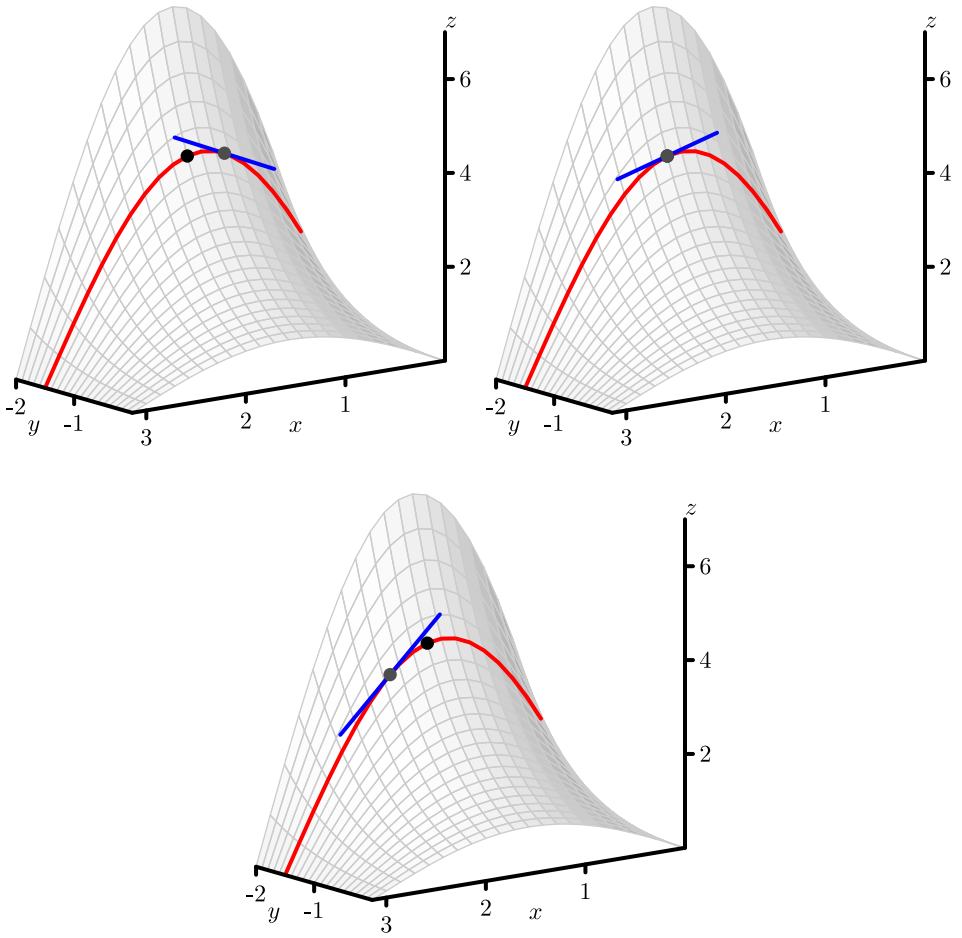


Figure 10.3.4. The tangent lines to a trace with increasing x.

Furthermore, we remember that the second derivative of a function at a point provides us with information about the concavity of the function at that point. Since the unmixed second-order partial derivative f_{xx} requires us to hold y constant and differentiate twice with respect to x, we may simply view f_{xx} as the second derivative of a trace of f where y is fixed. As such, f_{xx} will measure the concavity of this trace.

Consider, for example, $f(x, y) = \sin(x)e^{-y}$. Figure 10.3.4 shows the graph of this function along with the trace given by $y = -1.5$. Also shown are three tangent lines to this trace, with increasing x-values from left to right among the three plots in Figure 10.3.4.

That the slope of the tangent line is decreasing as x increases is reflected, as it is in one-variable calculus, in the fact that the trace is concave down. Indeed, we see that $f_x(x, y) = \cos(x)e^{-y}$ and so $f_{xx}(x, y) = -\sin(x)e^{-y} < 0$, since $e^{-y} > 0$ for all values of y, including $y = -1.5$.

In the following activity, we further explore what second-order partial derivatives tell us about the geometric behavior of a surface.

Activity 10.3.3

We continue to consider the function f defined by $f(x, y) = \sin(x)e^{-y}$.

a. In Figure 10.3.5, we see the trace of $f(x,y) = \sin(x)e^{-y}$ that has x held constant with $x = 1.75$. We also see three different lines that are tangent to the trace of f in the y direction at values of y that are increasing from left to right in the figure. Write a couple of sentences that describe whether the slope of the tangent lines to this curve increase or decrease as y increases, and, after computing $f_{yy}(x,y)$, explain how this observation is related to the value of $f_{yy}(1.75,y)$. Be sure to address the notion of concavity in your response.(You need to be careful about the directions in which x and y are increasing.)

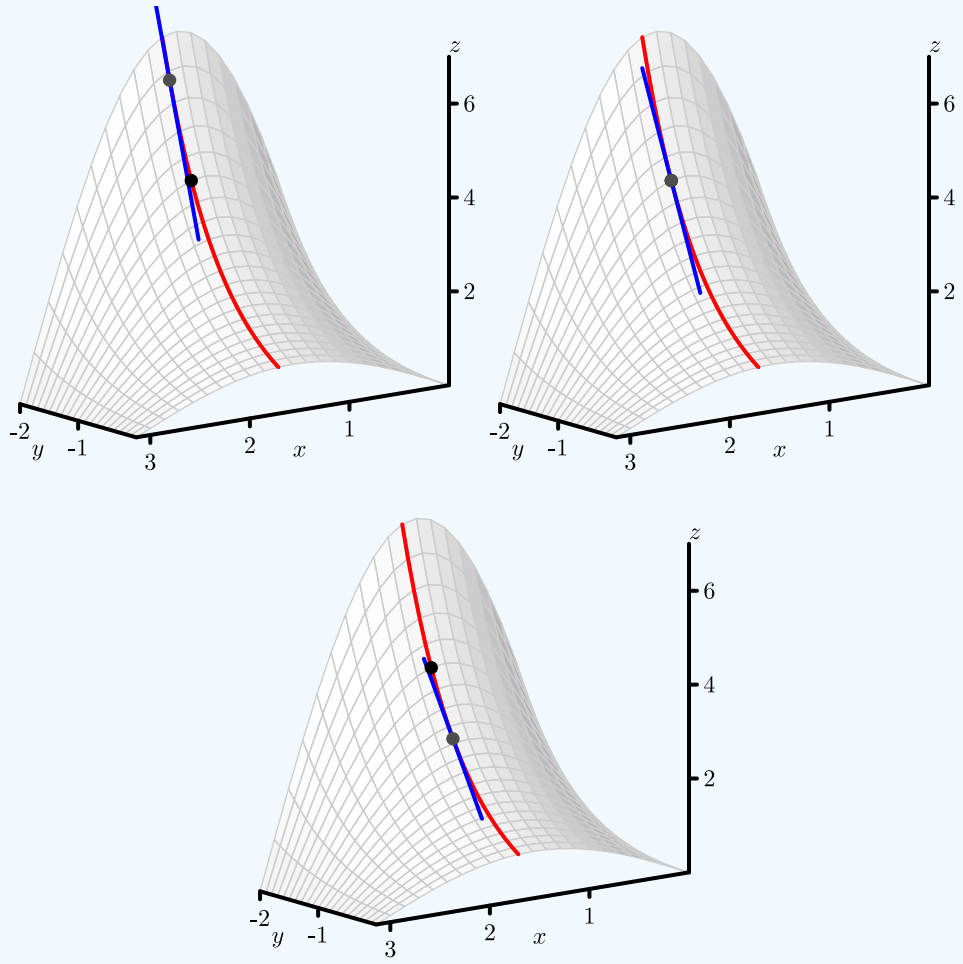


Figure 10.3.5. The tangent lines to a trace with increasing y.

b. In Figure 10.3.6, we start to think about the mixed partial derivative, f_{xy}. Here, we first hold y constant to generate the first-order partial derivative f_x, and then we hold x constant to compute f_{xy}. This leads to first thinking about a trace with x being constant, followed by slopes of tangent lines in the x-direction that slide along the original trace. You might think of sliding your pencil down the trace with x constant in a way that its slope indicates $(f_x)_y$ in order to further animate the three snapshots shown in the figure.

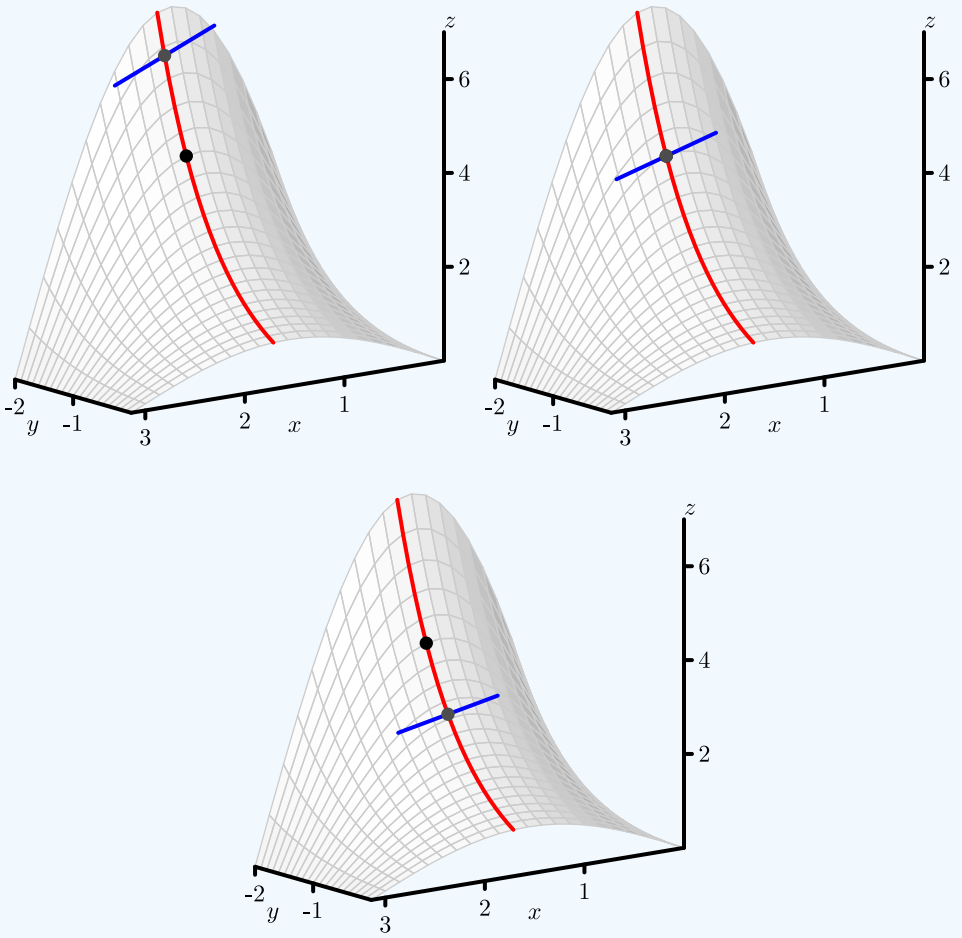


Figure 10.3.6. The trace of $z = f(x, y) = \sin(x)e^{-y}$ with $x = 1.75$, along with tangent lines in the y-direction at three different points.

Based on Figure 10.3.6, is $f_{xy}(1.75, -1.5)$ positive or negative? Why?

c. Determine the formula for $f_{xy}(x, y)$, and hence evaluate $f_{xy}(1.75, -1.5)$. How does this value compare with your observations in (b)?

d. We know that $f_{xx}(1.75, -1.5)$ measures the concavity of the $y = -1.5$ trace, and that $f_{yy}(1.75, -1.5)$ measures the concavity of the $x = 1.75$ trace. What do you think the quantity $f_{xy}(1.75, -1.5)$ measures?

e. On Figure 10.3.6, sketch the trace with $y = -1.5$, and sketch three tangent lines whose slopes correspond to the value of $f_{yx}(x, -1.5)$ for three different values of x, the middle of which is $x = -1.5$. Is $f_{yx}(1.75, -1.5)$ positive or negative? Why? What does $f_{yx}(1.75, -1.5)$ measure?

Just as with the first-order partial derivatives, we can approximate second-order partial derivatives in the situation where we have only partial information about the function.

Activity 10.3.4

As we saw in Activity 10.2.5, the wind chill $w(v, T)$, in degrees Fahrenheit, is a function of the wind speed, in miles per hour, and the air temperature, in degrees Fahrenheit. Some values of the wind chill are recorded in Table 10.3.1.

Table 10.3.1. Wind chill as a function of wind speed and temperature.

$v\backslash T$	-30	-25	-20	-15	-10	-5	0	5	10	15	20
5	-46	-40	-34	-28	-22	-16	-11	-5	1	7	13
10	-53	-47	-41	-35	-28	-22	-16	-10	-4	3	9
15	-58	-51	-45	-39	-32	-26	-19	-13	-7	0	6
20	-61	-55	-48	-42	-35	-29	-22	-15	-9	-2	4
25	-64	-58	-51	-44	-37	-31	-24	-17	-11	-4	3
30	-67	-60	-53	-46	-39	-33	-26	-19	-12	-5	1
35	-69	-62	-55	-48	-41	-34	-27	-21	-14	-7	0
40	-71	-64	-57	-50	-43	-36	-29	-22	-15	-8	-1

1. Estimate the partial derivatives $w_T(20,-15)$, $w_T(20,-10)$, and $w_T(20,-5)$. Use these results to estimate the second-order partial $w_{TT}(20,-10)$.
2. In a similar way, estimate the second-order partial $w_{vv}(20,-10)$.
3. Estimate the partial derivatives $w_T(20,-10)$, $w_T(25,-10)$, and $w_T(15,-10)$, and use your results to estimate the partial $w_{Tv}(20,-10)$.
4. In a similar way, estimate the partial derivative $w_{vT}(20,-10)$.
5. Write several sentences that explain what the values $w_{TT}(20,-10)$, $w_{vv}(20,-10)$, and $w_{Tv}(20,-10)$ indicate regarding the behavior of $w(v,T)$.

As we have found in Activities 10.3.3 and Activity 10.3.4, we may think of f_{xy} as measuring the "twist" of the graph as we increase y along a particular trace where x is held constant. In the same way, f_{yx} measures how the graph twists as we increase x. If we remember that Clairaut's theorem tells us that $f_{xy} = f_{yx}$, we see that the amount of twisting is the same in both directions. This twisting is perhaps more easily seen in Figure 10.3.7, which shows the graph of $f(x,y) = -xy$, for which $f_{xy} = -1$.

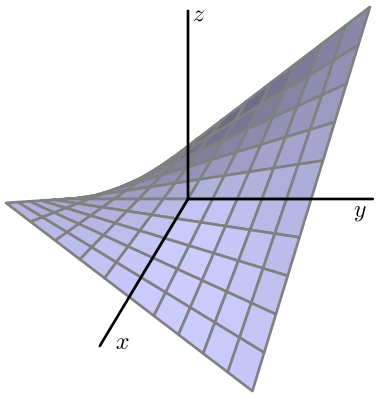


Figure 10.3.7. The graph of $f(x,y) = -xy$.

Summary

- There are four second-order partial derivatives of a function f of two independent variables x and y:
$$f_{xx} = (f_x)_x, f_{xy} = (f_x)_y, f_{yx} = (f_y)_x, \text{ and } f_{yy} = (f_y)_y.$$
- The unmixed second-order partial derivatives, f_{xx} and f_{yy}, tell us about the concavity of the traces. The mixed second-order partial derivatives, f_{xy} and f_{yx}, tell us how the graph of f twists.

This page titled 10.3: Second-Order Partial Derivatives is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.3: Second-Order Partial Derivatives by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

10.4: Linearization- Tangent Planes and Differentials

Motivating Questions

- What does it mean for a function of two variables to be locally linear at a point?
- How do we find the equation of the plane tangent to a locally linear function at a point?
- What is the differential of a multivariable function of two variables and what are its uses?

One of the central concepts in single variable calculus is that the graph of a differentiable function, when viewed on a very small scale, looks like a line. We call this line the tangent line and measure its slope with the derivative. In this section, we will extend this concept to functions of several variables.

Let's see what happens when we look at the graph of a two-variable function on a small scale. To begin, let's consider the function f defined by

$$f(x,y) = 6 - \frac{x^2}{2} - y^2,$$

whose graph is shown in Figure 10.4.1.

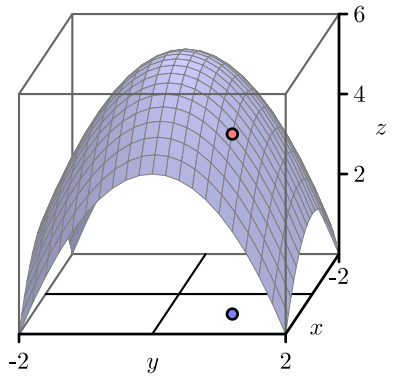


Figure 10.4.1. The graph of $f(x,y) = 6 - x^2/2 - y^2$.

We choose to study the behavior of this function near the point $(x_0, y_0) = (1,1)$. In particular, we wish to view the graph on an increasingly small scale around this point, as shown in the two plots in Figure 10.4.2

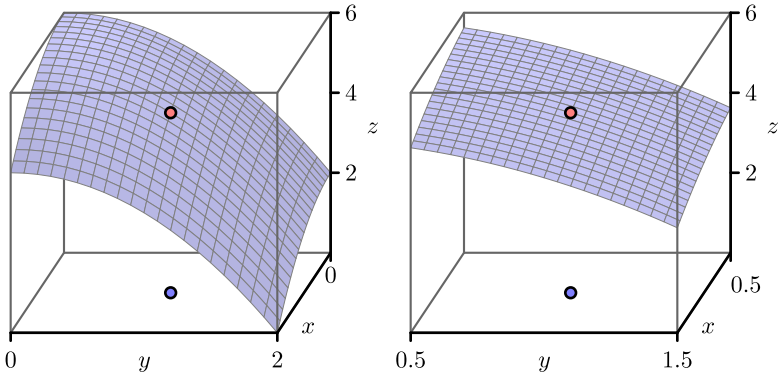


Figure 10.4.2. The graph of $f(x,y) = 6 - x^2/2 - y^2$.

Just as the graph of a differentiable single-variable function looks like a line when viewed on a small scale, we see that the graph

of this particular two-variable function looks like a plane, as seen in Figure 10.4.3. In the following preview activity, we explore how to find the equation of this plane.

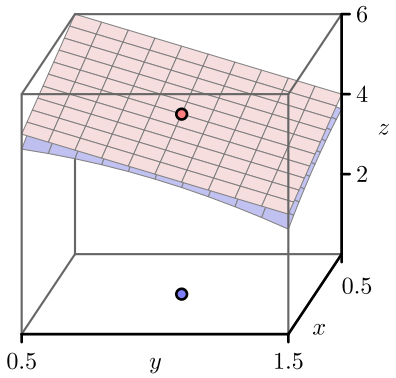


Figure 10.4.3. The graph of $f(x,y) = 6 - x^2/2 - y^2$.

In what follows, we will also use the important fact[1] that the plane passing through (x_0, y_0, z_0) may be expressed in the form $z = z_0 + a(x - x_0) + b(y - y_0)$, where a and b are constants.

As we saw in Section 9.5, the equation of a plane passing through the point (x_0, y_0, z_0) may be written in the form $A(x - x_0) + B(y - y_0) + C(z - z_0) = 0$. If the plane is not vertical, then $C \neq 0$, and we can rearrange this and hence write $C(z - z_0) = -A(x - x_0) - B(y - y_0)$ and thus

$$\begin{aligned} z &= z_0 - \frac{A}{C}(x - x_0) - \frac{B}{C}(y - y_0) \\ &= z_0 + a(x - x_0) + b(y - y_0) \end{aligned}$$

where $a = -A/C$ and $b = -B/C$, respectively.

Preview Activity 10.4.1

Let $f(x,y) = 6 - \frac{x^2}{2} - y^2$, and let $(x_0, y_0) = (1,1)$.

a. Evaluate $f(x,y) = 6 - \frac{x^2}{2} - y^2$ and its partial derivatives at (x_0, y_0); that is, find $f(1,1)$, $f_x(1,1)$, and $f_y(1,1)$.

b. We know one point on the tangent plane; namely, the z-value of the tangent plane agrees with the z-value on the graph of $f(x,y) = 6 - \frac{x^2}{2} - y^2$ at the point (x_0, y_0). In other words, both the tangent plane and the graph of the function f contain the point (x_0, y_0, z_0). Use this observation to determine z_0 in the expression $z = z_0 + a(x - x_0) + b(y - y_0)$.

c. Sketch the traces of $f(x,y) = 6 - \frac{x^2}{2} - y^2$ for $y = y_0 = 1$ and $x = x_0 = 1$ below in Figure 10.4.4.

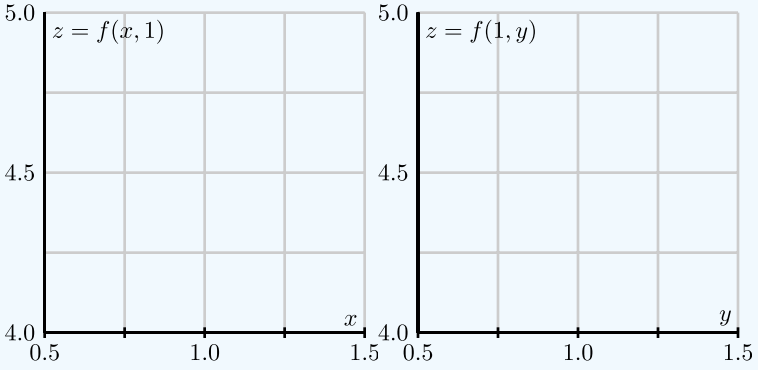


Figure 10.4.4. The traces of $f(x,y)$ with $y = y_0 = 1$ and $x = x_0 = 1$.

d. Determine the equation of the tangent line of the trace that you sketched in the previous part with $y = 1$ (in the x direction) at the point $x_0 = 1$.

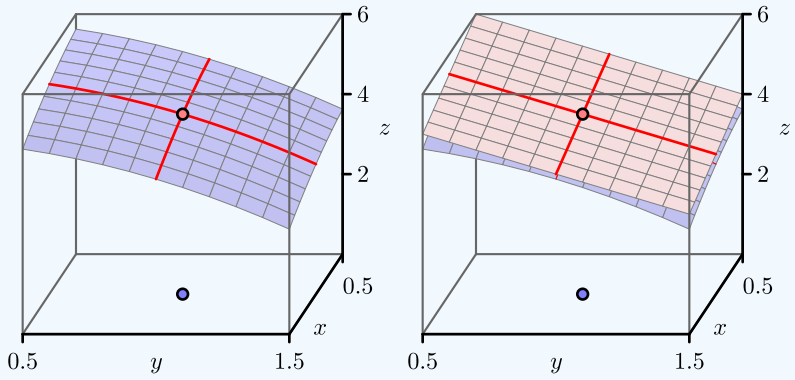


Figure 10.4.5. The traces of $f(x, y)$ and the tangent plane.

e. Figure 10.4.5 shows the traces of the function and the traces of the tangent plane. Explain how the tangent line of the trace of f, whose equation you found in the last part of this activity, is related to the tangent plane. How does this observation help you determine the constant a in the equation for the tangent plane $z = z_0 + a(x - x_0) + b(y - y_0)$? (Hint: How do you think $f_x(x_0, y_0)$ should be related to $z_x(x_0, y_0)$?)

f. In a similar way to what you did in (d), determine the equation of the tangent line of the trace with $x = 1$ at the point $y_0 = 1$. Explain how this tangent line is related to the tangent plane, and use this observation to determine the constant b in the equation for the tangent plane $z = z_0 + a(x - x_0) + b(y - y_0)$. (Hint: How do you think $f_y(x_0, y_0)$ should be related to $z_y(x_0, y_0)$?)

g. Finally, write the equation $z = z_0 + a(x - x_0) + b(y - y_0)$ of the tangent plane to the graph of $f(x, y) = 6 - x^2/2 - y^2$ at the point $(x_0, y_0) = (1, 1)$.

The Tangent Plane

Before stating the formula for the equation of the tangent plane at a point for a general function $f = f(x, y)$, we need to discuss a technical condition. As we have noted, when we look at the graph of a single-variable function on a small scale near a point x_0, we expect to see a line; in this case, we say that f *is locally linear near* x_0 since the graph looks like a linear function locally around x_0. Of course, there are functions, such as the absolute value function given by $f(x) = |x|$, that are not locally linear at every point. In single-variable calculus, we learn that if the derivative of a function exists at a point, then the function is guaranteed to be locally linear there.

In a similar way, we say that a two-variable function f *is locally linear near* (x_0, y_0) provided that the graph of f looks like a plane (its *tangent plane*) when viewed on a small scale near (x_0, y_0). How can we tell when a function of two variables is locally linear at a point?

It is not unreasonable to expect that if $f_x(a, b)$ and $f_y(a, b)$ exist for some function f at a point (a, b), then f is locally linear at (a, b). This is not sufficient, however. As an example, consider the function f defined by $f(x, y) = x^{1/3}y^{1/3}$. In Exercise 10.4.15 you are asked to show that $f_x(0, 0)$ and $f_y(0, 0)$ both exist, but that f is not locally linear at $(0, 0)$ (see Figure 10.4.12). So the existence of the two first order partial derivatives at a point does not guarantee local linearity at that point.

It would take us too far afield to provide a rigorous discussion of differentiability of functions of more than one variable (see Exercise 10.4.15 for a little more detail), so we will be content to define stronger, but more easily verified, conditions that ensure local linearity.

Differentiablity

If f is a function of the independent variables x and y and both f_x and f_y exist and are continuous in an open disk containing the point (x_0, y_0), then f is *continuously differentiable* at (x_0, y_0).

As a consequence, whenever a function $z = f(x, y)$ is continuously differentiable at a point (x_0, y_0), it follows that the function has a tangent plane at (x_0, y_0). Viewed up close, the tangent plane and the function are then virtually indistinguishable. (We won't formally define differentiability of multivariable functions here, and for our purposes continuous differentiability is the only condition we will ever need to use. It is important to note that continuous differentiability is a stronger condition than differentiability. All of the results we encounter will apply to differentiable functions, and so also apply to continuously differentiable functions.) In addition, as in Preview Activity $10.4.1$, we find the following general formula for the tangent plane.

The tangent plane

If $f(x, y)$ has continuous first-order partial derivatives, then the equation of the plane tangent to the graph of f at the point $(x_0, y_0, f(x_0, y_0))$ is

$$z = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0). \tag{10.4.1}$$

Important Note: As can be seen in Exercise $10.4.5.11$, it is possible that $f_x(x_0, y_0)$ and $f_y(x_0, y_0)$ can exist for a function f, and so the plane $z = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0)$ exists even though f is not locally linear at (x_0, y_0) (because the graph of f does not look linear when we zoom in around the point (x_0, y_0)). In such a case this plane is not tangent to the graph. Differentiability for a function of two variables implies the existence of a tangent plane, but the existence of the two first order partial derivatives of a function at a point does not imply differentiaility. This is quite different than what happens in single variable calculus.

Finally, one important note about the form of the equation for the tangent plane, $z = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0)$. Say, for example, that we have the particular tangent plane $z = 7 - 2(x - 3) + 4(y + 1)$. Observe that we can immediately read from this form that $f_x(3, -1) = -2$ and $f_y(3, -1) = 4$; furthermore, $f_x(3, -1) = -2$ is the slope of the trace to both f and the tangent plane in the x-direction at $(3, -1)$. In the same way, $f_y(3, -1) = 4$ is the slope of the trace of both f and the tangent plane in the y-direction at $(3, -1)$.

Activity $10.4.2$

1. Find the equation of the tangent plane to $f(x, y) = 2 + 4x - 3y$ at the point $(1, 2)$. Simplify as much as possible. Does the result surprise you? Explain.
2. Find the equation of the tangent plane to $f(x, y) = x^2 y$ at the point $(1, 2)$.

Linearization

In single variable calculus, an important use of the tangent line is to approximate the value of a differentiable function. Near the point x_0, the tangent line to the graph of f at x_0 is close to the graph of f near x_0, as shown in Figure $10.4.6$.

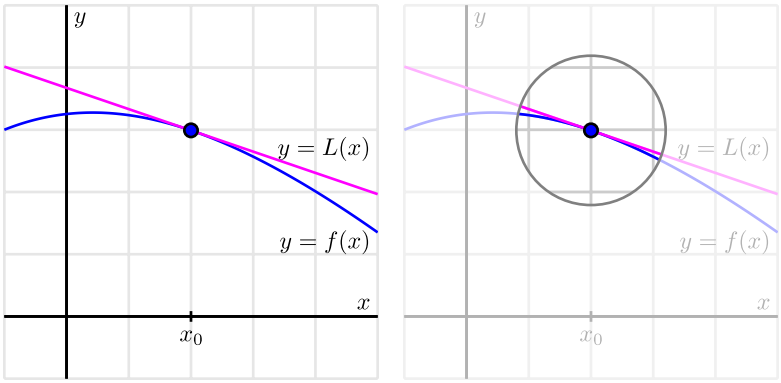


Figure 10.4.6. The linearization of the single-variable function $f(x)$.

In this single-variable setting, we let L denote the function whose graph is the tangent line, and thus

$$L(x) = f(x_0) + f'(x_0)(x - x_0)$$

Furthermore, observe that $f(x) \approx L(x)$ near x_0. We call L the *linearization* of f.

In the same way, the tangent plane to the graph of a differentiable function $z = f(x, y)$ at a point (x_0, y_0) provides a good approximation of $f(x, y)$ near (x_0, y_0). Here, we define the linearization, L, to be the two-variable function whose graph is the tangent plane, and thus

$$L(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0).$$

Finally, note that $f(x, y) \approx L(x, y)$ for points near (x_0, y_0). This is illustrated in Figure 10.4.7.

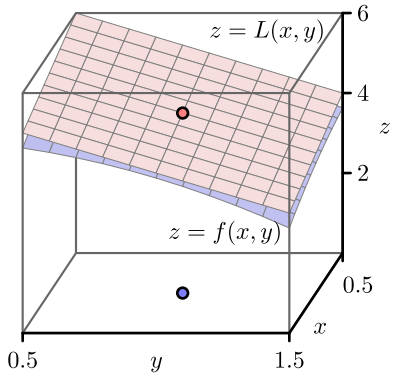


Figure 10.4.7. The linearization of $f(x, y)$.

Activity 10.4.3

In what follows, we find the linearization of several different functions that are given in algebraic, tabular, or graphical form.

a. Find the linearization $L(x, y)$ for the function g defined by

$$g(x, y) = \frac{x}{x^2 + y^2}$$

at the point $(1, 2)$. Then use the linearization to estimate the value of $g(0.8, 2.3)$.

b. Table 10.4.1 provides a collection of values of the wind chill $w(v, T)$, in degrees Fahrenheit, as a function of wind speed, in miles per hour, and temperature, also in degrees Fahrenheit.

Table 10.4.1. Wind chill as a function of wind speed and temperature.

$v\backslash T$	-30	-25	-20	-15	-10	-5	0	5	10	15	20
5	-46	-40	-34	-28	-22	-16	-11	-5	1	7	13
10	-53	-47	-41	-35	-28	-22	-16	-10	-4	3	9
15	-58	-51	-45	-39	-32	-26	-19	-13	-7	0	6
20	-61	-55	-48	-42	-35	-29	-22	-15	-9	-2	4
25	-64	-58	-51	-44	-37	-31	-24	-17	-11	-4	3
30	-67	-60	-53	-46	-39	-33	-26	-19	-12	-5	1
35	-69	-62	-55	-48	-41	-34	-27	-21	-14	-7	0
40	-71	-64	-57	-50	-43	-36	-29	-22	-15	-8	-1

Use the data to first estimate the appropriate partial derivatives, and then find the linearization $L(v, T)$ at the point $(20, -10)$. Finally, use the linearization to estimate $w(10, -10)$, $w(20, -12)$, and $w(18, -12)$. Compare your results to what you obtained in Activity 10.2.5

c. Figure 10.4.9 gives a contour plot of a continuously differentiable function f.

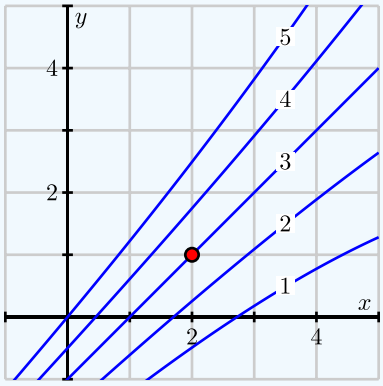


Figure 10.4.9. A contour plot of $f(x, y)$.

After estimating appropriate partial derivatives, determine the linearization $L(x, y)$ at the point $(2, 1)$, and use it to estimate $f(2.2, 1)$, $f(2, 0.8)$, and $f(2.2, 0.8)$.

Differentials

As we have seen, the linearization $L(x, y)$ enables us to estimate the value of $f(x, y)$ for points (x, y) near the base point (x_0, y_0). Sometimes, however, we are more interested in the *change* in f as we move from the base point (x_0, y_0) to another point (x, y).

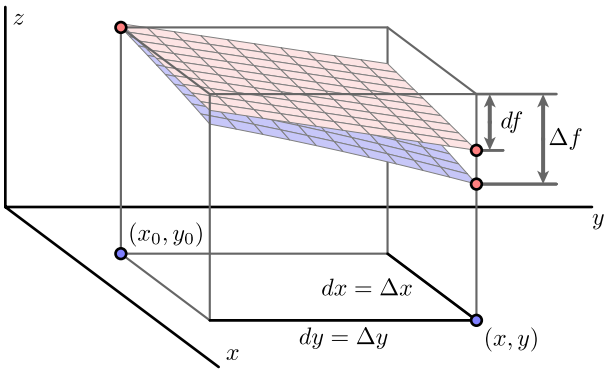


Figure 10.4.10. The differential df approximates the change in $f(x,y)$.

Figure 10.4.10 $illustratesthissituation.Supposeweareatthepoint\((x_0,y_0)$, and we know the value $f(x_0,y_0)$ of f at (x_0,y_0). If we consider the displacement $\langle \Delta x, \Delta y\rangle$ to a new point $(x,y)=(x_0+\Delta x, y_0+\Delta y)$, we would like to know how much the function has changed. We denote this change by Δf, where

$$\Delta f = f(x,y) - f(x_0,y_0).$$

A simple way to estimate the change Δf is to approximate it by df, which represents the change in the linearization $L(x,y)$ as we move from (x_0,y_0) to (x,y). This gives

$$\begin{aligned}\Delta f \approx df &= L(x,y) - f(x_0,y_0)\\ &= [f(x_0,y_0) + f_x(x_0,y_0)(x-x_0) + f_y(x_0,y_0)(y-y_0)] - f(x_0,y_0)\\ &= f_x(x_0,y_0)\Delta x + f_y(x_0,y_0)\Delta y.\end{aligned}$$

For consistency, we will denote the change in the independent variables as $dx=\Delta x$ and $dy=\Delta y$, and thus

$$\Delta f \approx df = f_x(x_0,y_0)\ dx + f_y(x_0,y_0)\ dy. \tag{10.4.2}$$

Expressed equivalently in Leibniz notation, we have

$$df = \frac{\partial f}{\partial x}\ dx + \frac{\partial f}{\partial y}\ dy. \tag{10.4.3}$$

We call the quantities dx, dy, and df *differentials*, and we think of them as measuring small changes in the quantities x, y, and f. Equations (10.4.2) and (10.4.3) express the relationship between these changes. Equation (10.4.3) resembles an important idea from single-variable calculus: when y depends on x, it follows in the notation of differentials that

$$dy = y'\ dx = \frac{dy}{dx}\ dx.$$

We will illustrate the use of differentials with an example.

Example 10.4.1

Suppose we have a machine that manufactures rectangles of width $x=20$ cm and height $y=10$ cm. However, the machine isn't perfect, and therefore the width could be off by $dx=\Delta x=0.2$ cm and the height could be off by $dy=\Delta y=0.4$ cm.

The area of the rectangle is

$$A(x,y)=xy,$$

so that the area of a perfectly manufactured rectangle is $A(20,10)=200$ square centimeters. Since the machine isn't perfect, we would like to know how much the area of a given manufactured rectangle could differ from the perfect rectangle. We will estimate the uncertainty in the area using (10.4.2), and find that

$$\Delta A \approx dA = A_x(20,10)\ dx + A_y(20,10)\ dy.$$

Since $A_x = y$ and $A_y = x$, we have

$$\Delta A \approx dA = 10\ dx + 20\ dy = 10 \cdot 0.2 + 20 \cdot 0.4 = 10.$$

That is, we estimate that the area in our rectangles could be off by as much as 10 square centimeters.

Activity 10.4.4

The questions in this activity explore the differential in several different contexts.

1. Suppose that the elevation of a landscape is given by the function h, where we additionally know that $h(3,1) = 4.35$, $h_x(3,1) = 0.27$, and $h_y(3,1) = -0.19$. Assume that x and y are measured in miles in the easterly and northerly directions, respectively, from some base point $(0,0)$. Your GPS device says that you are currently at the point $(3,1)$. However, you know that the coordinates are only accurate to within 0.2 units; that is, $dx = \Delta x = 0.2$ and $dy = \Delta y = 0.2$. Estimate the uncertainty in your elevation using differentials.
2. The pressure, volume, and temperature of an ideal gas are related by the equation
$$P = P(T,V) = 8.31T/V,$$
where P is measured in kilopascals, V in liters, and T in kelvin. Find the pressure when the volume is 12 liters and the temperature is 310 K. Use differentials to estimate the change in the pressure when the volume increases to 12.3 liters and the temperature decreases to 305 K.
3. Refer to Table 10.4.8, the table of values of the wind chill $w(v,T)$, in degrees Fahrenheit, as a function of temperature, also in degrees Fahrenheit, and wind speed, in miles per hour. Suppose your anemometer says the wind is blowing at 25 miles per hour and your thermometer shows a reading of -15° degrees. However, you know your thermometer is only accurate to within 2° degrees and your anemometer is only accurate to within 3 miles per hour. What is the wind chill based on your measurements? Estimate the uncertainty in your measurement of the wind chill.

Summary

- A function f of two independent variables is locally linear at a point (x_0, y_0) if the graph of f looks like a plane as we zoom in on the graph around the point (x_0, y_0). In this case, the equation of the tangent plane is given by
$$z = f(x_0,y_0) + f_x(x_0,y_0)(x - x_0) + f_y(x_0,y_0)(y - y_0).$$
- The tangent plane $L(x,y) = f(x_0,y_0) + f_x(x_0,y_0)(x - x_0) + f_y(x_0,y_0)(y - y_0)$, when considered as a function, is called the linearization of a differentiable function f at (x_0, y_0) and may be used to estimate values of $f(x,y)$; that is, $f(x,y) \approx L(x,y)$ for points (x,y) near (x_0, y_0).
- A function f of two independent variables is differentiable at (x_0, y_0) provided that both f_x and f_y exist and are continuous in an open disk containing the point (x_0, y_0).
- The differential df of a function $f = f(x,y)$ is related to the differentials dx and dy by
$$df = f_x(x_0,y_0)dx + f_y(x_0,y_0)dy.$$
We can use this relationship to approximate small changes in f that result from small changes in x and y.

This page titled 10.4: Linearization- Tangent Planes and Differentials is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.4: Linearization- Tangent Planes and Differentials by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

10.5: The Chain Rule

Motivating Questions

- What is the Chain Rule and how do we use it to find a derivative?
- How can we use a tree diagram to guide us in applying the Chain Rule?

In single-variable calculus, we encountered situations in which some quantity z depends on y and, in turn, y depends on x. A change in x produces a change in y, which consequently produces a change in z. Using the language of differentials that we saw in the previous section, these changes are naturally related by

$$dz = \frac{dz}{dy}\, dy \text{ and } dy = \frac{dy}{dx}\, dx.$$

In terms of instantaneous rates of change, we then have

$$dz = \frac{dz}{dy}\frac{dy}{dx}\, dx = \frac{dz}{dx}\, dx$$

and thus

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}.$$

This most recent equation we call the *Chain Rule*.

In the case of a function f of two variables where $z = f(x, y)$, it might be that both x and y depend on another variable t. A change in t then produces changes in both x and y, which then cause z to change. In this section we will see how to find the change in z that is caused by a change in t, leading us to multivariable versions of the Chain Rule involving both regular and partial derivatives.

Preview Activity 10.5.1

Suppose you are driving around in the xy-plane in such a way that your position $\mathbf{r}(t)$ at time t is given by function

$$\mathbf{r}(t) = \langle x(t), y(t)\rangle = \langle 2 - t^2, t^3 + 1\rangle.$$

The path taken is shown on the left of Figure 10.5.1.

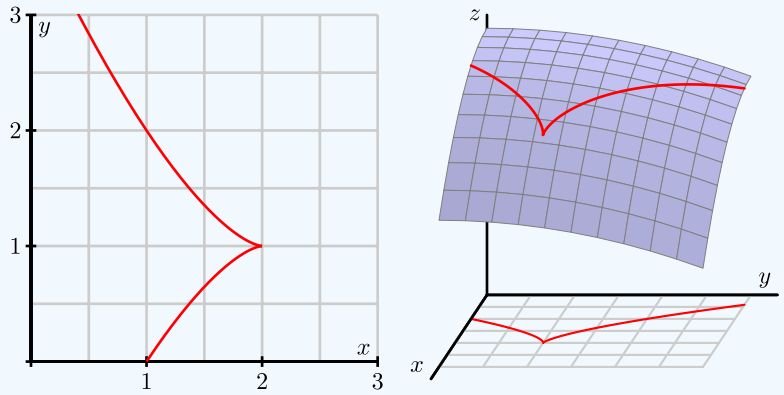


Figure 10.5.1. Left: Your position in the plane. Right: The corresponding temperature.

Suppose, furthermore, that the temperature at a point in the plane is given by

$$T(x,y) = 10 - \frac{1}{2}x^2 - \frac{1}{5}y^2,$$

and note that the surface generated by T is shown on the right of Figure 10.5.1. Therefore, as time passes, your position $(x(t), y(t))$ changes, and, as your position changes, the temperature $T(x,y)$ also changes.

1. The position function $\mathbf{r}$ provides a parameterization $x = x(t)$ and $y = y(t)$ of the position at time t. By substituting $x(t)$ for x and $y(t)$ for y in the formula for T, we can write $T = T(x(t), y(t))$ as a function of t. Make these substitutions to write T as a function of t and then use the Chain Rule from single variable calculus to find $\frac{dT}{dt}$. (Do not do any algebra to simplify the derivative, either before taking the derivative, nor after.)
2. Now we want to understand how the result from part (a) can be obtained from T as a multivariable function. Recall from the previous section that small changes in x and y produce a change in T that is approximated by

   $$\Delta T \approx T_x \Delta x + T_y \Delta y.$$

   The Chain Rule tells us about the instantaneous rate of change of T, and this can be found as

   $$\lim_{\Delta t \to 0} \frac{\Delta T}{\Delta t} = \lim_{\Delta t \to 0} \frac{T_x \Delta x + T_y \Delta y}{\Delta t}. \tag{10.5.1}$$

   Use Equation 10.5.1 to explain why the instantaneous rate of change of T that results from a change in t is

   $$\frac{dT}{dt} = \frac{\partial T}{\partial x}\frac{dx}{dt} + \frac{\partial T}{\partial y}\frac{dy}{dt}. \tag{10.5.2}$$

3. Using the original formulas for T, x, and y in the problem statement, calculate all of the derivatives in Equation 10.5.2 (with T_x and T_y in terms of x and y, and x' and y' in terms of t), and hence write the right-hand side of Equation 10.5.2 in terms of x, y, and t.
4. Compare the results of parts (a) and (c). Write a couple of sentences that identify specifically how each term in (c) relates to a corresponding terms in (a). This connection between parts (a) and (c) provides a multivariable version of the Chain Rule.

The Chain Rule

As Preview Activity 10.5.1 suggests, the following version of the Chain Rule holds in general.

Chain Rule

Let $z = f(x,y)$, where f is a differentiable function of the independent variables x and y, and let x and y each be differentiable functions of an independent variable t. Then

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}. \tag{10.5.3}$$

It is important to note the differences among the derivatives in Equation 10.5.3. Since z is a function of the two variables x and y, the derivatives in the Chain Rule for z with respect to x and y are partial derivatives. However, since $x = x(t)$ and $y = y(t)$ are functions of the single variable t, their derivatives are the standard derivatives of functions of one variable. When we compose z with $x(t)$ and $y(t)$, we then have z as a function of the single variable t, making the derivative of z with respect to t a standard derivative from single variable calculus as well.

To understand why this Chain Rule works in general, suppose that some quantity z depends on x and y so that

$$dz = \frac{\partial z}{\partial x}\,dx + \frac{\partial z}{\partial y}\,dy. \tag{10.5.4}$$

Next, suppose that x and y each depend on another quantity t, so that

$$dx = \frac{dx}{dt}\, dt \text{ and } dy = \frac{dy}{dt}\, dt. \tag{10.5.5}$$

Combining Equations 10.5.4 and 10.5.5, we find that

$$dz = \frac{\partial z}{\partial x}\frac{dx}{dt}\, dt + \frac{\partial z}{\partial y}\frac{dy}{dt}\, dt = \frac{dz}{dt}\, dt,$$

which is the Chain Rule in this particular context, as expressed in Equation 10.5.3.

Activity 10.5.2

In the following questions, we apply the Chain Rule in several different contexts.

1. Suppose that we have a function z defined by $z(x, y) = x^2 + xy^3$. In addition, suppose that x and y are restricted to points that move around the plane by following a circle of radius 2 centered at the origin that is parameterized by
$$x(t) = 2\cos(t), \text{ and } y(t) = 2\sin(t).$$
   1. Use the Chain Rule to find the resulting instantaneous rate of change $\frac{dz}{dt}$.
   2. Substitute $x(t)$ for x and $y(t)$ for y in the rule for z to write z in terms of t and calculate $\frac{dz}{dt}$ directly. Compare to the result of part (i.).
2. Suppose that the temperature on a metal plate is given by the function T with
$$T(x, y) = 100 - (x^2 + 4y^2),$$
where the temperature is measured in degrees Fahrenheit and x and y are each measured in feet.
   1. Find T_x and T_y. What are the units on these partial derivatives?
   2. Suppose an ant is walking along the x-axis at the rate of 2 feet per minute toward the origin. When the ant is at the point $(2, 0)$, what is the instantaneous rate of change in the temperature dT/dt that the ant experiences. Include units on your response.
   3. Suppose instead that the ant walks along an ellipse with $x = 6\cos(t)$ and $y = 3\sin(t)$, where t is measured in minutes. Find $\frac{dT}{dt}$ at $t = \pi/6$, $t = \pi/4$, and $t = \pi/3$. What does this seem to tell you about the path along which the ant is walking?
3. Suppose that you are walking along a surface whose elevation is given by a function f. Furthermore, suppose that if you consider how your location corresponds to points in the xy-plane, you know that when you pass the point $(2, 1)$, your velocity vector is $\mathbf{v} = \langle -1, 2\rangle$. If some contours of f are as shown in Figure 10.5.2, estimate the rate of change df/dt when you pass through $(2, 1)$.

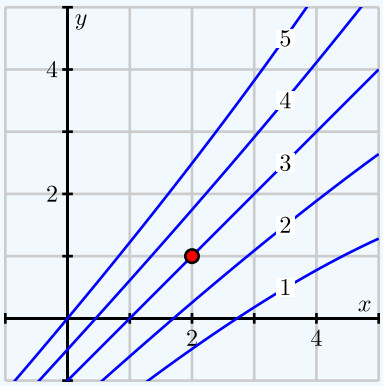


Figure 10.5.2. Some contours of f.

Tree Diagrams

Up to this point, we have applied the Chain Rule to situations where we have a function z of variables x and y, with both x and

y depending on another single quantity t. We may apply the Chain Rule, however, when x and y each depend on more than one quantity, or when z is a function of more than two variables. It can be challenging to keep track of all the dependencies among the variables, and thus a tree diagram can be a useful tool to organize our work. For example, suppose that z depends on x and y, and x and y both depend on t. We may represent these relationships using the tree diagram shown at left Figure 10.5.3. We place the dependent variable at the top of the tree and connect it to the variables on which it depends one level below. We then connect each of those variables to the variable on which each depends.

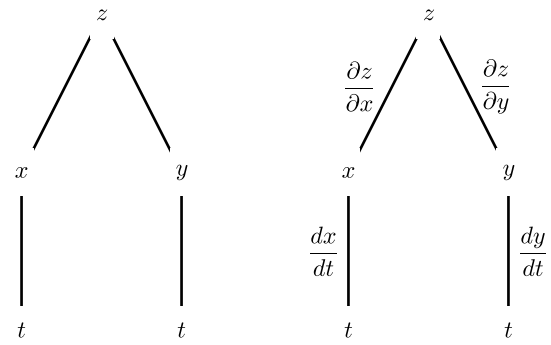


Figure 10.5.3. A tree diagram illustrating dependencies.

To represent the Chain Rule, we label every edge of the diagram with the appropriate derivative or partial derivative, as seen above in Figure 10.5.3. To calculate an overall derivative according to the Chain Rule, we construct the product of the derivatives along all paths connecting the variables and then add all of these products. For example, the diagram above in Figure 10.5.3 illustrates the Chain Rule

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}.$$

Activity 10.5.3

1. Figure 10.5.4 shows the tree diagram we construct when (a) z depends on w, x, and y, (b) w, x, and y each depend on u and v, and (c) u and v depend on t.

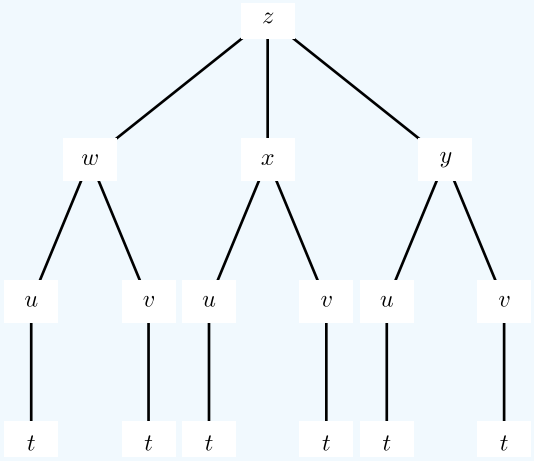


Figure 10.5.4. Three levels of dependencies

   1. Label the edges with the appropriate derivatives.
   2. Use the Chain Rule to write $\frac{dz}{dt}$.
2. Suppose that $z = x^2 - 2xy^2$ and that

$$x = r\cos(\theta)$$
$$y = r\sin(\theta).$$

1. Construct a tree diagram representing the dependencies of z on x and y and x and y on r and θ.
2. Use the tree diagram to find $\frac{\partial z}{\partial r}$.
3. Now suppose that $r = 3$ and $\theta = \pi/6$. Find the values of x and y that correspond to these given values of r and θ, and then use the Chain Rule to find the value of the partial derivative $\frac{\partial z}{\partial \theta}|_{(3,\frac{\pi}{6})}$.

Summary

- The Chain Rule is a tool for differentiating a composite for functions. In its simplest form, it says that if $f(x, y)$ is a function of two variables and $x(t)$ and $y(t)$ depend on t, then
$$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}.$$
- A tree diagram can be used to represent the dependence of variables on other variables. By following the links in the tree diagram, we can form chains of partial derivatives or derivatives that can be combined to give a desired partial derivative.

This page titled 10.5: The Chain Rule is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.5: The Chain Rule by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

10.6: Directional Derivatives and the Gradient

Motivating Questions

- The partial derivatives of a function f tell us the rate of change of f in the direction of the coordinate axes. How can we measure the rate of change of f in other directions?
- What is the gradient of a function and what does it tell us?

The partial derivatives of a function tell us the instantaneous rate at which the function changes as we hold all but one independent variable constant and allow the remaining independent variable to change. It is natural to wonder how we can measure the rate at which a function changes in directions other than parallel to a coordinate axes. In what follows, we investigate this question, and see how the rate of change in any given direction is connected to the rates of change given by the standard partial derivatives.

Preview Activity 10.6.1

Let's consider the function f defined by

$$f(x,y) = 30 - x^2 - \frac{1}{2}y^2,$$

and suppose that f measures the temperature, in degrees Celsius, at a given point in the plane, where x and y are measured in feet. Assume that the positive x-axis points due east, while the positive y-axis points due north. A contour plot of f is shown in Figure 10.6.1.

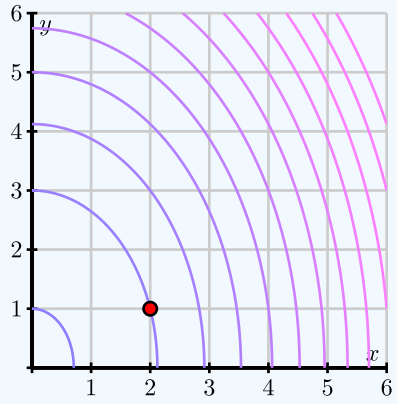


Figure 10.6.1. A contour plot of $f(x,y) = 30 - x^2 - \frac{1}{2}y^2$.

1. Suppose that a person is walking due east, and thus parallel to the x-axis. At what instantaneous rate is the temperature changing with respect to x at the moment the walker passes the point $(2,1)$? What are the units on this rate of change?
2. Next, determine the instantaneous rate of change of temperature with respect to distance at the point $(2,1)$ if the person is instead walking due north. Again, include units on your result.
3. Now, rather than walking due east or due north, let's suppose that the person is walking with velocity given by the vector $\mathbf{v} = \langle 3,4 \rangle$, where time is measured in seconds. Note that the person's speed is thus $|\mathbf{v}| = 5$ feet per second. Find parametric equations for the person's path; that is, parameterize the line through $(2,1)$ using the direction vector $\mathbf{v} = \langle 3,4 \rangle$. Let $x(t)$ denote the x-coordinate of the line, and $y(t)$ its y-coordinate. Make sure your parameterization places the walker at the point $(2,1)$ when $t = 0$.
4. With the parameterization in (c), we can now view the temperature f as not only a function of x and y, but also of time, t. Hence, use the chain rule to determine the value of $\frac{df}{dt}\Big|_{t=0}$. What are the units on your answer? What is the practical meaning of this result?

Directional Derivatives

Given a function $z = f(x, y)$, the partial derivative $f_x(x_0, y_0)$ measures the instantaneous rate of change of f as only the x variable changes; likewise, $f_y(x_0, y_0)$ measures the rate of change of f at (x_0, y_0) as only y changes. Note particularly that $f_x(x_0, y_0)$ is measured in "units of f per unit of change in x," and that the units on $f_y(x_0, y_0)$ are similar.

In Preview Activity 10.6.1, we saw how we could measure the rate of change of f in a situation where both x and y were changing; in that activity, however, we found that this rate of change was measured in "units of f per unit of *time*." In a given unit of time, we may move more than one unit of distance. In fact, in Preview Activity 10.6.1, in each unit increase in time we move a distance of $|\mathbf{v}| = 5$ feet. To generalize the notion of partial derivatives to any direction of our choice, we instead want to have a rate of change whose units are "units of f per unit of distance in the given direction."

In this light, in order to formally define the derivative in a particular direction of motion, we want to represent the change in f for a given *unit* change in the direction of motion. We can represent this unit change in direction with a unit vector, say $\mathbf{u} = \langle u_1, u_2 \rangle$. If we move a distance h in the direction of $\mathbf{u}$ from a fixed point (x_0, y_0), we then arrive at the new point $(x_0 + u_1 h, y_0 + u_2 h)$. It now follows that the slope of the secant line to the curve on the surface through (x_0, y_0) in the direction of $\mathbf{u}$ through the points (x_0, y_0) and $(x_0 + u_1 h, y_0 + u_2 h)$ is

$$m_{\text{sec}} = \frac{f(x_0 + u_1 h, y_0 + u_2 h) - f(x_0, y_0)}{h}. \tag{10.6.1}$$

To get the instantaneous rate of change of f in the direction $\mathbf{u} = \langle u_1, u_2 \rangle$, we must take the limit of the quantity in Equation 10.6.1 as $h \to 0$. Doing so results in the formal definition of the directional derivative.

Definition 10.6.1

Let $f = f(x, y)$ be given. The *derivative of f at the point (x, y) in the direction of the unit vector $\mathbf{u} = \langle u_1, u_2 \rangle$* is denoted $D_{\mathbf{u}} f(x, y)$ and is given by

$$D_{\mathbf{u}} f(x, y) = \lim_{h \to 0} \frac{f(x + u_1 h, y + u_2 h) - f(x, y)}{h} \tag{10.6.2}$$

for those values of x and y for which the limit exists.

The quantity $D_{\mathbf{u}} f(x, y)$ is called a *directional derivative*. When we evaluate the directional derivative $D_{\mathbf{u}} f(x, y)$ at a point (x_0, y_0), the result $D_{\mathbf{u}} f(x_0, y_0)$ tells us the instantaneous rate at which f changes at (x_0, y_0) per unit increase in the direction of the vector $\mathbf{u}$. In addition, the quantity $D_{\mathbf{u}} f(x_0, y_0)$ tells us the slope of the line tangent to the surface in the direction of $\mathbf{u}$ at the point $(x_0, y_0, f(x_0, y_0))$.

Computing the Directional Derivative

In a similar way to how we developed shortcut rules for standard derivatives in single variable calculus, and for partial derivatives in multivariable calculus, we can also find a way to evaluate directional derivatives without resorting to the limit definition found in Equation 10.6.2. We do so using a very similar approach to our work in Preview Activity 10.6.1.

Suppose we consider the situation where we are interested in the instantaneous rate of change of f at a point (x_0, y_0) in the direction $\mathbf{u} = \langle u_1, u_2 \rangle$, where $\mathbf{u}$ is a unit vector. The variables x and y are therefore changing according to the parameterization

$$x = x_0 + u_1 t \qquad \text{and} \qquad y = y_0 + u_2 t.$$

Observe that $\frac{dx}{dt} = u_1$ and $\frac{dy}{dt} = u_2$ for all values of t. Since $\mathbf{u}$ is a unit vector, it follows that a point moving along this line moves one unit of distance per one unit of time; that is, each single unit of time corresponds to movement of a single unit of distance in that direction. This observation allows us to use the Chain Rule to calculate the directional derivative, which measures the instantaneous rate of change of f with respect to change in the direction $\mathbf{u}$.

In particular, by the Chain Rule, it follows that

$$D_{\mathbf{u}}f(x_0,y_0) = f_x(x_0,y_0)\frac{dx}{dt}\Big|_{(x_0,y_0)} + f_y(x_0,y_0)\frac{dy}{dt}\Big|_{(x_0,y_0)}$$
$$= f_x(x_0,y_0)u_1 + f_y(x_0,y_0)u_2.$$

This now allows us to compute the directional derivative at an arbitrary point according to the following formula.

Calculating a directional derivative

Given a differentiable function $f = f(x,y)$ and a unit vector $\mathbf{u} = \langle u_1, u_2\rangle$, we may compute $D_{\mathbf{u}}f(x,y)$ by

$$D_{\mathbf{u}}f(x,y) = f_x(x,y)u_1 + f_y(x,y)u_2. \tag{10.6.3}$$

Note well: To use Equation 10.6.3, we must have a *unit* vector $\mathbf{u} = \langle u_1, u_2\rangle$ in the direction of motion. In the event that we have a direction prescribed by a non-unit vector, we must first scale the vector to have length 1.

Activity 10.6.2

Let $f(x,y) = 3xy - x^2y^3$.

1. Determine $f_x(x,y)$ and $f_y(x,y)$.
2. Use Equation 10.6.3 to determine $D_{\mathbf{i}}f(x,y)$ and $D_{\mathbf{j}}f(x,y)$. What familiar function is $D_{\mathbf{i}}f$? What familiar function is $D_{\mathbf{j}}f$? (Recall that $\mathbf{i}$ is the unit vector in the positive x-direction and $\mathbf{j}$ is the unit vector in the positive y-direction.)
3. Use Equation 10.6.3 to find the derivative of f in the direction of the vector $\mathbf{v} = \langle 2,3\rangle$ at the point $(1,-1)$. Remember that a unit direction vector is needed.

The Gradient

Via the Chain Rule, we have seen that for a given function $f = f(x,y)$, its instantaneous rate of change in the direction of a unit vector $\mathbf{u} = \langle u_1, u_2\rangle$ is given by

$$D_{\mathbf{u}}f(x_0,y_0) = f_x(x_0,y_0)u_1 + f_y(x_0,y_0)u_2. \tag{10.6.4}$$

Recalling that the dot product of two vectors $\mathbf{v} = \langle v_1, v_2\rangle$ and $\mathbf{u} = \langle u_1, u_2\rangle$ is computed by

$$\mathbf{v}\cdot\mathbf{u} = v_1u_1 + v_2u_2,$$

we see that we may recast Equation 10.6.4 in a way that has geometric meaning. In particular, we see that $D_{\mathbf{u}}f(x_0,y_0)$ is the dot product of the vector $\langle f_x(x_0,y_0), f_y(x_0,y_0)\rangle$ and the vector $\mathbf{u}$.

We call this vector formed by the partial derivatives of f the *gradient* of f and denote it

$$\nabla f(x_0,y_0) = \langle f_x(x_0,y_0), f_y(x_0,y_0)\rangle.$$

We read ∇f as "the gradient of f," "grad f" or "del f".[1] Notice that ∇f varies from point to point, and also provides an alternate formulation of the directional derivative.

The directional derivative and the gradient

Given a differentiable function $f = f(x,y)$ and a unit vector $\mathbf{u} = \langle u_1, u_2\rangle$, we may compute $D_{\mathbf{u}}f(x,y)$ by

$$D_{\mathbf{u}}f(x,y) = \nabla f(x,y)\cdot\mathbf{u}. \tag{10.6.5}$$

In the following activity, we investigate some of what the gradient tells us about the behavior of a function f.

The symbol ∇ is called *nabla*, which comes from a Greek word for a certain type of harp that has a similar shape.

Activity 10.6.3

Let's consider the function f defined by $f(x, y) = x^2 - y^2$. Some contours for this function are shown in Figure 10.6.3.

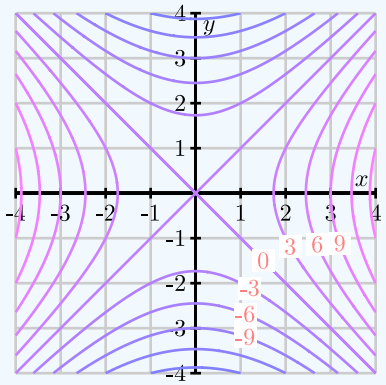


Figure 10.6.3. Contours of $f(x, y) = x^2 - y^2$.

1. Find the gradient $\nabla f(x, y)$.
2. For each of the following points (x_0, y_0), evaluate the gradient $\nabla f(x_0, y_0)$ and sketch the gradient vector with its tail at (x_0, y_0). Some of the vectors are too long to fit onto the plot, but we'd like to draw them to scale; to do so, scale each vector by a factor of 1/4.
   - $(x_0, y_0) = (2, 0)$
   - $(x_0, y_0) = (0, 2)$
   - $(x_0, y_0) = (2, 2)$
   - $(x_0, y_0) = (2, 1)$
   - $(x_0, y_0) = (-3, 2)$
   - $(x_0, y_0) = (-2, -4)$
   - $(x_0, y_0) = (0, 0)$
3. What do you notice about the relationship between the gradient at (x_0, y_0) and the contour passing through that point?
4. Does f increase or decrease in the direction of $\nabla f(x_0, y_0)$? Provide a justification for your response.

As a vector, $\nabla f(x_0, y_0)$ defines a direction and a length. As we will soon see, both of these convey important information about the behavior of f near (x_0, y_0).

The Direction of the Gradient

Remember that the dot product also conveys information about the angle between the two vectors. If θ is the angle between $\nabla f(x_0, y_0)$ and $\mathbf{u}$ (where $\mathbf{u}$ is a unit vector), then we also have that

$$D_{\mathbf{u}} f(x_0, y_0) = \nabla f(x_0, y_0) \cdot \mathbf{u} = |\nabla f(x_0, y_0)||\mathbf{u}| \cos(\theta).$$

In particular, when θ is a right angle, as shown on the left of Figure 10.6.4, then $D_{\mathbf{u}} f(x_0, y_0) = 0$, because $\cos(\theta) = 0$. Since the value of the directional derivative is 0, this means that f is unchanging in this direction, and hence $\mathbf{u}$ must be tangent to the contour of f that passes through (x_0, y_0). In other words, $\nabla f(x_0, y_0)$ is orthogonal to the contour through (x_0, y_0). This shows that the gradient vector at a given point is always perpendicular to the contour passing through the point, confirming that what we saw in part (c) of Activity 10.6.3 holds in general.

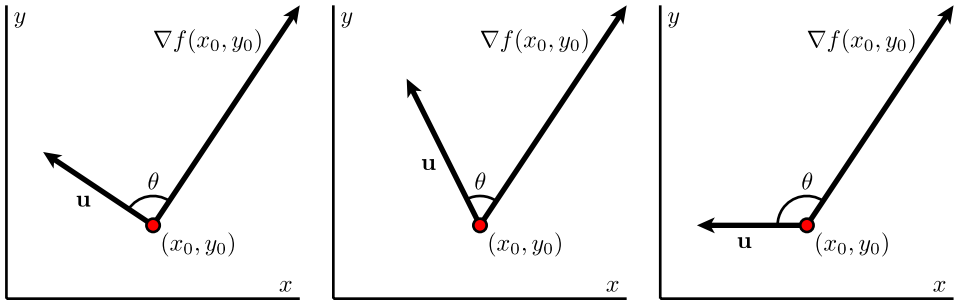


Figure 10.6.4. The sign of $D_{\mathbf{u}}f(x_0, y_0)$ is determined by θ.

Moreover, when θ is an acute angle, it follows that $\cos(\theta) > 0$ so since

$$D_{\mathbf{u}}f(x_0, y_0) = |\nabla f(x_0, y_0)||\mathbf{u}|\cos(\theta),$$

and therefore $D_{\mathbf{u}}f(x_0, y_0) > 0$, as shown in the middle image in Figure 10.6.4. This means that f is increasing in any direction where θ is acute. In a similar way, when θ is an obtuse angle, then $\cos(\theta) < 0$ so $D_{\mathbf{u}}f(x_0, y_0) < 0$, as seen on the right in Figure 10.6.4. This means that f is decreasing in any direction for which θ is obtuse.

Finally, as we can see in the following activity, we may also use the gradient to determine the directions in which the function is increasing and decreasing most rapidly.

Activity 10.6.4

In this activity we investigate how the gradient is related to the directions of greatest increase and decrease of a function. Let f be a differentiable function and $\mathbf{u}$ a unit vector.

1. Let θ be the angle between $\nabla f(x_0, y_0)$ and $\mathbf{u}$. Use the relationship between the dot product and the angle between two vectors to explain why
$$D_{\mathbf{u}}f(x_0, y_0) = |\langle f_x(x_0, y_0), f_y(x_0, y_0)\rangle|\cos(\theta). \tag{10.6.6}$$
2. At the point (x_0, y_0), the only quantity in Equation 10.6.6 that can change is θ (which determines the direction $\mathbf{u}$ of travel). Explain why $\theta = 0$ makes the quantity
$$|\langle f_x(x_0, y_0), f_y(x_0, y_0)\rangle|\cos(\theta)$$
as large as possible.
3. When $\theta = 0$, in what direction does the unit vector $\mathbf{u}$ point relative to $\nabla f(x_0, y_0)$? Why? What does this tell us about the direction of greatest increase of f at the point (x_0, y_0)?
4. In what direction, relative to $\nabla f(x_0, y_0)$, does f decrease most rapidly at the point (x_0, y_0)?
5. State the unit vectors $\mathbf{u}$ and $\mathbf{v}$ (in terms of $\nabla f(x_0, y_0)$) that provide the directions of greatest increase and decrease for the function f at the point (x_0, y_0). What important assumption must we make regarding $\nabla f(x_0, y_0)$ in order for these vectors to exist?

The Length of the Gradient

Having established in Activity 10.6.4 that the direction in which a function increases most rapidly at a point (x_0, y_0) is the unit vector $\mathbf{u}$ in the direction of the gradient, (that is, $\mathbf{u} = \frac{1}{|\nabla f(x_0, y_0)|}\nabla f(x_0, y_0)$, provided that $\nabla f(x_0, y_0) \neq \mathbf{0}$), it is also natural to ask, "in the direction of greatest increase for f at (x_0, y_0), what is the *value* of the rate of increase?" In this situation, we are asking for the value of $D_{\mathbf{u}}f(x_0, y_0)$ where $\mathbf{u} = \frac{1}{|\nabla f(x_0, y_0)|}\nabla f(x_0, y_0)$.

Using the now familiar way to compute the directional derivative, we see that

$$D_{\mathbf{u}}f(x_0,y_0) = \nabla f(x_0,y_0)\cdot\left(\frac{1}{|\nabla f(x_0,y_0)|}\nabla f(x_0,y_0)\right).$$

Next, we recall two important facts about the dot product: (i) $\mathbf{w}\cdot(c\mathbf{v}) = c(\mathbf{w}\cdot\mathbf{v})$ for any scalar c, and (ii) $\mathbf{w}\cdot\mathbf{w} = |\mathbf{w}|^2$. Applying these properties to the most recent equation involving the directional derivative, we find that

$$D_{\mathbf{u}}f(x_0,y_0) = \frac{1}{|\nabla f(x_0,y_0)|}(\nabla f(x_0,y_0)\cdot\nabla f(x_0,y_0)) = \frac{1}{|\nabla f(x_0,y_0)|}|\nabla f(x_0,y_0)|^2.$$

Finally, since $\nabla f(x_0,y_0)$ is a nonzero vector, its length $|\nabla f(x_0,y_0)|$ is a nonzero scalar, and thus we can simplify the preceding equation to establish that

$$D_{\mathbf{u}}f(x_0,y_0) = |\nabla f(x_0,y_0)|.$$

We summarize our most recent work by stating two important facts about the gradient.

Important facts about the gradient

Let f be a differentiable function and (x_0,y_0) a point for which $\nabla f(x_0,y_0)\neq\mathbf{0}$. Then $\nabla f(x_0,y_0)$ points in the direction of greatest increase of f at (x_0,y_0), and the instantaneous rate of change of f in that direction is the length of the gradient vector. That is, if $\mathbf{u} = \frac{1}{|\nabla f(x_0,y_0)|}\nabla f(x_0,y_0)$, then $\mathbf{u}$ is a unit vector in the direction of greatest increase of f at (x_0,y_0), and $D_{\mathbf{u}}f(x_0,y_0) = |\nabla f(x_0,y_0)|$.

Activity 10.6.5

Consider the function f defined by $f(x,y) = -x+2xy-y$.

a. Find the gradient $\nabla f(1,2)$ and sketch it on Figure 10.6.5.

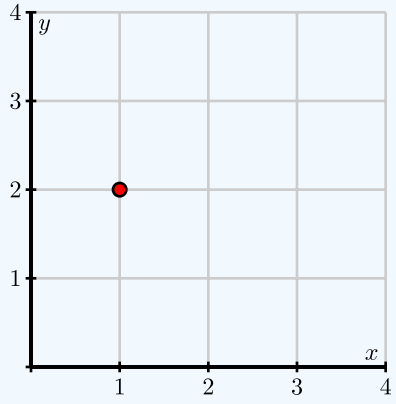


Figure 10.6.5. A plot for the gradient $\nabla f(1,2)$.

b. Sketch the unit vector $\mathbf{z} = \left\langle -\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right\rangle$ on Figure 10.6.5 with its tail at $(1,2)$. Now find the directional derivative $D_{\mathbf{z}}f(1,2)$.

c. What is the slope of the graph of f in the direction $\mathbf{z}$? What does the sign of the directional derivative tell you?

d. Consider the vector $\mathbf{v} = \langle 2,-1\rangle$ and sketch $\mathbf{v}$ on Figure 10.6.5 with its tail at $(1,2)$. Find a unit vector $\mathbf{w}$ pointing in the same direction of $\mathbf{v}$. Without computing $D_{\mathbf{w}}f(1,2)$, what do you know about the sign of this directional derivative? Now verify your observation by computing $D_{\mathbf{w}}f(1,2)$.

e. In which direction (that is, for what unit vector $\mathbf{u}$) is $D_{\mathbf{u}}f(1,2)$ the greatest? What is the slope of the graph in this direction?

f. Corresponding, in which direction is $D_{\mathbf{u}}f(1,2)$ least? What is the slope of the graph in this direction?

g. Sketch two unit vectors $\mathbf{u}$ for which $D_{\mathbf{u}}f(1,2) = 0$ and then find component representations of these vectors.

h. Suppose you are standing at the point $(3,3)$. In which direction should you move to cause f to increase as rapidly as possible? At what rate does f increase in this direction?

Applications

The gradient finds many natural applications. For example, situations often arise — for instance, constructing a road through the mountains or planning the flow of water across a landscape — where we are interested in knowing the direction in which a function is increasing or decreasing most rapidly.

For example, consider a two-dimensional version of how a heat-seeking missile might work.(This application is borrowed from United States Air Force Academy Department of Mathematical Sciences.) Suppose that the temperature surrounding a fighter jet can be modeled by the function T defined by

$$T(x,y) = \frac{100}{1+(x-5)^2+4(y-2.5)^2},$$

where (x,y) is a point in the plane of the fighter jet and $T(x,y)$ is measured in degrees Celsius. Some contours and gradients ∇T are shown on the left in Figure 10.6.6.

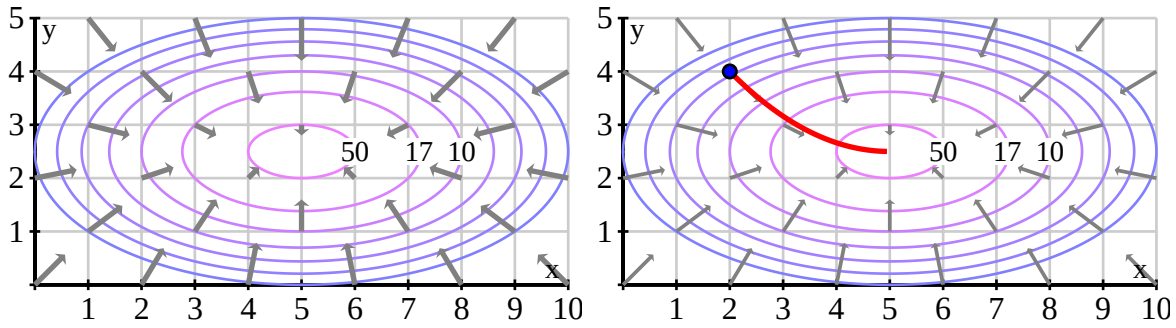


Figure 10.6.6. Contours and gradient for $T(x,y)$ and the missile's path.

A heat-seeking missile will always travel in the direction in which the temperature increases most rapidly; that is, it will always travel in the direction of the gradient ∇T. If a missile is fired from the point $(2,4)$, then its path will be that shown on the right in Figure 10.6.6.

In the final activity of this section, we consider several questions related to this context of a heat-seeking missile, and foreshadow some upcoming work in Section 10.7.

Activity 10.6.6

1. The temperature $T(x,y)$ has its maximum value at the fighter jet's location. State the fighter jet's location and explain how Figure 10.6.6 tells you this.
2. Determine ∇T at the fighter jet's location and give a justification for your response.
3. Suppose that a different function f has a local maximum value at (x_0,y_0). Sketch the behavior of some possible contours near this point. What is $\nabla f(x_0,y_0)$? (Hint: What is the direction of greatest increase in f at the local maximum?)
4. Suppose that a function g has a local minimum value at (x_0,y_0). Sketch the behavior of some possible contours near this point. What is $\nabla g(x_0,y_0)$?
5. If a function g has a local minimum at (x_0,y_0), what is the direction of greatest increase of g at (x_0,y_0)?

Summary

- The directional derivative of f at the point (x, y) in the direction of the unit vector $\mathbf{u} = \langle u_1, u_2 \rangle$ is

$$D_{\mathbf{u}} f(x, y) = \lim_{h \to 0} \frac{f(x + u_1 h, y + u_2 h) - f(x, y)}{h}$$

  for those values of x and y for which the limit exists. In addition, $D_{\mathbf{u}} f(x, y)$ measures the slope of the graph of f when we move in the direction $\mathbf{u}$. Alternatively, $D_{\mathbf{u}} f(x_0, y_0)$ measures the instantaneous rate of change of f in the direction $\mathbf{u}$ at (x_0, y_0).

- The gradient of a function $f = f(x, y)$ at a point (x_0, y_0) is the vector

$$\nabla f(x_0, y_0) = \langle f_x(x_0, y_0), f_y(x_0, y_0) \rangle .$$

- The directional derivative in the direction $\mathbf{u}$ may be computed by

$$D_{\mathbf{u}} f(x_0, y_0) = \nabla f(x_0, y_0) \cdot \mathbf{u}.$$

- At any point where the gradient is nonzero, gradient is orthogonal to the contour through that point and points in the direction in which f increases most rapidly; moreover, the slope of f in this direction equals the length of the gradient $|\nabla f(x_0, y_0)|$. Similarly, the opposite of the gradient points in the direction of greatest decrease, and that rate of decrease is the opposite of the length of the gradient.

This page titled 10.6: Directional Derivatives and the Gradient is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.6: Directional Derivatives and the Gradient by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

10.7: Optimization

Motivating Questions

- How can we find the points at which $f(x, y)$ has a local maximum or minimum?
- How can we determine whether critical points of $f(x, y)$ are local maxima or minima?
- How can we find the absolute maximum and minimum of $f(x, y)$ on a closed and bounded domain?

We learn in single-variable calculus that the derivative is a useful tool for finding the local maxima and minima of functions, and that these ideas may often be employed in applied settings. In particular, if a function f, such as the one shown in Figure $10.7.1$ is everywhere differentiable, we know that the tangent line is horizontal at any point where f has a local maximum or minimum. This, of course, means that the derivative f' is zero at any such point. Hence, one way that we seek extreme values of a given function is to first find where the derivative of the function is zero.

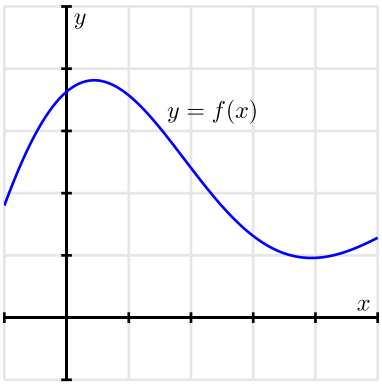


Figure $10.7.1$. The graph of $y = f(x)$.

In multivariable calculus, we are often similarly interested in finding the greatest and/or least value(s) that a function may achieve. Moreover, there are many applied settings in which a quantity of interest depends on several different variables. In the following preview activity, we begin to see how some key ideas in multivariable calculus can help us answer such questions by thinking about the geometry of the surface generated by a function of two variables.

Preview Activity $10.7.1$

Let $z = f(x, y)$ be a differentiable function, and suppose that at the point (x_0, y_0), f achieves a local maximum. That is, the value of $f(x_0, y_0)$ is greater than the value of $f(x, y)$ for all (x, y) nearby (x_0, y_0). You might find it helpful to sketch a rough picture of a possible function f that has this property.

1. If we consider the trace given by holding $y = y_0$ constant, then the single-variable function defined by $f(x, y_0)$ must have a local maximum at x_0. What does this say about the value of the partial derivative $f_x(x_0, y_0)$?
2. In the same way, the trace given by holding $x = x_0$ constant has a local maximum at $y = y_0$. What does this say about the value of the partial derivative $f_y(x_0, y_0)$?
3. What may we now conclude about the gradient $\nabla f(x_0, y_0)$ at the local maximum? How is this consistent with the statement "f increases most rapidly in the direction $\nabla f(x_0, y_0)$?"
4. How will the tangent plane to the surface $z = f(x, y)$ appear at the point $(x_0, y_0, f(x_0, y_0))$?
5. By first computing the partial derivatives, find any points at which $f(x, y) = 2x - x^2 - (y + 2)^2$ may have a local maximum.

Extrema and Critical Points

One of the important applications of single-variable calculus is the use of derivatives to identify local extremes of functions (that is, local maxima and local minima). Using the tools we have developed so far, we can naturally extend the concept of local maxima and minima to several-variable functions.

Definition 10.7.1

Let f be a function of two variables x and y.

- The function f has a *local maximum* at a point (x_0, y_0) provided that $f(x, y) \leq f(x_0, y_0)$ for all points (x, y) near (x_0, y_0). In this situation we say that $f(x_0, y_0)$ is a *local maximum value*.
- The function f has a *local minimum* at a point (x_0, y_0) provided that $f(x, y) \geq f(x_0, y_0)$ for all points (x, y) near (x_0, y_0). In this situation we say that $f(x_0, y_0)$ is a *local minimum value*.
- An *absolute maximum point* is a point (x_0, y_0) for which $f(x, y) \leq f(x_0, y_0)$ for all points (x, y) in the domain of f. The value of f at an absolute maximum point is the *maximum value* of f.
- An *absolute minimum point* is a point such that $f(x, y) \geq f(x_0, y_0)$ for all points (x, y) in the domain of f. The value of f at an absolute minimum point is the *minimum value* of f.

We use the term *extremum point* to refer to any point (x_0, y_0) at which f has a local maximum or minimum. In addition, the function value $f(x_0, y_0)$ at an extremum is called an *extremal value*. Figure 10.7.2 illustrates the graphs of two functions that have an absolute maximum and minimum, respectively, at the origin $(x_0, y_0) = (0, 0)$.

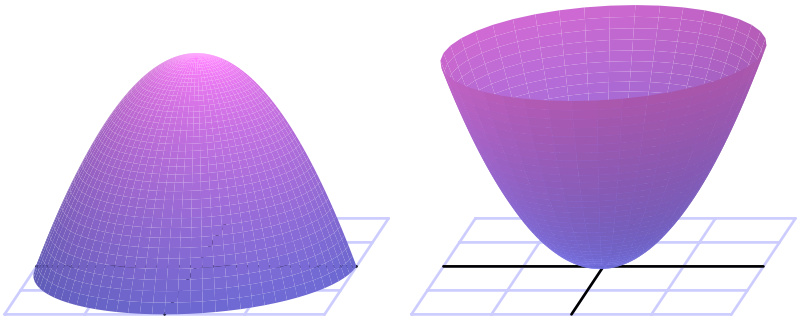


Figure 10.7.2. An absolute maximum and an absolute minimum

In single-variable calculus, we saw that the extrema of a continuous function f always occur at *critical points*, values of x where f fails to be differentiable or where $f'(x) = 0$. Said differently, critical points provide the locations where extrema of a function may appear. Our work in Preview Activity 10.7.1 suggests that something similar happens with two-variable functions.

Suppose that a continuous function f has an extremum at (x_0, y_0). In this case, the trace $f(x, y_0)$ has an extremum at x_0, which means that x_0 is a critical value of $f(x, y_0)$. Therefore, either $f_x(x_0, y_0)$ does not exist or $f_x(x_0, y_0) = 0$. Similarly, either $f_y(x_0, y_0)$ does not exist or $f_y(x_0, y_0) = 0$. This implies that the extrema of a two-variable function occur at points that satisfy the following definition.

Definition 10.7.2

A *critical point* (x_0, y_0) of a function $f = f(x, y)$ is a point in the domain of f at which $f_x(x_0, y_0) = 0$ and $f_y(x_0, y_0) = 0$, or such that one of $f_x(x_0, y_0)$ or $f_y(x_0, y_0)$ fails to exist.

We can therefore find critical points of a function f by computing partial derivatives and identifying any values of (x, y) for which one of the partials doesn't exist or for which both partial derivatives are simultaneously zero. For the latter, note that we have to solve the system of equations

$$f_x(x,y) = 0$$
$$f_y(x,y) = 0.$$

Activity 10.7.2

Find the critical points of each of the following functions. Then, using appropriate technology, plot the graphs of the surfaces near each critical point and compare the graph to your work.

1. $f(x,y) = 2 + x^2 + y^2$
2. $f(x,y) = 2 + x^2 - y^2$
3. $f(x,y) = 2x - x^2 - \frac{1}{4}y^2$
4. $f(x,y) = |x| + |y|$
5. $f(x,y) = 2xy - 4x + 2y - 3.$

Classifying Critical Points: The Second Derivative Test

While the extrema of a continuous function f always occur at critical points, it is important to note that not every critical point leads to an extremum. Recall, for instance, $f(x) = x^3$ from single variable calculus. We know that $x_0 = 0$ is a critical point since $f'(x_0) = 0$, but $x_0 = 0$ is neither a local maximum nor a local minimum of f.

A similar situation may arise in a multivariable setting. Consider the function f defined by $f(x,y) = x^2 - y^2$ whose graph and contour plot are shown in Figure 10.7.3. Because $\nabla f = \langle 2x, -2y \rangle$, we see that the origin $(x_0, y_0) = (0,0)$ is a critical point. However, this critical point is neither a local maximum or minimum; the origin is a local minimum on the trace defined by $y = 0$, while the origin is a local maximum on the trace defined by $x = 0$. We call such a critical point a *saddle point* due to the shape of the graph near the critical point.

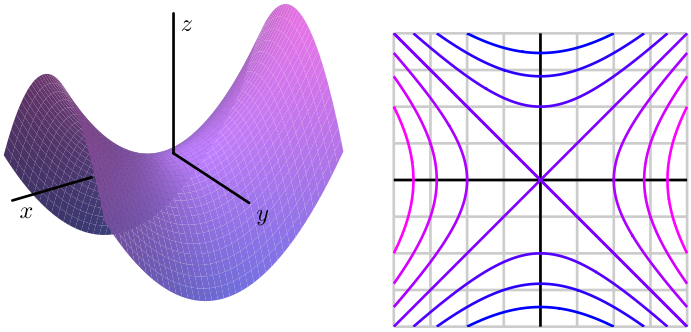


Figure 10.7.3. A saddle point.

As in single-variable calculus, we would like to have some sort of test to help us identify whether a critical point is a local maximum, local maximum, or neither.

Activity 10.7.3

Recall that the *Second Derivative Test* for single-variable functions states that if x_0 is a critical point of a function f so that $f'(x_0) = 0$ and if $f''(x_0)$ exists, then

- if $f''(x_0) < 0$, x_0 is a local maximum,
- if $f''(x_0) > 0$, x_0 is a local minimum, and
- if $f''(x_0) = 0$, this test yields no information.

Our goal in this activity is to understand a similar test for classifying extreme values of functions of two variables. Consider the following three functions:

$$f_1(x,y) = 4 - x^2 - y^2, \quad f_2(x,y) = x^2 + y^2, \quad f_3(x,y) = x^2 - y^2.$$

You can verify that each function has a critical point at the origin $(0,0)$. You should check this.

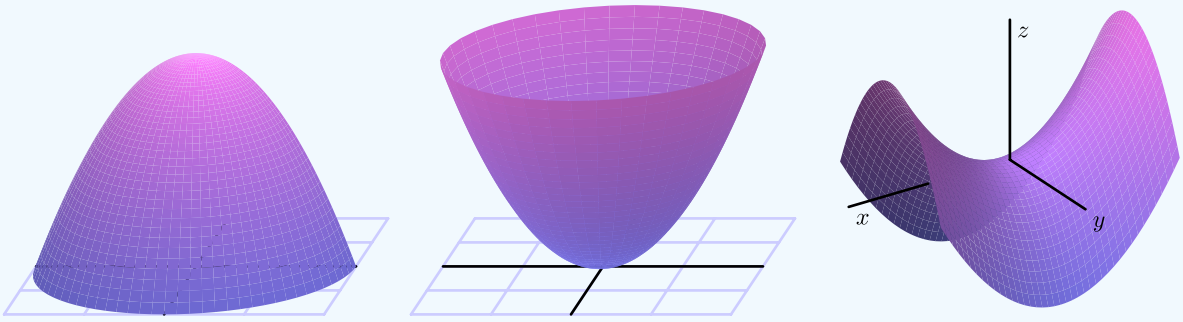


Figure 10.7.4. Three surfaces.

1. The graphs of these three functions are shown in Figure 10.7.4, with $z = 4 - x^2 - y^2$ at left, $z = x^2 + y^2$ in the middle, and $z = x^2 - y^2$ at right. Use the graphs to decide if a function has a local maximum, local minimum, saddle point, or none of the above at the origin.
2. There is no single second derivative of a function of two variables, so we consider a quantity that combines the second order partial derivatives. Let $D = f_{xx}f_{yy} - f_{xy}^2$. Calculate D at the origin for each of the functions f_1, f_2, and f_3. What difference do you notice between the values of D when a function has a maximum or minimum value at the origin versus when a function has a saddle point at the origin?
3. Now consider the cases where $D > 0$. It is in these cases that a function has a local maximum or minimum at a point. What is necessary in these cases is to find a condition that will distinguish between a maximum and a minimum. In the cases where $D > 0$ at the origin, evaluate $f_{xx}(0,0)$. What value does $f_{xx}(0,0)$ have when f has a local maximum value at the origin? When f has a local minimum value at the origin? Explain why. (Hint: This should look very similar to the Second Derivative Test for functions of a single variable.) What would happen if we considered the values of $f_{yy}(0,0)$ instead?

Activity 10.7.3 provides the basic ideas for the Second Derivative Test for functions of two variables.

The Second Derivative Test

Suppose (x_0, y_0) is a critical point of the function f for which $f_x(x_0, y_0) = 0$ and $f_y(x_0, y_0) = 0$. Let D be the quantity defined by

$$D = f_{xx}(x_0, y_0)f_{yy}(x_0, y_0) - f_{xy}(x_0, y_0)^2.$$

- If $D > 0$ and $f_{xx}(x_0, y_0) < 0$, then f has a local maximum at (x_0, y_0).
- If $D > 0$ and $f_{xx}(x_0, y_0) > 0$, then f has a local minimum at (x_0, y_0).
- If $D < 0$, then f has a saddle point at (x_0, y_0).
- If $D = 0$, then this test yields no information about what happens at (x_0, y_0).

The quantity D is called the *discriminant* of the function f at (x_0, y_0).

To properly understand the origin of the Second Derivative Test, we could introduce a "second-order directional derivative." If this second-order directional derivative were negative in every direction, for instance, we could guarantee that the critical point is a local maximum. A complete justification of the Second Derivative Test requires key ideas from linear algebra that are beyond the scope of this course, so instead of presenting a detailed explanation, we will accept this test as stated. In Activity 10.7.4, we apply

the test to more complicated examples.

Activity 10.7.4

Find the critical points of the following functions and use the Second Derivative Test to classify the critical points.

1. $f(x,y) = 3x^3 + y^2 - 9x + 4y$
2. $f(x,y) = xy + \frac{2}{x} + \frac{4}{y}$
3. $f(x,y) = x^3 + y^3 - 3xy.$

As we learned in single-variable calculus, finding extremal values of functions can be particularly useful in applied settings. For instance, we can often use calculus to determine the least expensive way to construct something or to find the most efficient route between two locations. The same possibility holds in settings with two or more variables.

Activity 10.7.5

While the quantity of a product demanded by consumers is often a function of the price of the product, the demand for a product may also depend on the price of other products. For instance, the demand for blue jeans at Old Navy may be affected not only by the price of the jeans themselves, but also by the price of khakis.

Suppose we have two goods whose respective prices are p_1 and p_2. The demand for these goods, q_1 and q_2, depend on the prices as

Extra close brace or missing open brace

The seller would like to set the prices p_1 and p_2 in order to maximize revenue. We will assume that the seller meets the full demand for each product. Thus, if we let R be the revenue obtained by selling q_1 items of the first good at price p_1 per item and q_2 items of the second good at price p_2 per item, we have

$$R = p_1 q_1 + p_2 q_2.$$

We can then write the revenue as a function of just the two variables p_1 and p_2 by using Equations 10.7.1 and 10.7.2, giving us

$$\begin{aligned} R(p_1, p_2) &= p_1(150 - 2p_1 - p_2) + p_2(200 - p_1 - 3p_2) \\ &= 150p_1 + 200p_2 - 2p_1p_2 - 2p_1^2 - 3p_2^2. \end{aligned}$$

A graph of R as a function of p_1 and p_2 is shown in Figure 10.7.5.

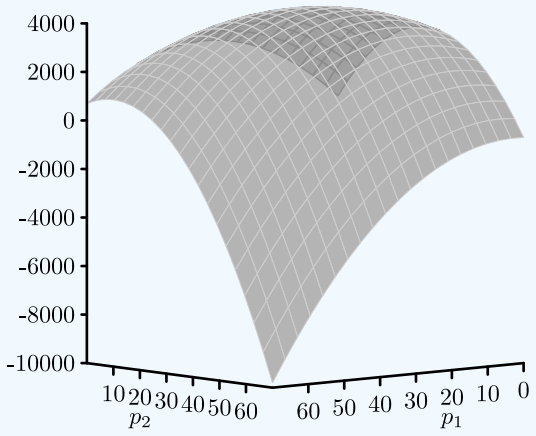


Figure 10.7.5. A revenue function.

1. Find all critical points of the revenue function, R. (Hint: You should obtain a system of two equations in two unknowns which can be solved by elimination or substitution.)
2. Apply the Second Derivative Test to determine the type of any critical point(s).
3. Where should the seller set the prices p_1 and p_2 to maximize the revenue?

Optimization on a Restricted Domain

The Second Derivative Test helps us classify critical points of a function, but it does not tell us if the function actually has an absolute maximum or minimum at each such point. For single-variable functions, the Extreme Value Theorem told us that a continuous function on a closed interval $[a, b]$ always has both an absolute maximum and minimum on that interval, and that these absolute extremes must occur at either an endpoint or at a critical point. Thus, to find the absolute maximum and minimum, we determine the critical points in the interval and then evaluate the function at these critical points and at the endpoints of the interval. A similar approach works for functions of two variables.

For functions of two variables, closed and bounded regions play the role that closed intervals did for functions of a single variable. A closed region is a region that contains its boundary (the unit disk $x^2 + y^2 \leq 1$ is closed, while its interior $x^2 + y^2 < 1$ is not, for example), while a bounded region is one that does not stretch to infinity in any direction. Just as for functions of a single variable, continuous functions of several variables that are defined on closed, bounded regions must have absolute maxima and minima in those regions.

The Extreme Value Theorem

Let $f = f(x, y)$ be a continuous function on a closed and bounded region R. Then f has an absolute maximum and an absolute minimum in R.

The absolute extremes must occur at either a critical point in the interior of R or at a boundary point of R. We therefore must test both possibilities, as we demonstrate in the following example.

Example 10.7.1

Suppose the temperature T at each point on the circular plate $x^2 + y^2 \leq 1$ is given by

$$T(x, y) = 2x^2 + y^2 - y.$$

The domain $R = \{(x, y) : x^2 + y^2 \leq 1\}$ is a closed and bounded region, as shown on the left of Figure 10.7.9, so the Extreme Value Theorem assures us that T has an absolute maximum and minimum on the plate. The graph of T over its domain R is shown in Figure 10.7.6. We will find the hottest and coldest points on the plate.

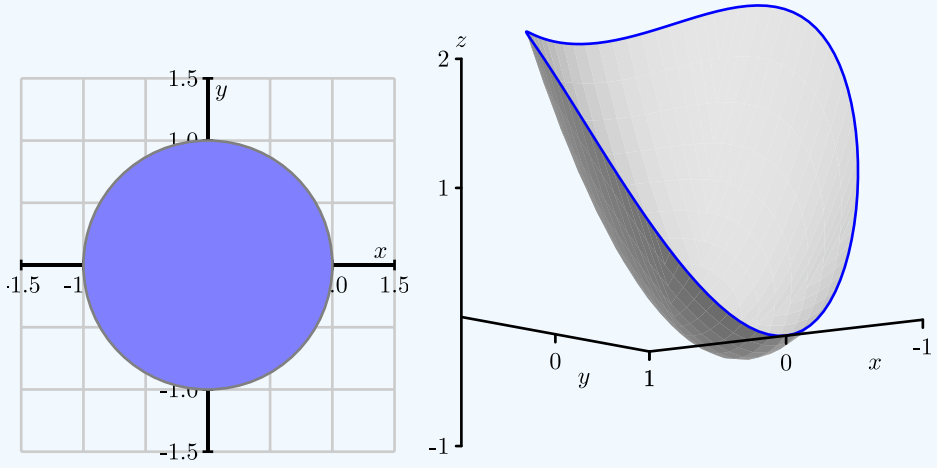


Figure 10.7.6. Domain of the temperature $T(x,y) = 2x^2 + y^2 - y$ and its graph.

If the absolute maximum or minimum occurs inside the disk, it will be at a critical point so we begin by looking for critical points inside the disk. To do this, notice that critical points are given by the conditions $T_x = 4x = 0$ and $T_y = 2y - 1 = 0$. This means that there is one critical point of the function at the point $(x_0, y_0) = (0, 1/2)$, which lies inside the disk.

We now find the hottest and coldest points on the boundary of the disk, which is the circle of radius 1. As we have seen, the points on the unit circle can be parametrized as

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

where $0 \leq t \leq 2\pi$. The temperature at a point on the circle is then described by

$$T(x(t), y(t)) = 2\cos^2(t) + \sin^2(t) - \sin(t).$$

To find the hottest and coldest points on the boundary, we look for the critical points of this single-variable function on the interval $0 \leq t \leq 2\pi$. We have

$$\begin{aligned} \frac{dT}{dt} &= -4\cos(t)\sin(t) + 2\cos(t)\sin(t) - \cos(t) \\ &= -2\cos(t)\sin(t) - \cos(t) = \cos(t)(-2\sin(t) - 1) \\ &= 0. \end{aligned}$$

This shows that we have critical points when $\cos(t) = 0$ or $\sin(t) = -1/2$. This occurs when $t = \pi/2,\ 3\pi/2,\ 7\pi/6$, and $11\pi/6$. Since we have $x(t) = \cos(t)$ and $y(t) = \sin(t)$, the corresponding points are

- $(x,y) = (0,1)$ when $t = \frac{\pi}{2}$,
- $(x,y) = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)$ when $t = \frac{11\pi}{6}$.

- $(x,y) = (0,-1)$ when $t = \frac{3\pi}{2}$,
- $(x,y) = \left(-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)$ when $t = \frac{7\pi}{6}$.

These are the critical points of T on the boundary and so this collection of points includes the hottest and coldest points on the boundary.

We now have a list of candidates for the hottest and coldest points: the critical point in the interior of the disk and the critical points on the boundary. We find the hottest and coldest points by evaluating the temperature at each of these points, and find that

- $T\left(0, \frac{1}{2}\right) = -\frac{1}{4}$,
- $T\left(-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) = \frac{9}{4}$,

- $T(0,1) = 0$,
- $T\left(-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) = \frac{9}{4}$.

- $T(0,-1) = 2$,

So the maximum value of T on the disk $x^2 + y^2 \leq 1$ is $\frac{9}{4}$, which occurs at the two points $\left(\pm\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)$ on the boundary, and the minimum value of T on the disk is $-\frac{1}{4}$ which occurs at the critical point $\left(0, \frac{1}{2}\right)$ in the interior of R.

From this example, we see that we use the following procedure for determining the absolute maximum and absolute minimum of a function on a closed and bounded domain.

- Find all critical points of the function in the interior of the domain.
- Find all the critical points of the function on the boundary of the domain. Working on the boundary of the domain reduces this part of the problem to one or more single variable optimization problems. Note that there may be endpoints on portions of the boundary that need to be considered.
- Evaluate the function at each of the points found in Steps 1 and 2.
- The maximum value of the function is the largest value obtained in Step 3, and the minimum value of the function is the smallest value obtained in Step 3.

Activity 10.7.6

Let $f(x,y) = x^2 - 3y^2 - 4x + 6y$ with triangular domain R whose vertices are at $(0,0)$, $(4,0)$, and $(0,4)$. The domain R and a graph of f on the domain appear in Figure 10.7.7.

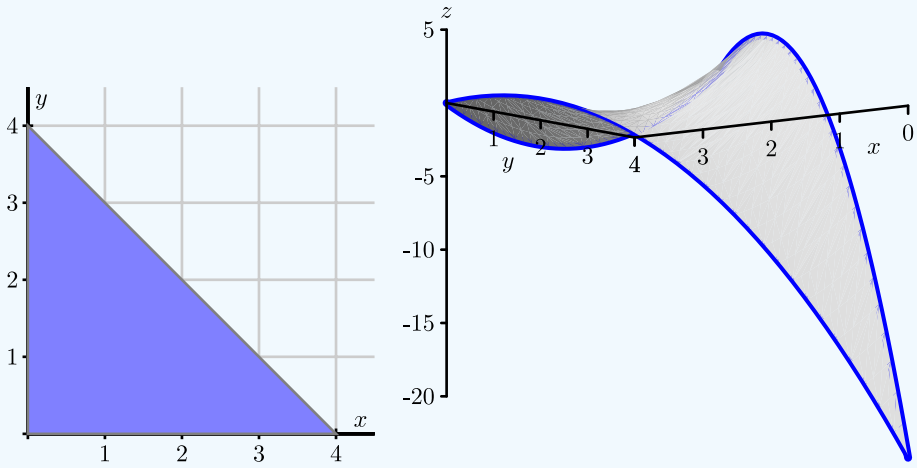


Figure 10.7.7. The domain of $f(x,y) = x^2 - 3y^2 - 4x + 6y$ and its graph.

1. Find all of the critical points of f in R.
2. Parameterize the horizontal leg of the triangular domain, and find the critical points of f on that leg. (Hint: You may need to consider endpoints.)
3. Parameterize the vertical leg of the triangular domain, and find the critical points of f on that leg. (Hint: You may need to consider endpoints.)
4. Parameterize the hypotenuse of the triangular domain, and find the critical points of f on the hypotenuse. (Hint: You may need to consider endpoints.)
5. Find the absolute maximum and absolute minimum values of f on R.

Summary

- To find the extrema of a function $f = f(x,y)$, we first find the critical points, which are points where one of the partials of f fails to exist, or where $f_x = 0$ and $f_y = 0$.
- The Second Derivative Test helps determine whether a critical point is a local maximum, local minimum, or saddle point.
- If f is defined on a closed and bounded domain, we find the absolute maxima and minima by finding the critical points in the interior of the domain, finding the critical points on the boundary, and testing the value of f at both sets of critical points.

This page titled 10.7: Optimization is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.7: Optimization by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

10.8: Constrained Optimization - Lagrange Multipliers

Motivating Questions

- What geometric condition enables us to optimize a function $f = f(x, y)$ subject to a constraint given by $g(x, y) = k$, where k is a constant?
- How can we exploit this geometric condition to find the extreme values of a function subject to a constraint?

We previously considered how to find the extreme values of functions on both unrestricted domains and on closed, bounded domains. Other types of optimization problems involve maximizing or minimizing a quantity subject to an external constraint. In these cases the extreme values frequently won't occur at the points where the gradient is zero, but rather at other points that satisfy an important geometric condition. These problems are often called *constrained optimization* problems and can be solved with the method of Lagrange Multipliers, which we study in this section.

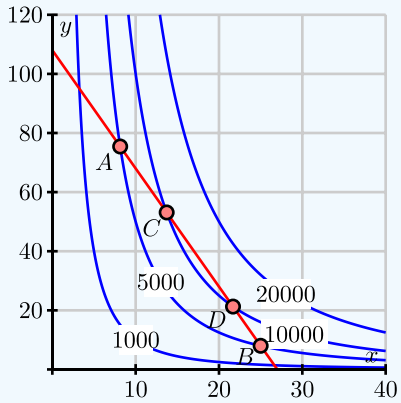
Preview Activity 10.8.1

According to U.S. postal regulations, the girth plus the length of a parcel sent by mail may not exceed 108 inches, where by "girth" we mean the perimeter of the smallest end. Our goal is to find the largest possible volume of a rectangular parcel with a square end that can be sent by mail. (We solved this applied optimization problem in single variable *Active Calculus*, so it may look familiar. We take a different approach in this section, and this approach allows us to view most applied optimization problems from single variable calculus as constrained optimization problems, as well as provide us tools to solve a greater variety of optimization problems.) If we let x be the length of the side of one square end of the package and y the length of the package, then we want to maximize the volume $f(x, y) = x^2 y$ of the box subject to the constraint that the girth $(4x)$ plus the length (y) is as large as possible, or $4x + y = 108$. The equation $4x + y = 108$ is thus an external constraint on the variables.

1. The constraint equation involves the function g that is given by

$$g(x, y) = 4x + y.$$

   Explain why the constraint is a contour of g, and is therefore a two-dimensional curve.



1.

Figure 10.8.1. Contours of f and the constraint equation $g(x, y) = 108$.

2. Figure 10.8.1 shows the graph of the constraint equation $g(x, y) = 108$ along with a few contours of the volume function f. Since our goal is to find the maximum value of f subject to the constraint $g(x, y) = 108$, we want to find the point on our constraint curve that intersects the contours of f at which f has its largest value.
   1. Points A and B in Figure 10.8.1 lie on a contour of f and on the constraint equation $g(x, y) = 108$. Explain why neither A nor B provides a maximum value of f that satisfies the constraint.

2. Points C and D in Figure 10.8.1 lie on a contour of f and on the constraint equation $g(x, y) = 108$. Explain why neither C nor D provides a maximum value of f that satisfies the constraint.
3. Based on your responses to parts i. and ii., draw the contour of f on which you believe f will achieve a maximum value subject to the constraint $g(x, y) = 108$. Explain why you drew the contour you did.

3. Recall that $g(x, y) = 108$ is a contour of the function g, and that the gradient of a function is always orthogonal to its contours. With this in mind, how should ∇f and ∇g be related at the optimal point? Explain.

Constrained Optimization and Lagrange Multipliers

In Preview Activity 10.8.1, we considered an optimization problem where there is an external constraint on the variables, namely that the girth plus the length of the package cannot exceed 108 inches. We saw that we can create a function g from the constraint, specifically $g(x, y) = 4x + y$. The constraint equation is then just a contour of g, $g(x, y) = c$, where c is a constant (in our case 108). Figure 10.8.2 illustrates that the volume function f is maximized, subject to the constraint $g(x, y) = c$, when the graph of $g(x, y) = c$ is tangent to a contour of f. Moreover, the value of f on this contour is the sought maximum value.

Figure 10.8.2. Contours of f and the constraint contour.

To find this point where the graph of the constraint is tangent to a contour of f, recall that ∇f is perpendicular to the contours of f and ∇g is perpendicular to the contour of g. At such a point, the vectors ∇g and ∇f are parallel, and thus we need to determine the points where this occurs. Recall that two vectors are parallel if one is a nonzero scalar multiple of the other, so we therefore look for values of a parameter λ that make

$$\nabla f = \lambda \nabla g. \tag{10.8.1}$$

The constant λ is called a *Lagrange multiplier*.

To find the values of λ that satisfy Equation 10.8.1 for the volume function in Preview Activity 10.8.1, we calculate both ∇f and ∇g. Observe that

$$\nabla f = 2xy\mathbf{i} + x^2\mathbf{j} \quad \text{and} \quad \nabla g = 4\mathbf{i} + \mathbf{j},$$

and thus we need a value of λ so that

$$2xy\mathbf{i} + x^2\mathbf{j} = \lambda(4\mathbf{i} + \mathbf{j}).$$

Equating components in the most recent equation and incorporating the original constraint, we have three equations

$$2xy = \lambda(4) \tag{10.8.2}$$
$$x^2 = \lambda(1) \tag{10.8.3}$$
$$4x + y = 108 \tag{10.8.4}$$

in the three unknowns x, y, and λ. First, note that if $\lambda = 0$, then Equation 10.8.3 shows that $x = 0$. From this, Equation 10.8.4 tells us that $y = 108$. So the point $(0, 108)$ is a point we need to consider. Next, provided that $\lambda \neq 0$ (from which it follows that $x \neq 0$ by Equation 10.8.3, we may divide both sides of Equation 10.8.2 by the corresponding sides of Equation 10.8.3 to eliminate λ, and thus find that

$$\frac{2y}{x} = 4, \text{ so}$$
$$y = 2x.$$

Substituting into Equation 10.8.4 gives us

$$4x + 2x = 108$$

or

$$x = 18.$$

Thus we have $y = 2x = 36$ and $\lambda = x^2 = 324$ as another point to consider. So the points at which the gradients of f and g are parallel, and thus at which f may have a maximum or minimum subject to the constraint, are $(0, 108)$ and $(18, 36)$. By evaluating the function f at these points, we see that we maximize the volume when the length of the square end of the box is 18 inches and the length is 36 inches, for a maximum volume of $f(18, 36) = 11664$ cubic inches. Since $f(0, 108) = 0$, we obtain a minimum value at this point.

We summarize the process of Lagrange multipliers as follows.

The method of Lagrange multipliers

The general technique for optimizing a function $f = f(x, y)$ subject to a constraint $g(x, y) = c$ is to solve the system $\nabla f = \lambda \nabla g$ and $g(x, y) = c$ for x, y, and λ. We then evaluate the function f at each point (x, y) that results from a solution to the system in order to find the optimum values of f subject to the constraint.

Activity 10.8.2

A cylindrical soda can holds about 355 cc of liquid. In this activity, we want to find the dimensions of such a can that will minimize the surface area. For the sake of simplicity, assume the can is a perfect cylinder.

1. What are the variables in this problem? Based on the context, what restriction(s), if any, are there on these variables?
2. What quantity do we want to optimize in this problem? What equation describes the constraint? (You need to decide which of these functions plays the role of f and which plays the role of g in our discussion of Lagrange multipliers.)
3. Find λ and the values of your variables that satisfy Equation 10.8.1 in the context of this problem.
4. Determine the dimensions of the pop can that give the desired solution to this constrained optimization problem.

The method of Lagrange multipliers also works for functions of more than two variables.

Activity 10.8.3

Use the method of Lagrange multipliers to find the dimensions of the least expensive packing crate with a volume of 240 cubic feet when the material for the top costs \$2 per square foot, the bottom is \$3 per square foot and the sides are \$1.50 per square foot.

The method of Lagrange multipliers also works for functions of three variables. That is, if we have a function $f = f(x, y, z)$ that we want to optimize subject to a constraint $g(x, y, z) = k$, the optimal point (x, y, z) lies on the level surface S defined by the constraint $g(x, y, z) = k$. As we did in Preview Activity 10.8.1, we can argue that the optimal value occurs at the level surface $f(x, y, z) = c$ that is tangent to S. Thus, the gradients of f and g are parallel at this optimal point. So, just as in the two variable case, we can optimize $f = f(x, y, z)$ subject to the constraint $g(x, y, z) = k$ by finding all points (x, y, z) that satisfy $\nabla f = \lambda \nabla g$ and $g(x, y, z) = k$.

Summary

- The extrema of a function $f = f(x, y)$ subject to a constraint $g(x, y) = c$ occur at points for which the contour of f is tangent to the curve that represents the constraint equation. This occurs when
$$\nabla f = \lambda \nabla g.$$
- We use the condition $\nabla f = \lambda \nabla g$ to generate a system of equations, together with the constraint $g(x, y) = c$, that may be solved for x, y, and λ. Once we have all the solutions, we evaluate f at each of the (x, y) points to determine the extrema.

This page titled 10.8: Constrained Optimization - Lagrange Multipliers is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.8: Constrained Optimization - Lagrange Multipliers by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

11: MULTIPLE INTEGRALS

http://scholarworks.gvsu.edu/books/14/

11.1: Double Riemann Sums and Double Integrals over Rectangles

11.2: Iterated Integrals

11.3: Double Integrals over General Regions

11.4: Applications of Double Integrals

11.5: Double Integrals in Polar Coordinates

11.6: Surfaces Defined Parametrically and Surface Area

11.7: Triple Integrals

11.8: Triple Integrals in Cylindrical and Spherical Coordinates

11.9: Change of Variables

This page titled 11: Multiple Integrals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 11: Multiple Integrals by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

11.1: Double Riemann Sums and Double Integrals over Rectangles

Motivating Questions

- What is a double Riemann sum?
- How is the double integral of a continuous function $f = f(x, y)$ defined?
- What are two things the double integral of a function can tell us?

In single-variable calculus, recall that we approximated the area under the graph of a positive function f on an interval $[a, b]$ by adding areas of rectangles whose heights are determined by the curve. The general process involved subdividing the interval $[a, b]$ into smaller subintervals, constructing rectangles on each of these smaller intervals to approximate the region under the curve on that subinterval, then summing the areas of these rectangles to approximate the area under the curve. We will extend this process in this section to its three-dimensional analogs, double Riemann sums and double integrals over rectangles.

Preview Activity 11.1.1

In this activity we introduce the concept of a double Riemann sum.

1. Review the concept of the Riemann sum from single-variable calculus. Then, explain how we define the definite integral $\int_a^b f(x)\ dx$ of a continuous function of a single variable x on an interval $[a, b]$. Include a sketch of a continuous function on an interval $[a, b]$ with appropriate labeling in order to illustrate your definition.
2. In our upcoming study of integral calculus for multivariable functions, we will first extend the idea of the single-variable definite integral to functions of two variables over rectangular domains. To do so, we will need to understand how to partition a rectangle into subrectangles. Let R be rectangular domain $R = \{(x, y) : 0 \leq x \leq 6, 2 \leq y \leq 4\}$ (we can also represent this domain with the notation $[0, 6] \times [2, 4]$), as pictured in Figure 11.1.1.

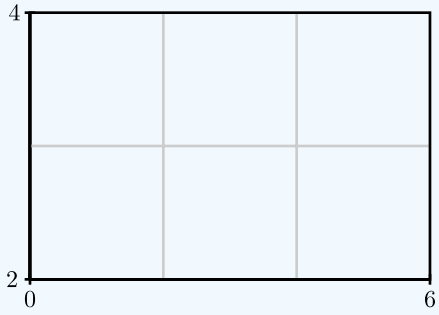


Figure 11.1.1. Rectangular domain R with subrectangles.

To form a partition of the full rectangular region, R, we will partition both intervals $[0, 6]$ and $[2, 4]$; in particular, we choose to partition the interval $[0, 6]$ into three uniformly sized subintervals and the interval $[2, 4]$ into two evenly sized subintervals as shown in Figure 11.1.1. In the following questions, we discuss how to identify the endpoints of each subinterval and the resulting subrectangles.

1. Let $0 = x_0 < x_1 < x_2 < x_3 = 6$ be the endpoints of the subintervals of $[0, 6]$ after partitioning. What is the length Δx of each subinterval $[x_{i-1}, x_i]$ for i from 1 to 3?
2. Explicitly identify x_0, x_1, x_2, and x_3. On Figure 11.1.1 or your own version of the diagram, label these endpoints.
3. Let $2 = y_0 < y_1 < y_2 = 4$ be the endpoints of the subintervals of $[2, 4]$ after partitioning. What is the length Δy of each subinterval $[y_{j-1}, y_j]$ for j from 1 to 2? Identify y_0, y_1, and y_2 and label these endpoints on Figure 11.1.1.
4. Let R_{ij} denote the subrectangle $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$. Appropriately label each subrectangle in your drawing of Figure 11.1.1. How does the total number of subrectangles depend on the partitions of the intervals $[0, 6]$ and $[2, 4]$?
5. What is area ΔA of each subrectangle?

Double Riemann Sums over Rectangles

For the definite integral in single-variable calculus, we considered a continuous function over a closed, bounded interval $[a, b]$. In multivariable calculus, we will eventually develop the idea of a definite integral over a closed, bounded region (such as the interior of a circle). We begin with a simpler situation by thinking only about rectangular domains, and will address more complicated domains in Section 11.3.

Let $f = f(x, y)$ be a continuous function defined on a rectangular domain $R = \{(x, y) : a \leq x \leq b, c \leq y \leq d\}$. As we saw in Preview Activity 11.1.1, the domain is a rectangle R and we want to partition R into subrectangles. We do this by partitioning each of the intervals $[a, b]$ and $[c, d]$ into subintervals and using those subintervals to create a partition of R into subrectangles. In the first activity, we address the quantities and notations we will use in order to define double Riemann sums and double integrals.

Activity 11.1.2

Let $f(x, y) = 100 - x^2 - y^2$ be defined on the rectangular domain $R = [a, b] \times [c, d]$. Partition the interval $[a, b]$ into four uniformly sized subintervals and the interval $[c, d]$ into three evenly sized subintervals as shown in Figure 11.1.2. As we did in Preview Activity 11.1.1, we will need a method for identifying the endpoints of each subinterval and the resulting subrectangles.

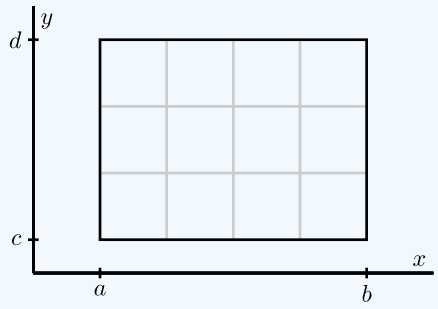


Figure 11.1.2. Rectangular domain with subrectangles.

1. Let $a = x_0 < x_1 < x_2 < x_3 < x_4 = b$ be the endpoints of the subintervals of $[a, b]$ after partitioning. Label these endpoints in Figure 11.1.2.
2. What is the length Δx of each subinterval $[x_{i-1}, x_i]$? Your answer should be in terms of a and b.
3. Let $c = y_0 < y_1 < y_2 < y_3 = d$ be the endpoints of the subintervals of $[c, d]$ after partitioning. Label these endpoints in Figure 11.1.2.
4. What is the length Δy of each subinterval $[y_{j-1}, y_j]$? Your answer should be in terms of c and d.
5. The partitions of the intervals $[a, b]$ and $[c, d]$ partition the rectangle R into subrectangles. How many subrectangles are there?
6. Let R_{ij} denote the subrectangle $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$. Label each subrectangle in Figure 11.1.2.
7. What is area ΔA of each subrectangle?
8. Now let $[a, b] = [0, 8]$ and $[c, d] = [2, 6]$. Let (x^*_{11}, y^*_{11}) be the point in the upper right corner of the subrectangle R_{11}. Identify and correctly label this point in Figure 11.1.2. Calculate the product

$$f(x^*_{11}, y^*_{11})\Delta A.$$

   Explain, geometrically, what this product represents.

9. For each i and j, choose a point (x^*_{ij}, y^*_{ij}) in the subrectangle $R_{i,j}$. Identify and correctly label these points in Figure 11.1.2 . Explain what the product

$$f(x^*_{ij}, y^*_{ij})\Delta A$$

   represents.

10. If we were to add all the values $f(x_{ij}^*, y_{ij}^*)\Delta A$ for each i and j, what does the resulting number approximate about the surface defined by f on the domain R? (You don't actually need to add these values.)
11. Write a double sum using summation notation that expresses the arbitrary sum from part (j).

Double Riemann Sums and Double Integrals

Now we use the process from the most recent activity to formally define double Riemann sums and double integrals.

Definition 11.1.1

Let f be a continuous function on a rectangle $R = \{(x, y) : a \le x \le b, c \le y \le d\}$. A *double Riemann sum for f over R* is created as follows.

- Partition the interval $[a, b]$ into m subintervals of equal length $\Delta x = \frac{b-a}{m}$. Let $x_0, x_1, \ldots, x_m$ be the endpoints of these subintervals, where $a = x_0 < x_1 < x_2 < \cdots < x_m = b$.
- Partition the interval $[c, d]$ into n subintervals of equal length $\Delta y = \frac{d-c}{n}$. Let $y_0, y_1, \ldots, y_n$ be the endpoints of these subintervals, where $c = y_0 < y_1 < y_2 < \cdots < y_n = d$.
- These two partitions create a partition of the rectangle R into mn subrectangles R_{ij} with opposite vertices (x_{i-1}, y_{j-1}) and (x_i, y_j) for i between 1 and m and j between 1 and n. These rectangles all have equal area $\Delta A = \Delta x \cdot \Delta y$.
- Choose a point (x_{ij}^*, y_{ij}^*) in each rectangle R_{ij}. Then, a double Riemann sum for f over R is given by

$$\sum_{j=1}^{n}\sum_{i=1}^{m} f(x_{ij}^*, y_{ij}^*) \cdot \Delta A.$$

If $f(x, y) \ge 0$ on the rectangle R, we may ask to find the volume of the solid bounded above by f over R, as illustrated on the left of Figure 11.1.3. This volume is approximated by a Riemann sum, which sums the volumes of the rectangular boxes shown on the right of Figure 11.1.3.

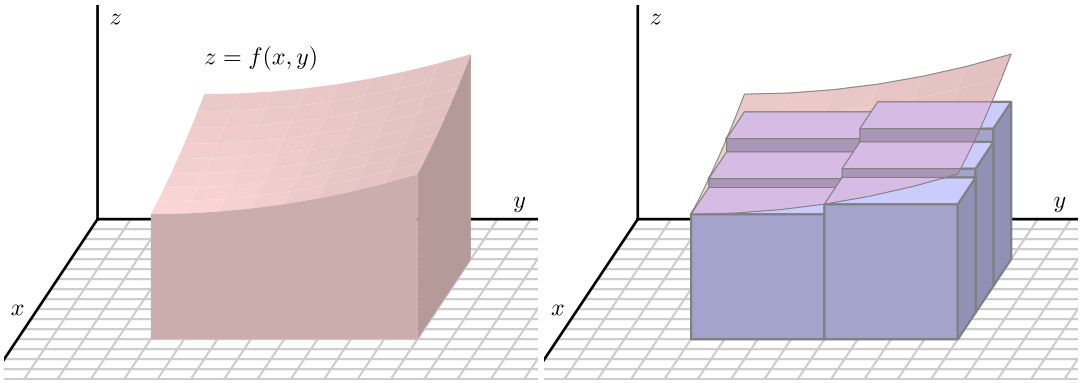


Figure 11.1.3. The volume under a graph approximated by a Riemann Sum.

As we let the number of subrectangles increase without bound (in other words, as both m and n in a double Riemann sum go to infinity), as illustrated in Figure 11.1.4, the sum of the volumes of the rectangular boxes approaches the volume of the solid bounded above by f over R. The value of this limit, provided it exists, is the double integral.

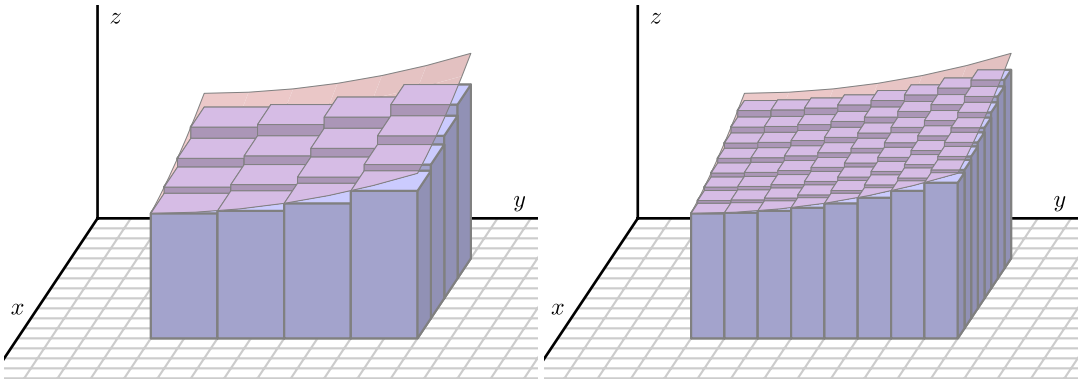


Figure 11.1.4. Finding better approximations by using smaller subrectangles.

Definition 11.1.2

Let R be a rectangular region in the xy-plane and f a continuous function over R. With terms defined as in a double Riemann sum, the *double integral of f over R* is

$$\iint_R f(x,y)\,dA = \lim_{m,n\to\infty} \sum_{j=1}^{n}\sum_{i=1}^{m} f(x_{ij}^*, y_{ij}^*)\cdot \Delta A.$$

Some textbooks use the notation $\int_R f(x,y)\ dA$ for a double integral. You will see this in some of the WeBWorK problems.

Interpretation of Double Riemann Sums and Double integrals.

At the moment, there are two ways we can interpret the value of the double integral.

- Suppose that $f(x,y)$ assumes both positive and negatives values on the rectangle R, as shown on the left of Figure 11.1.5. When constructing a Riemann sum, for each i and j, the product $f(x_{ij}^*, y_{ij}^*)\cdot\Delta A$ can be interpreted as a "signed" volume of a box with base area ΔA and "signed" height $f(x_{ij}^*, y_{ij}^*)$. Since f can have negative values, this "height" could be negative. The sum

  $$\sum_{j=1}^{n}\sum_{i=1}^{m} f(x_{ij}^*, y_{ij}^*)\cdot\Delta A$$

  can then be interpreted as a sum of "signed" volumes of boxes, with a negative sign attached to those boxes whose heights are below the xy-plane.

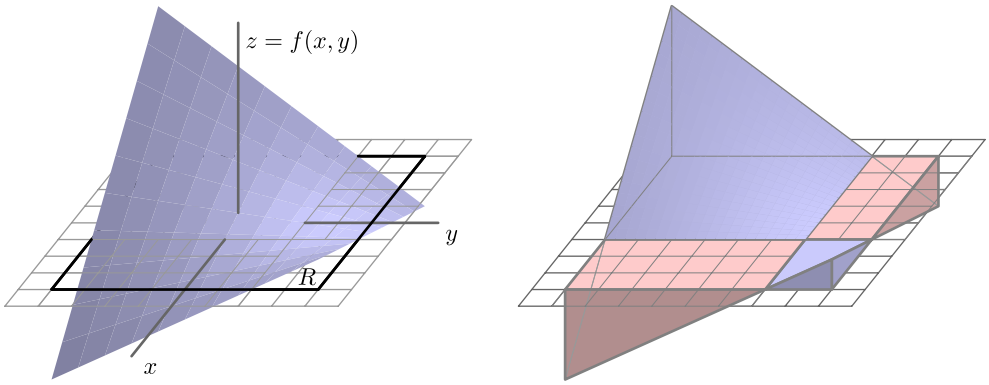


Figure 11.1.5. The integral measures signed volume.

We can then realize the double integral $\iint_R f(x,y)\ dA$ as a difference in volumes: $\iint_R f(x,y)\ dA$ tells us the volume of the solids the graph of f bounds above the xy-plane over the rectangle R minus the volume of the solids the graph of f bounds below the xy-plane under the rectangle R. This is shown on the right of Figure 11.1.5.

- The average of the finitely many mn values $f(x_{ij}^*, y_{ij}^*)$ that we take in a double Riemann sum is given by

$$\text{Avg}_{mn} = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m} f(x_{ij}^*, y_{ij}^*).$$

If we take the limit as m and n go to infinity, we obtain what we define as the average value of f over the region R, which is connected to the value of the double integral. First, to view Avg_{mn} as a double Riemann sum, note that

$$\Delta x = \frac{b-a}{m} \qquad \text{and} \qquad \Delta y = \frac{d-c}{n}.$$

Thus,

$$\frac{1}{mn} = \frac{\Delta x \cdot \Delta y}{(b-a)(d-c)} = \frac{\Delta A}{A(R)},$$

where $A(R)$ denotes the area of the rectangle R. Then, the average value of the function f over R, $f_{\text{AVG}(R)}$, is given by

$$\begin{aligned} f_{\text{AVG}(R)} &= \lim_{m,n\to\infty} \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m} f(x_{ij}^*, y_{ij}^*) \\ &= \lim_{m,n\to\infty} \frac{1}{A(R)}\sum_{j=1}^{n}\sum_{i=1}^{m} f(x_{ij}^*, y_{ij}^*)\cdot \Delta A \\ &= \frac{1}{A(R)}\iint_R f(x,y)\,dA. \end{aligned}$$

Therefore, the double integral of f over R divided by the area of R gives us the average value of the function f on R. Finally, if $f(x,y) \geq 0$ on R, we can interpret this average value of f on R as the height of the box with base R that has the same volume as the volume of the surface defined by f over R.

Activity 11.1.3

Let $f(x,y) = x + 2y$ and let $R = [0,2] \times [1,3]$.

1. Draw a picture of R. Partition $[0,2]$ into 2 subintervals of equal length and the interval $[1,3]$ into two subintervals of equal length. Draw these partitions on your picture of R and label the resulting subrectangles using the labeling scheme we established in the definition of a double Riemann sum.
2. For each i and j, let (x_{ij}^*, y_{ij}^*) be the midpoint of the rectangle R_{ij}. Identify the coordinates of each (x_{ij}^*, y_{ij}^*). Draw these points on your picture of R.
3. Calculate the Riemann sum

$$\sum_{j=1}^{n}\sum_{i=1}^{m} f(x_{ij}^*, y_{ij}^*)\cdot \Delta A$$

using the partitions we have described. If we let (x_{ij}^*, y_{ij}^*) be the midpoint of the rectangle R_{ij} for each i and j, then the resulting Riemann sum is called a *midpoint sum*.

4. Give two interpretations for the meaning of the sum you just calculated.

Activity 11.1.4

Let $f(x,y) = \sqrt{4-y^2}$ on the rectangular domain $R = [1,7] \times [-2,2]$. Partition $[1,7]$ into 3 equal length subintervals and $[-2,2]$ into 2 equal length subintervals. A table of values of f at some points in R is given in Table 11.1.1, and a graph of f with the indicated partitions is shown in Figure 11.1.6.

Table 11.1.1. *Table of values of* $f(x,y)=\sqrt{4-y^2}$.

	-2	-1	0	1	2
1	0	$\sqrt{3}$	2	$\sqrt{3}$	0
2	0	$\sqrt{3}$	2	$\sqrt{3}$	0
3	0	$\sqrt{3}$	2	$\sqrt{3}$	0
4	0	$\sqrt{3}$	2	$\sqrt{3}$	0
5	0	$\sqrt{3}$	2	$\sqrt{3}$	0
6	0	$\sqrt{3}$	2	$\sqrt{3}$	0
7	0	$\sqrt{3}$	2	$\sqrt{3}$	0

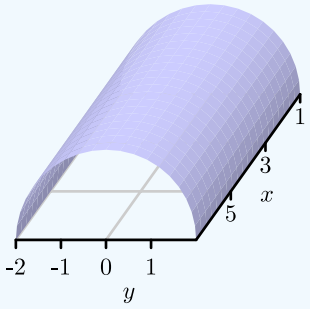


Figure 11.1.6. Graph of $f(x,y)=\sqrt{4-y^2}$ on R.

1. Sketch the region R in the plane using the values in Table 11.1.1 as the partitions.
2. Calculate the double Riemann sum using the given partition of R and the values of f in the upper right corner of each subrectangle.
3. Use geometry to calculate the exact value of $\iint_R f(x,y)\ dA$ and compare it to your approximation. Describe one way we could obtain a better approximation using the given data.

We conclude this section with a list of properties of double integrals. Since similar properties are satisfied by single-variable integrals and the arguments for double integrals are essentially the same, we omit their justification.

Properties of Double Integrals

Let f and g be continuous functions on a rectangle $R=\{(x,y): a\le x\le b, c\le y\le d\}$, and let k be a constant. Then

1. $\iint_R (f(x,y)+g(x,y))\ dA=\iint_R f(x,y)\ dA+\iint_R g(x,y)\ dA$.
2. $\iint_R kf(x,y)\ dA=k\iint_R f(x,y)\ dA$.
3. If $f(x,y)\ge g(x,y)$ on R, then $\iint_R f(x,y)\ dA\ge \iint_R g(x,y)\ dA$.

Summary

- Let f be a continuous function on a rectangle $R=\{(x,y): a\le x\le b, c\le y\le d\}$. The double Riemann sum for f over R is created as follows.
  - Partition the interval $[a,b]$ into m subintervals of equal length $\Delta x=\frac{b-a}{m}$. Let $x_0,\ x_1,\ \ldots,\ x_m$ be the endpoints of these subintervals, where $a=x_0<x_1<x_2<\cdots<x_m=b$.
  - Partition the interval $[c,d]$ into n subintervals of equal length $\Delta y=\frac{d-c}{n}$. Let $y_0,\ y_1,\ \ldots,\ y_n$ be the endpoints of these subintervals, where $c=y_0<y_1<y_2<\cdots<y_n=d$.
  - These two partitions create a partition of the rectangle R into mn subrectangles R_{ij} with opposite vertices (x_{i-1},y_{j-1}) and (x_i,y_j) for i between 1 and m and j between 1 and n. These rectangles all have equal area $\Delta A=\Delta x\cdot\Delta y$.
  - Choose a point (x_{ij}^*,y_{ij}^*) in each rectangle R_{ij}. Then a double Riemann sum for f over R is given by

$$\sum_{j=1}^{n}\sum_{i=1}^{m} f(x_{ij}^*, y_{ij}^*)\cdot \Delta A.$$

- With terms defined as in the Double Riemann Sum, the double integral of f over R is

$$\iint_R f(x,y)\,dA = \lim_{m,n\to\infty}\sum_{j=1}^{n}\sum_{i=1}^{m} f(x_{ij}^*, y_{ij}^*)\cdot \Delta A.$$

- Two interpretations of the double integral $\iint_R f(x,y)\ dA$ are:
  - The volume of the solids the graph of f bounds above the xy-plane over the rectangle R minus the volume of the solids the graph of f bounds below the xy-plane under the rectangle R;
  - Dividing the double integral of f over R by the area of R gives us the average value of the function f on R. If $f(x,y)\geq 0$ on R, we can interpret this average value of f on R as the height of the box with base R that has the same volume as the volume of the surface defined by f over R.

This page titled 11.1: Double Riemann Sums and Double Integrals over Rectangles is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.1: Double Riemann Sums and Double Integrals over Rectangles by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

11.2: Iterated Integrals

Motivating Questions

- How do we evaluate a double integral over a rectangle as an iterated integral, and why does this process work?

Recall that we defined the double integral of a continous function $f = f(x, y)$ over a rectanlge $R = [a, b] \times [c, d]$ as

$$\iint_R f(x, y)\, dA = \lim_{m,n\to\infty} \sum_{j=1}^{n} \sum_{i=1}^{m} f\left(x_{ij}^*, y_{ij}^*\right) \cdot \Delta A,$$

where the different variables and notation are as described in Section 11.1. Thus $\iint_R f(x, y)\ dA$ is a limit of double Riemann sums, but while this definition tells us exactly what a double integral is, it is not very helpful for determining the value of a double integral. Fortunately, there is a way to view a double integral as an *iterated integral*, which will make computations feasible in many cases.

The viewpoint of an iterated integral is closely connected to an important idea from single-variable calculus. When we studied solids of revolution, such as the one shown in Figure 11.2.1, we saw that in some circumstances we could slice the solid perpendicular to an axis and have each slice be approximately a circular disk. From there, we were able to find the volume of each disk, and then use an integral to add the volumes of the slices. In what follows, we are able to use single integrals to generalize this approach to handle even more general geometric shapes.

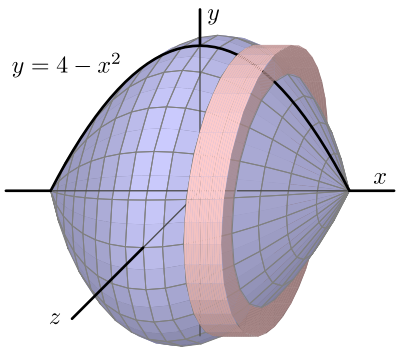


Figure 11.2.1. A solid of revolution.

Preview Activity 11.2.1

Let $f(x, y) = 25 - x^2 - y^2$ on the rectangular domain $R = [-3, 3] \times [-4, 4]$.

As with partial derivatives, we may treat one of the variables in f as constant and think of the resulting function as a function of a single variable. Now we investigate what happens if we integrate instead of differentiate.

1. Choose a fixed value of x in the interior of $[-3, 3]$. Let

$$A(x) = \int_{-4}^{4} f(x, y)\, dy.$$

   What is the geometric meaning of the value of $A(x)$ relative to the surface defined by f. (Hint: Think about the trace determined by the fixed value of x, and consider how $A(x)$ is related to the image at left in Figure 11.2.2.)

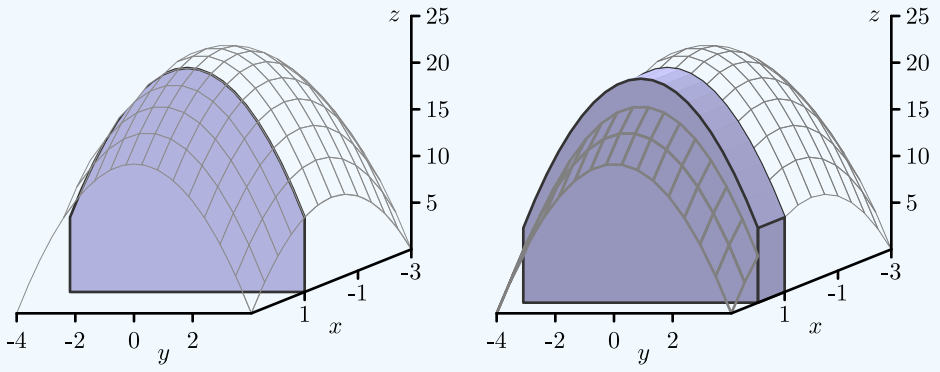


Figure 11.2.2. Left: A cross section with fixed x. Right: A cross section with fixed x and Δx.

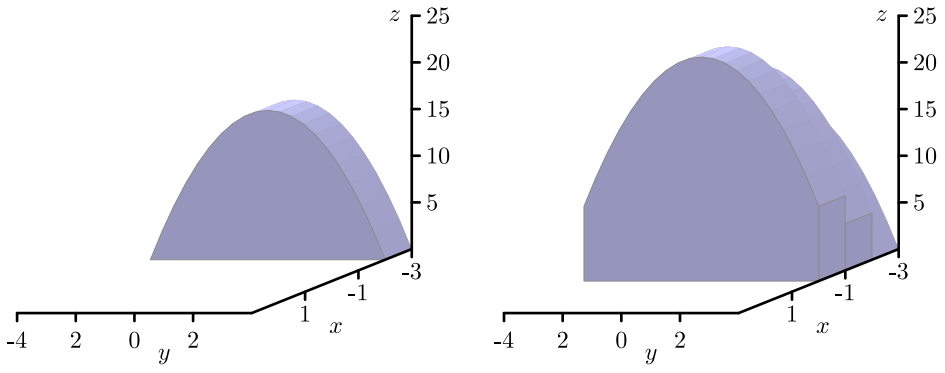
2. For a fixed value of x, say x_i^*, what is the geometric meaning of $A(x_i^*)\ \Delta x$? (Hint: Consider how $A(x_i^*)\Delta x$ is related to the image at right in Figure 11.2.2.)
3. Since f is continuous on R, we can define the function $A = A(x)$ at every value of x in $[-3, 3]$. Now think about subdividing the x-interval $[-3, 3]$ into m subintervals, and choosing a value x_i^* in each of those subintervals. What will be the meaning of the sum $\sum_{i=1}^{m} A(x_i^*)\ \Delta x$?
4. Explain why $\int_{-3}^{3} A(x)\ dx$ will determine the exact value of the volume under the surface $z = f(x, y)$ over the rectangle R.

Iterated Integrals

The ideas that we explored in Preview Activity 11.2.1 work more generally and lead to the idea of an iterated integral. Let f be a continuous function on a rectangular domain $R = [a, b] \times [c, d]$, and let

$$A(x) = \int_c^d f(x, y)\, dy.$$

The function $A = A(x)$ determines the value of the cross sectional area (by area we mean "signed" area) in the y direction for the fixed value of x of the solid bounded between the surface defined by f and the xy-plane.



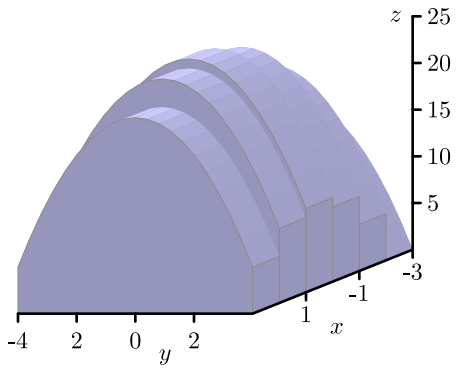


Figure 11.2.3. Summing volumes of cross section slices.

The value of this cross sectional area is determined by the input x in A. Since A is a function of x, it follows that we can integrate A with respect to x. In doing so, we use a partition of $[a, b]$ and make an approximation to the integral given by

$$\int_a^b A(x)\,dx \approx \sum_{i=1}^{m} A(x_i^*)\Delta x,$$

where x_i^* is any number in the subinterval $[x_{i-1}, x_i]$. Each term $A(x_i^*)\Delta x$ in the sum represents an approximation of a fixed cross sectional slice of the surface in the y direction with a fixed width of Δx as illustrated in Figure 11.2.3. We add the signed volumes of these slices as shown in the frames in Figure 11.2.3 to obtain an approximation of the total signed volume.

As we let the number of subintervals in the x direction approach infinity, we can see that the Riemann sum $\sum_{i=1}^{m} A(x_i^*)\Delta x$ approaches a limit and that limit is the sum of signed volumes bounded by the function f on R. Therefore, since $A(x)$ is itself determined by an integral, we have

$$\iint_R f(x,y)\,dA = \lim_{m\to\infty} \sum_{i=1}^{m} A(x_i^*)\Delta x = \int_a^b A(x)\,dx = \int_a^b \left(\int_c^d f(x,y)\,dy \right) dx.$$

Hence, we can compute the double integral of f over R by first integrating f with respect to y on $[c, d]$, then integrating the resulting function of x with respect to x on $[a, b]$. The nested integral

$$\int_a^b \left(\int_c^d f(x,y)\,dy \right) dx = \int_a^b \int_c^d f(x,y)\,dy\,dx$$

is called an *iterated integral*, and we see that each double integral may be represented by two single integrals.

We made a choice to integrate first with respect to y. The same argument shows that we can also find the double integral as an iterated integral integrating with respect to x first, or

$$\iint_R f(x,y)\,dA = \int_c^d \left(\int_a^b f(x,y)\,dx \right) dy = \int_c^d \int_a^b f(x,y)\,dx\,dy.$$

The fact that integrating in either order results in the same value is known as Fubini's Theorem.

Fubini's Theorem

If $f = f(x, y)$ is a continuous function on a rectangle $R = [a, b] \times [c, d]$, then

$$\iint_R f(x,y)\,dA = \int_c^d \int_a^b f(x,y)\,dx\,dy = \int_a^b \int_c^d f(x,y)\,dy\,dx.$$

Fubini's theorem enables us to evaluate iterated integrals without resorting to the limit definition. Instead, working with one

integral at a time, we can use the Fundamental Theorem of Calculus from single-variable calculus to find the exact value of each integral, starting with the inner integral.

Activity 11.2.2

Let $f(x,y) = 25 - x^2 - y^2$ on the rectangular domain $R = [-3,3] \times [-4,4]$.

1. Viewing x as a fixed constant, use the Fundamental Theorem of Calculus to evaluate the integral
$$A(x) = \int_{-4}^{4} f(x,y)\,dy.$$
Note that you will be integrating with respect to y, and holding x constant. Your result should be a function of x only.
2. Next, use your result from (a) along with the Fundamental Theorem of Calculus to determine the value of $\int_{-3}^{3} A(x)\ dx$.
3. What is the value of $\iint_R f(x,y)\ dA$? What are two different ways we may interpret the meaning of this value?

Activity 4

Let $f(x,y) = x + y^2$ on the rectangle $R = [0,2] \times [0,3]$.

1. Evaluate $\iint_R f(x,y)\ dA$ using an iterated integral. Choose an order for integration by deciding whether you want to integrate first with respect to x or y.
2. Evaluate $\iint_R f(x,y)\ dA$ using the iterated integral whose order of integration is the opposite of the order you chose in (a).

Summary

- We can evaluate the double integral $\iint_R f(x,y)\ dA$ over a rectangle $R = [a,b] \times [c,d]$ as an iterated integral in one of two ways:
  - $\int_a^b \left(\int_c^d f(x,y)\,dy \right)\ dx$, or
  - $\int_c^d \left(\int_a^b f(x,y)\,dx \right)\ dy$.

  This process works because each inner integral represents a cross-sectional (signed) area and the outer integral then sums all of the cross-sectional (signed) areas. Fubini's Theorem guarantees that the resulting value is the same, regardless of the order in which we integrate.

This page titled 11.2: Iterated Integrals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.2: Iterated Integrals by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

11.3: Double Integrals over General Regions

Motivating Questions

- How do we define a double integral over a non-rectangular region?
- What general form does an iterated integral over a non-rectangular region have?

Recall that we defined the double integral of a continuous function $f = f(x, y)$ over a rectangle $R = [a, b] \times [c, d]$ as

$$\iint_R f(x, y)\, dA = \lim_{m,n\to\infty} \sum_{j=1}^{n} \sum_{i=1}^{m} f(x_{ij}^*, y_{ij}^*) \cdot \Delta A,$$

where the notation is as described in Section 11.1. Furthermore, we have seen that we can evaluate a double integral $\iint_R f(x, y)\, dA$ over R as an iterated integral of either of the forms

$$\int_a^b \int_c^d f(x, y)\, dy\, dx \qquad \text{or} \qquad \int_c^d \int_a^b f(x, y)\, dx\, dy.$$

It is natural to wonder how we might define and evaluate a double integral over a non-rectangular region; we explore one such example in the following preview activity.

Preview Activity 11.3.1

A tetrahedron is a three-dimensional figure with four faces, each of which is a triangle. A picture of the tetrahedron T with vertices $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ is shown at left in Figure 11.3.1. If we place one vertex at the origin and let vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ be determined by the edges of the tetrahedron that have one end at the origin, then a formula that tells us the volume V of the tetrahedron is

$$V = \frac{1}{6}|\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|. \tag{11.3.1}$$

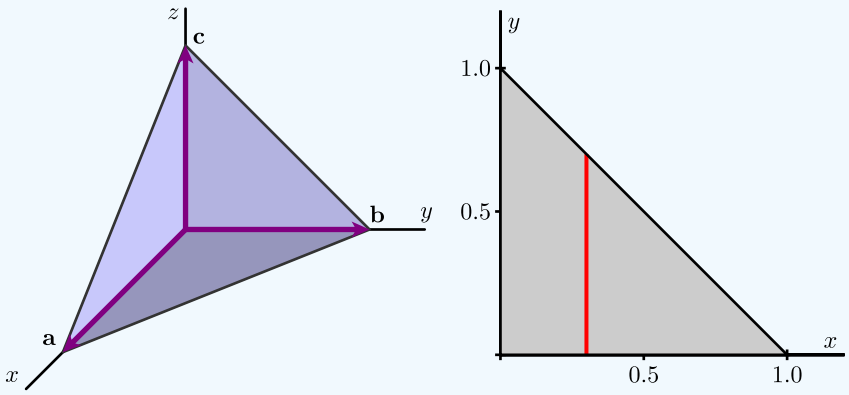


Figure 11.3.1. Left: The tetrahedron T. Right: Projecting T onto the xy-plane.

1. Use the formula (11.3.1) to find the volume of the tetrahedron T.
2. Instead of memorizing or looking up the formula for the volume of a tetrahedron, we can use a double integral to calculate the volume of the tetrahedron T. To see how, notice that the top face of the tetrahedron T is the plane whose equation is

$$z = 1 - (x + y).$$

Provided that we can use an iterated integral on a non-rectangular region, the volume of the tetrahedron will be given by an iterated integral of the form

$$\int_{x=?}^{x=?} \int_{y=?}^{y=?} 1-(x+y)\,dy\,dx.$$

The issue that is new here is how we find the limits on the integrals; note that the outer integral's limits are in x, while the inner ones are in y, since we have chosen $dA = dy\ dx$. To see the domain over which we need to integrate, think of standing way above the tetrahedron looking straight down on it, which means we are projecting the entire tetrahedron onto the xy-plane. The resulting domain is the triangular region shown at right in Figure 11.3.1. Explain why we can represent the triangular region with the inequalities

$$0 \le y \le 1-x \quad \text{and} \quad 0 \le x \le 1.$$

(Hint: Consider the cross sectional slice shown at right in Figure 11.3.1.)

3. Explain why it makes sense to now write the volume integral in the form

$$\int_{x=?}^{x=?} \int_{y=?}^{y=?} 1-(x+y)\,dy\,dx = \int_{x=0}^{x=1} \int_{y=0}^{y=1-x} 1-(x+y)\,dy\,dx.$$

4. Use the Fundamental Theorem of Calculus to evaluate the iterated integral

$$\int_{x=0}^{x=1} \int_{y=0}^{y=1-x} 1-(x+y)\,dy\,dx$$

and compare to your result from part (a). (As with iterated integrals over rectangular regions, start with the inner integral.)

Double Integrals over General Regions

So far, we have learned that a double integral over a rectangular region may be interpreted in one of two ways:

- $\iint_R f(x,y)\ dA$ tells us the volume of the solids the graph of f bounds above the xy-plane over the rectangle R minus the volume of the solids the graph of f bounds below the xy-plane under the rectangle R;
- $\frac{1}{A(R)} \iint_R f(x,y)\ dA$, where $A(R)$ is the area of R tells us the average value of the function f on R. If $f(x,y) \ge 0$ on R, we can interpret this average value of f on R as the height of the box with base R that has the same volume as the volume of the surface defined by f over R.

As we saw in Preview Activity 11.1.1, a function $f = f(x,y)$ may be considered over regions other than rectangular ones, and thus we want to understand how to set up and evaluate double integrals over non-rectangular regions. Note that if we can, then the two interpretations of the double integral noted above will naturally extend to solid regions with non-rectangular bases.

So, suppose f is a continuous function on a closed, bounded domain D. For example, consider D as the circular domain shown at left in Figure 11.3.2.

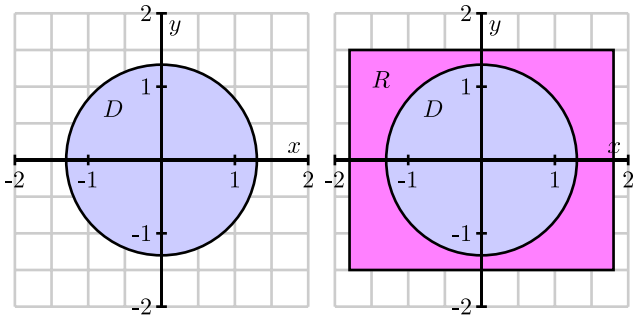


Figure 11.3.2. Left: A non-rectangular domain. Right: Enclosing this domain in a rectangle.

We can enclose D in a rectangular domain R as shown at right in Figure 11.3.2 and extend the function f to be defined over R in order to be able to use the definition of the double integral over a rectangle. We extend f in such a way that its values at the points in R that are not in D contribute 0 to the value of the integral. In other words, define a function $F = F(x, y)$ on R as

$$F(x,y) = \begin{cases} f(x,y), & \text{if } (x,y) \in D, \\ 0, & \text{if } (x,y) \notin D \end{cases}.$$

We then say that the double integral of f over D is the same as the double integral of F over R, and thus

$$\iint_D f(x,y)\,dA = \iint_R F(x,y)\,dA.$$

In practice, we just ignore everything that is in R but not in D, since these regions contribute 0 to the value of the integral.

Just as with double integrals over rectangles, a double integral over a domain D can be evaluated as an iterated integral. If the region D can be described by the inequalities $g_1(x) \le y \le g_2(x)$ and $a \le x \le b$, where $g_1 = g_1(x)$ and $g_2 = g_2(x)$ are functions of only x, then

$$\iint_D f(x,y)\,dA = \int_{x=a}^{x=b} \int_{y=g_1(x)}^{y=g_2(x)} f(x,y)\,dy\,dx.$$

Alternatively, if the region D is described by the inequalities $h_1(y) \le x \le h_2(y)$ and $c \le y \le d$, where $h_1 = h_1(y)$ and $h_2 = h_2(y)$ are functions of only y, we have

$$\iint_D f(x,y)\,dA = \int_{y=c}^{y=d} \int_{x=h_1(y)}^{x=h_2(y)} f(x,y)\,dx\,dy.$$

The structure of an iterated integral is of particular note:

In an iterated double integral:

- the limits on the outer integral must be constants;
- the limits on the inner integral must be constants or in terms of only the remaining variable — that is, if the inner integral is with respect to y, then its limits may only involve x and constants, and vice versa.

We next consider a detailed example.

Example 11.3.3

Let $f(x,y) = x^2 y$ be defined on the triangle D with vertices $(0,0)$, $(2,0)$, and $(2,3)$ as shown at left in Figure 11.3.3.

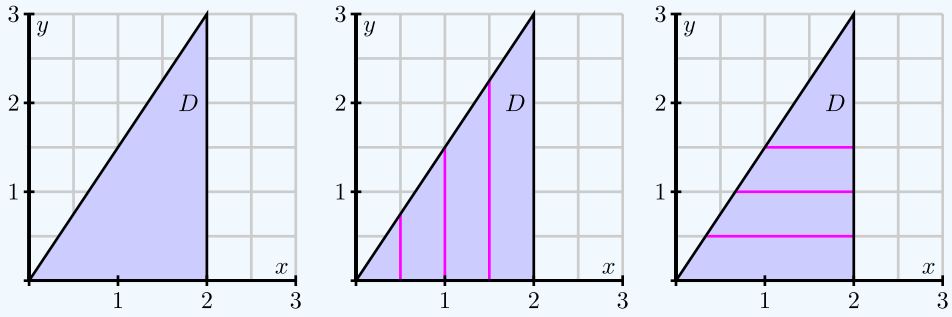


Figure 11.3.3. A triangular domain and slices in the y and x directions.

To evaluate $\iint_D f(x,y)\ dA$, we must first describe the region D in terms of the variables x and y. We take two approaches.

Approach 1: Integrate first with respect to y. available at the end of this section.

We see, of course, that in the situation where D can be described in two different ways, the order in which we choose to set up and evaluate the double integral doesn't matter, and the same value results in either case.

The meaning of a double integral over a non-rectangular region, D, parallels the meaning over a rectangular region. In particular,

- $\iint_D f(x,y)\ dA$ tells us the volume of the solids the graph of f bounds above the xy-plane over the closed, bounded region D minus the volume of the solids the graph of f bounds below the xy-plane under the region D;
- $\frac{1}{A(D)}\iint_R f(x,y)\ dA$, where $A(D)$ is the area of D tells us the average value of the function f on D. If $f(x,y) \geq 0$ on D, we can interpret this average value of f on D as the height of the solid with base D and constant cross-sectional area D that has the same volume as the volume of the surface defined by f over D.

Activity 11.3.2

Consider the double integral $\iint_D (4-x-2y)\ dA$, where D is the triangular region with vertices (0,0), (4,0), and (0,2).

1. Write the given integral as an iterated integral of the form $\iint_D (4-x-2y)\ dy\ dx$. Draw a labeled picture of D with relevant cross sections.
2. Write the given integral as an iterated integral of the form $\iint_D (4-x-2y)\ dx\ dy$. Draw a labeled picture of D with relevant cross sections.
3. Evaluate the two iterated integrals from (a) and (b), and verify that they produce the same value. Give at least one interpretation of the meaning of your result.

Activity 11.3.3

Consider the iterated integral $\int_{x=0}^{x=1}\int_{y=x}^{y=\sqrt{x}} (4x+10y)\ dy\ dx$.

1. Sketch the region of integration, D, for which
$$\iint_D (4x+10y)\,dA = \int_{x=0}^{x=1}\int_{y=x}^{y=\sqrt{x}} (4x+10y)\,dy\,dx.$$
2. Determine the equivalent iterated integral that results from integrating in the opposite order ($dx\ dy$, instead of $dy\ dx$). That is, determine the limits of integration for which
$$\iint_D (4x+10y)\,dA = \int_{y=?}^{y=?}\int_{x=?}^{x=?} (4x+10y)\,dx\,dy.$$
3. Evaluate one of the two iterated integrals above. Explain what the value you obtained tells you.
4. Set up and evaluate a single definite integral to determine the exact area of D, $A(D)$.
5. Determine the exact average value of $f(x,y) = 4x+10y$ over D.

Activity 11.3.4

Consider the iterated integral $\int_{x=0}^{x=4}\int_{y=x/2}^{y=2} e^{y^2}\ dy\ dx$.

1. Explain why we cannot find a simple antiderivative for e^{y^2} with respect to y, and thus are unable to evaluate $\int_{x=0}^{x=4}\int_{y=x/2}^{y=2} e^{y^2}\ dy\ dx$ in the indicated order using the Fundamental Theorem of Calculus.
2. Given that $\iint_D e^{y^2}\ dA = \int_{x=0}^{x=4}\int_{y=x/2}^{y=2} e^{y^2}\ dy\ dx$, sketch the region of integration, D.
3. Rewrite the given iterated integral in the opposite order, using $dA = dx\ dy$. (Hint: You may need more than one integral.)
4. Use the Fundamental Theorem of Calculus to evaluate the iterated integral you developed in (c). Write one sentence to explain the meaning of the value you found.
5. What is the important lesson this activity offers regarding the order in which we set up an iterated integral?

Summary

- For a double integral $\iint_D f(x,y)\ dA$ over a non-rectangular region D, we enclose D in a rectangle R and then extend integrand f to a function F so that $F(x,y)=0$ at all points in R outside of D and $F(x,y)=f(x,y)$ for all points in D. We then define $\iint_D f(x,y)\ dA$ to be equal to $\iint_R F(x,y)\ dA$.
- In an iterated double integral, the limits on the outer integral must be constants while the limits on the inner integral must be constants or in terms of only the remaining variable. In other words, an iterated double integral has one of the following forms (which result in the same value):

$$\int_{x=a}^{x=b}\int_{y=g_1(x)}^{y=g_2(x)} f(x,y)\,dy\,dx,$$

where $g_1=g_1(x)$ and $g_2=g_2(x)$ are functions of x only and the region D is described by the inequalities $g_1(x)\le y\le g_2(x)$ and $a\le x\le b$ or

$$\int_{y=c}^{y=d}\int_{x=h_1(y)}^{x=h_2(y)} f(x,y)\,dx\,dy,$$

where $h_1=h_1(y)$ and $h_2=h_2(y)$ are functions of y only and the region D is described by the inequalities $h_1(y)\le x\le h_2(y)$ and $c\le y\le d$.

Answers

Example 11.3.3

In this case we choose to evaluate the double integral as an iterated integral in the form

$$\iint_D x^2y\,dA=\int_{x=a}^{x=b}\int_{y=g_1(x)}^{y=g_2(x)} x^2y\,dy\,dx,$$

and therefore we need to describe D in terms of inequalities

$$g_1(x)\le y\le g_2(x)\qquad\text{and}\qquad a\le x\le b.$$

Since we are integrating with respect to y first, the iterated integral has the form

$$\iint_D x^2y\,dA=\int_{x=a}^{x=b} A(x)\,dx,$$

where $A(x)$ is a cross sectional area in the y direction. So we are slicing the domain perpendicular to the x-axis and want to understand what a cross sectional area of the overall solid will look like. Several slices of the domain are shown in the middle image in Figure 11.3.4. On a slice with fixed x value, the y values are bounded below by 0 and above by the y coordinate on the hypotenuse of the right triangle. Thus, $g_1(x)=0$; to find $y=g_2(x)$, we need to write the hypotenuse as a function of x. The hypotenuse connects the points (0,0) and (2,3) and hence has equation $y=\frac{3}{2}x$. This gives the upper bound on y as $g_2(x)=\frac{3}{2}x$. The leftmost vertical cross section is at $x=0$ and the rightmost one is at $x=2$, so we have $a=0$ and $b=2$. Therefore,

$$\iint_D x^2y\,dA=\int_{x=0}^{x=2}\int_{y=0}^{y=\frac{3}{2}x} x^2y\,dy\,dx.$$

We evaluate the iterated integral by applying the Fundamental Theorem of Calculus first to the inner integral, and then to the outer one, and find that

$$\begin{aligned}\int_{x=0}^{x=2}\int_{y=0}^{y=\frac{3}{2}x} x^2y\,dy\,dx &= \int_{x=0}^{x=2}\left[x^2\cdot\frac{y^2}{2}\right]\Big|_{y=0}^{y=\frac{3}{2}x}\,dx\\ &=\int_{x=0}^{x=2}\frac{9}{8}x^4\,dx\\ &=\frac{9}{8}\frac{x^5}{5}\Big|_{x=0}^{x=2}\\ &=\left(\frac{9}{8}\right)\left(\frac{32}{5}\right)\\ &=\frac{36}{5}.\end{aligned}$$

In this case, we choose to evaluate the double integral as an iterated integral in the form

$$\iint_D x^2y\,dA=\int_{y=c}^{y=d}\int_{x=h_1(y)}^{x=h_2(y)} x^2y\,dx\,dy$$

and thus need to describe D in terms of inequalities

$$h_1(y)\le x\le h_2(y)\qquad\text{and}\qquad c\le y\le d.$$

Since we are integrating with respect to x first, the iterated integral has the form

$$\iint_D x^2y\,dA=\int_c^d A(y)\,dy,$$

where $A(y)$ is a cross sectional area of the solid in the x direction. Several slices of the domain — perpendicular to the y-axis — are shown at right in Figure 11.3.3. On a slice with fixed y value, the x values are bounded below by the x coordinate on the hypotenuse of the right triangle and above by 2. So $h_2(y) = 2$; to find $h_1(y)$, we need to write the hypotenuse as a function of y. Solving the earlier equation we have for the hypotenuse ($y = \frac{3}{2}x$) for x gives us $x = \frac{2}{3}y$. This makes $h_1(y) = \frac{2}{3}y$. The lowest horizontal cross section is at $y = 0$ and the uppermost one is at $y = 3$, so we have $c = 0$ and $d = 3$. Therefore,

$$\iint_D x^2 y \, dA = \int_{y=0}^{y=3} \int_{x=(2/3)y}^{x=2} x^2 y \, dx \, dy.$$

We evaluate the resulting iterated integral as before by twice applying the Fundamental Theorem of Calculus, and find that

$$\begin{aligned}
\int_{y=0}^{y=3} \int_{x=\frac{2}{3}y}^{2} x^2 y \, dx \, dy &= \int_{y=0}^{y=3} \left[\frac{x^3}{3}\right]\Big|_{x=\frac{2}{3}y}^{x=2} y \, dx \\
&= \int_{y=0}^{y=3} \left[\frac{8}{3}y - \frac{8}{81}y^4\right] dy \\
&= \left[\frac{8}{3}\frac{y^2}{2} - \frac{8}{81}\frac{y^5}{5}\right]\Big|_{y=0}^{y=3} \\
&= \left(\frac{8}{3}\right)\left(\frac{9}{2}\right) - \left(\frac{8}{81}\right)\left(\frac{243}{5}\right) \\
&= 12 - \frac{24}{5} \\
&= \frac{36}{5}.
\end{aligned}$$

This page titled 11.3: Double Integrals over General Regions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.3: Double Integrals over General Regions by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

11.4: Applications of Double Integrals

Motivating Questions

- If we have a mass density function for a lamina (thin plate), how does a double integral determine the mass of the lamina?
- How may a double integral be used to find the area between two curves?
- Given a mass density function on a lamina, how can we find the lamina's center of mass?
- What is a joint probability density function? How do we determine the probability of an event if we know a probability density function?

So far, we have interpreted the double integral of a function f over a domain D in two different ways. First, $\iint_D f(x,y)\ dA$ tells us a difference of volumes — the volume the surface defined by f bounds above the xy-plane on D minus the volume the surface bounds below the xy-plane on D. In addition, $\frac{1}{A(D)}\iint_D f(x,y)\ dA$ determines the average value of f on D. In this section, we investigate several other applications of double integrals, using the integration process as seen in Preview Activity 11.4.1: we partition into small regions, approximate the desired quantity on each small region, then use the integral to sum these values exactly in the limit.

The following preview activity explores how a double integral can be used to determine the density of a thin plate with a mass density distribution. Recall that in single-variable calculus, we considered a similar problem and computed the mass of a one-dimensional rod with a mass-density distribution. There, as here, the key idea is that if density is constant, mass is the product of density and volume.

Preview Activity 11.4.1

Suppose that we have a flat, thin object (called a *lamina*) whose density varies across the object. We can think of the density on a lamina as a measure of mass per unit area. As an example, consider a circular plate D of radius 1 cm centered at the origin whose density δ varies depending on the distance from its center so that the density in grams per square centimeter at point (x,y) is

$$\delta(x,y) = 10 - 2(x^2+y^2).$$

1. Suppose that we partition the plate into subrectangles R_{ij}, where $1 \le i \le m$ and $1 \le j \le n$, of equal area ΔA, and select a point (x_{ij}^*, y_{ij}^*) in R_{ij} for each i and j. What is the meaning of the quantity $\delta(x_{ij}^*, y_{ij}^*)\Delta A$?
2. State a double Riemann sum that provides an approximation of the mass of the plate.
3. Explain why the double integral

$$\iint_D \delta(x,y)\,dA$$

tells us the exact mass of the plate.

4. Determine an iterated integral which, if evaluated, would give the exact mass of the plate. Do not actually evaluate the integral. (This integral is considerably easier to evaluate in polar coordinates, which we will learn more about in Section 11.5.)

Mass

Density is a measure of some quantity per unit area or volume. For example, we can measure the human population density of some region as the number of humans in that region divided by the area of that region. In physics, the mass density of an object is the mass of the object per unit area or volume. As suggested by Preview Activity 11.4.1, the following holds in general.

The mass of a lamina

If $\delta(x,y)$ describes the density of a lamina defined by a planar region D, then the *mass* of D is given by the double integral $\iint_D \delta(x,y)\ dA$.

Activity 11.4.2

Let D be a half-disk lamina of radius 3 in quadrants IV and I, centered at the origin as shown in Figure 11.4.1. Assume the density at point (x,y) is given by $\delta(x,y) = x$. Find the exact mass of the lamina.

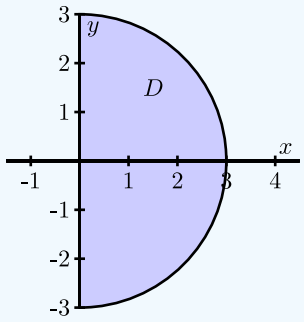


Figure 11.4.1. A half disk lamina.

Area

If we consider the situation where the mass-density distribution is constant, we can also see how a double integral may be used to determine the area of a region. Assuming that $\delta(x,y) = 1$ over a closed bounded region D, where the units of δ are "mass per unit of area," it follows that $\iint_D 1\ dA$ is the mass of the lamina. But since the density is constant, the numerical value of the integral is simply the area.

As the following activity demonstrates, we can also see this fact by considering a three-dimensional solid whose height is always 1.

Activity 11.4.3

Suppose we want to find the area of the bounded region D between the curves

$$y = 1 - x^2 \qquad \text{and} \qquad y = x - 1.$$

A picture of this region is shown in Figure 11.4.2.

a. The volume of a solid with constant height is given by the area of the base times the height. Hence, we may interpret the area of the region D as the volume of a solid with base D and of uniform height 1. That is, the area of D is given by $\iint_D 1\ dA$. Write an iterated integral whose value is $\iint_D 1\ dA$. (Hint: Which order of integration might be more efficient? Why?)

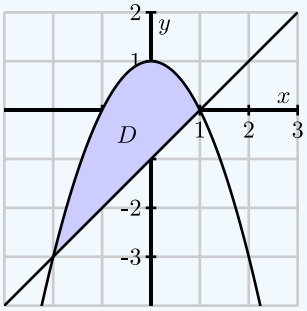


Figure 11.4.2. The graphs of $y = 1 - x^2$ and $y = x - 1$.

b. Evaluate the iterated integral from (a). What does the result tell you?

We now formally state the conclusion from our earlier discussion and Activity 11.4.3.

The double integral and area

Given a closed, bounded region D in the plane, the area of D, denoted $A(D)$, is given by the double integral

$$A(D) = \iint_D 1 \, dA.$$

Center of Mass

The center of mass of an object is a point at which the object will balance perfectly. For example, the center of mass of a circular disk of uniform density is located at its center. For any object, if we throw it through the air, it will spin around its center of mass and behave as if all the mass is located at the center of mass.

In order to understand the role that integrals play in determining the center of mass of an object with a nonuniform mass distribution, we start by finding the center of mass of a collection of N distinct point-masses in the plane.

Let $m_1, m_2, \ldots, m_N$ be N masses located in the plane. Think of these masses as connected by rigid rods of negligible weight from some central point (x, y). A picture with four masses is shown in Figure 11.4.3. Now imagine balancing this system by placing it on a thin pole at the point (x, y) perpendicular to the plane containing the masses. Unless the masses are perfectly balanced, the system will fall off the pole. The point $(\overline{x}, \overline{y})$ at which the system will balance perfectly is called the *center of mass* of the system. Our goal is to determine the center of mass of a system of discrete masses, then extend this to a continuous lamina.

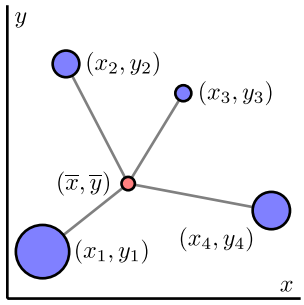


Figure 11.4.3. A center of mass $(\overline{x}, \overline{y})$ of four masses.

Each mass exerts a force (called a *moment*) around the lines $x = \overline{x}$ and $y = \overline{y}$ that causes the system to tilt in the direction of the mass. These moments are dependent on the mass and the distance from the given line. Let (x_1, y_1) be the location of mass m_1,

(x_2, y_2) the location of mass m_2, etc. In order to balance perfectly, the moments in the x direction and in the y direction must be in equilibrium. We determine these moments and solve the resulting system to find the equilibrium point $(\overline{x}, \overline{y})$ at the center of mass.

The force that mass m_1 exerts to tilt the system from the line $y = \overline{y}$ is

$$m_1 g(\overline{y} - y_1),$$

where g is the gravitational constant. Similarly, the force mass m_2 exerts to tilt the system from the line $y = \overline{y}$ is

$$m_2 g(\overline{y} - y_2).$$

In general, the force that mass m_k exerts to tilt the system from the line $y = \overline{y}$ is

$$m_k g(\overline{y} - y_k).$$

For the system to balance, we need the forces to sum to 0, so that

$$\sum_{k=1}^{N} m_k g(\overline{y} - y_k) = 0.$$

Solving for $\overline{y}$, we find that

$$\overline{y} = \frac{\sum_{k=1}^{N} m_k y_k}{\sum_{k=1}^{N} m_k}.$$

A similar argument shows that

$$\overline{x} = \frac{\sum_{k=1}^{N} m_k x_k}{\sum_{k=1}^{N} m_k}.$$

The value $M_x = \sum_{k=1}^{N} m_k y_k$ is called the *total moment* with respect to the x-axis; $M_y = \sum_{k=1}^{N} m_k x_k$ is the *total moment* with respect to the y-axis. Hence, the respective quotients of the moments to the total mass, M, determines the center of mass of a point-mass system:

$$(\overline{x}, \overline{y}) = \left(\frac{M_y}{M}, \frac{M_x}{M} \right).$$

Now, suppose that rather than a point-mass system, we have a continuous lamina with a variable mass-density $\delta(x, y)$. We may estimate its center of mass by partitioning the lamina into mn subrectangles of equal area ΔA, and treating the resulting partitioned lamina as a point-mass system. In particular, we select a point (x_{ij}^*, y_{ij}^*) in the ijth subrectangle, and observe that the quanity

$$\delta(x_{ij}^*, y_{ij}^*)\Delta A$$

is density times area, so $\delta(x_{ij}^*, y_{ij}^*)\Delta A$ approximates the mass of the small portion of the lamina determined by the subrectangle R_{ij}.

We now treat $\delta(x_{ij}^*, y_{ij}^*)\Delta A$ as a point mass at the point (x_{ij}^*, y_{ij}^*). The coordinates $(\overline{x}, \overline{y})$ of the center of mass of these mn point masses are thus given by

$$\overline{x} = \frac{\sum_{j=1}^{n} \sum_{i=1}^{m} x_{ij}^* \delta(x_{ij}^*, y_{ij}^*)\Delta A}{\sum_{j=1}^{n} \sum_{i=1}^{m} \delta(x_{ij}^*, y_{ij}^*)\Delta A} \qquad \text{and} \qquad \overline{y} = \frac{\sum_{j=1}^{n} \sum_{i=1}^{m} y_{ij}^* \delta(x_{ij}^*, y_{ij}^*)\Delta A}{\sum_{j=1}^{n} \sum_{i=1}^{m} \delta(x_{ij}^*, y_{ij}^*)\Delta A}.$$

If we take the limit as m and n go to infinity, we obtain the exact center of mass $(\overline{x}, \overline{y})$ of the continuous lamina.

The center of mass of a lamina

The coordinates $(\overline{x}, \overline{y})$ of the *center of mass of a lamina* D with density $\delta = \delta(x, y)$ are given by

$$\overline{x} = \frac{\iint_D x\delta(x,y)\,dA}{\iint_D \delta(x,y)\,dA} \qquad \text{and} \qquad \overline{y} = \frac{\iint_D y\delta(x,y)\,dA}{\iint_D \delta(x,y)\,dA}.$$

The center of mass of a lamina can then be thought of as a weighted average of all of the points in the lamina with the weights given by the density at each point. The *centroid* of a lamina is the just the average of all of the points in the lamina, or the center of mass if the density at each point is 1.

The numerators of $\overline{x}$ and $\overline{y}$ are called the respective *moments* of the lamina about the coordinate axes. Thus, the moment of a lamina D with density $\delta = \delta(x, y)$ about the y-axis is

$$M_y = \iint_D x\delta(x,y)\,dA$$

and the moment of D about the x-axis is

$$M_x = \iint_D y\delta(x,y)\,dA.$$

If M is the mass of the lamina, it follows that the center of mass is

$$(\overline{x}, \overline{y}) = \left(\frac{M_y}{M}, \frac{M_x}{M}\right).$$

Activity 11.4.4

In this activity we determine integrals that represent the center of mass of a lamina D described by the triangular region bounded by the x-axis and the lines $x = 1$ and $y = 2x$ in the first quadrant if the density at point (x, y) is $\delta(x, y) = 6x + 6y + 6$. A picture of the lamina is shown in Figure 11.4.4.

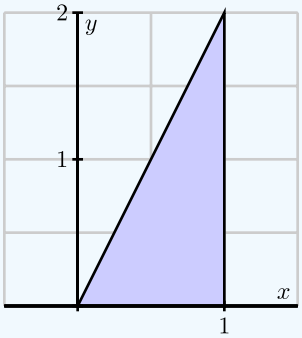


Figure 11.4.4. The lamina bounded by the x-axis and the lines $x = 1$ and $y = 2x$ in the first quadrant.

1. Set up an iterated integral that represents the mass of the lamina.
2. Assume the mass of the lamina is 14. Set up two iterated integrals that represent the coordinates of the center of mass of the lamina.

Probability

Calculating probabilities is a very important application of integration in the physical, social, and life sciences. To understand the basics, consider the game of darts in which a player throws a dart at a board and tries to hit a particular target. Let us suppose that a dart board is in the form of a disk D with radius 10 inches. If we assume that a player throws a dart at random, and is not

aiming at any particular point, then it is equally probable that the dart will strike any single point on the board. For instance, the probability that the dart will strike a particular 1 square inch region is $\frac{1}{100\pi}$, or the ratio of the area of the desired target to the total area of D (assuming that the dart thrower always hits the board itself at some point). Similarly, the probability that the dart strikes a point in the disk D_3 of radius 3 inches is given by the area of D_3 divided by the area of D. In other words, the probability that the dart strikes the disk D_3 is

$$\frac{9\pi}{100\pi} = \iint_{D_3} \frac{1}{100\pi}\, dA.$$

The integrand, $\frac{1}{100\pi}$, may be thought of as a *distribution function*, describing how the dart strikes are distributed across the board. In this case the distribution function is constant since we are assuming a uniform distribution, but we can easily envision situations where the distribution function varies. For example, if the player is fairly good and is aiming for the bulls eye (the center of D), then the distribution function f could be skewed toward the center, say

$$f(x, y) = Ke^{-(x^2+y^2)}$$

for some constant positive K. If we assume that the player is consistent enough so that the dart always strikes the board, then the probability that the dart strikes the board somewhere is 1, and the distribution function f will have to satisfy [1]

$$\iint_D f(x, y)\, dA = 1.$$

This makes $K = \frac{1}{\pi(1-e^{-100})}$, which you can check.

For such a function f, the probability that the dart strikes in the disk D_1 of radius 1 would be

$$\iint_{D_1} f(x, y)\, dA.$$

Indeed, the probability that the dart strikes in any region R that lies within D is given by

$$\iint_R f(x, y)\, dA.$$

The preceding discussion highlights the general idea behind calculating probabilities. We assume we have a *joint probability density function* f, a function of two independent variables x and y defined on a domain D that satisfies the conditions

- $f(x, y) \geq 0$ for all x and y in D,
- the probability that x is between some values a and b while y is between some values c and d is given by
$$\int_a^b \int_c^d f(x, y)\, dy\, dx,$$
- The probability that the point (x, y) is in D is 1, that is
$$\iint_D f(x, y)\, dA = 1. \tag{11.4.1}$$

Note that it is possible that D could be an infinite region and the limits on the integral in Equation 11.4.1 could be infinite. When we have such a probability density function $f = f(x, y)$, the probability that the point (x, y) is in some region R contained in the domain D (the notation we use here is "$P((x, y)^{\text{int}} R)$") is determined by

$$P((x, y)^{\text{int}} R) = \iint_R f(x, y)\, dA.$$

Activity 11.4.5

A firm manufactures smoke detectors. Two components for the detectors come from different suppliers — one in Michigan and one in Ohio. The company studies these components for their reliability and their data suggests that if x is the life span (in years) of a randomly chosen component from the Michigan supplier and y the life span (in years) of a randomly chosen component from the Ohio supplier, then the joint probability density function f might be given by

$$f(x,y) = e^{-x}e^{-y}.$$

1. Theoretically, the components might last forever, so the domain D of the function f is the set D of all (x,y) such that $x \geq 0$ and $y \geq 0$. To show that f is a probability density function on D we need to demonstrate that
$$\int\int_D f(x,y)\,dA = 1,$$
or that
$$\int_0^\infty \int_0^\infty f(x,y)\,dy\,dx = 1.$$
Use your knowledge of improper integrals to verify that f is indeed a probability density function.
2. Assume that the smoke detector fails only if both of the supplied components fail. To determine the probability that a randomly selected detector will fail within one year, we will need to determine the probability that the life span of each component is between 0 and 1 years. Set up an appropriate iterated integral, and evaluate the integral to determine the probability.
3. What is the probability that a randomly chosen smoke detector will fail between years 3 and 7?
4. Suppose that the manufacturer determines that one of the components is more likely to fail than the other, and hence conjectures that the probability density function is instead $f(x,y) = Ke^{-x}e^{-2y}$. What is the value of K?

Summary

- The mass of a lamina D with a mass density function $\delta = \delta(x,y)$ is $\iint_D \delta(x,y)\ dA$.
- The area of a region D in the plane has the same numerical value as the volume of a solid of uniform height 1 and base D, so the area of D is given by $\iint_D 1\ dA$.
- The center of mass, $(\overline{x},\overline{y})$, of a continuous lamina with a variable density $\delta(x,y)$ is given by
$$\overline{x} = \frac{\iint_D x\delta(x,y)\,dA}{\iint_D \delta(x,y)\,dA} \qquad \text{and} \qquad \overline{y} = \frac{\iint_D y\delta(x,y)\,dA}{\iint_D \delta(x,y)\,dA}.$$
- Given a joint probability density function f is a function of two independent variables x and y defined on a domain D, if R is some subregion of D, then the probability that (x,y) is in R is given by
$$\iint_R f(x,y)\,dA.$$

This page titled 11.4: Applications of Double Integrals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.4: Applications of Double Integrals by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

11.5: Double Integrals in Polar Coordinates

Motivating Questions

- What are the polar coordinates of a point in two-space?
- How do we convert between polar coordinates and rectangular coordinates?
- What is the area element in polar coordinates?
- How do we convert a double integral in rectangular coordinates to a double integral in polar coordinates?

While we have naturally defined double integrals in the rectangular coordinate system, starting with domains that are rectangular regions, there are many of these integrals that are difficult, if not impossible, to evaluate. For example, consider the domain D that is the unit circle and $f(x,y) = e^{-x^2-y^2}$. To integrate f over D, we would use the iterated integral

$$\iint_D f(x,y)\,dA = \int_{x=-1}^{x=1} \int_{y=-\sqrt{1-x^2}}^{y=\sqrt{1-x^2}} e^{-x^2-y^2}\,dy\,dx.$$

For this particular integral, regardless of the order of integration, we are unable to find an antiderivative of the integrand; in addition, even if we were able to find an antiderivative, the inner limits of integration involve relatively complicated functions.

It is useful, therefore, to be able to translate to other coordinate systems where the limits of integration and evaluation of the involved integrals is simpler. In this section we provide a quick discussion of one such system — polar coordinates — and then introduce and investigate their ramifications for double integrals. The rectangular coordinate system allows us to consider domains and graphs relative to a rectangular grid. The polar coordinate system is an alternate coordinate system that allows us to consider domains less suited to rectangular coordinates, such as circles.

Preview Activity 11.5.1

The coordinates of a point determine its location. In particular, the rectangular coordinates of a point P are given by an ordered pair (x,y), where x is the (signed) distance the point lies from the y-axis to P and y is the (signed) distance the point lies from the x-axis to P. In polar coordinates, we locate the point by considering the distance the point lies from the origin, $O = (0,0)$, and the angle the line segment from the origin to P forms with the positive x-axis.

1. Determine the rectangular coordinates of the following points:
   1. The point P that lies 1 unit from the origin on the positive x-axis.
   2. The point Q that lies 2 units from the origin and such that $\overline{OQ}$ makes an angle of $\frac{\pi}{2}$ with the positive x-axis.
   3. The point R that lies 3 units from the origin such that $\overline{OR}$ makes an angle of $\frac{2\pi}{3}$ with the positive x-axis.
2. Part (a) indicates that the two pieces of information completely determine the location of a point: either the traditional (x,y) coordinates, or alternately, the distance r from the point to the origin along with the angle θ that the line through the origin and the point makes with the positive x-axis. We write "(r,θ)" to denote the point's location in its polar coordinate representation. Find polar coordinates for the points with the given rectangular coordinates.
   1. $(0,-1)$ ii. $(-2,0)$ iii. $(-1,1)$
3. For each of the following points whose coordinates are given in polar form, determine the rectangular coordinates of the point.
   1. $(5,\frac{\pi}{4})$ ii. $(2,\frac{5\pi}{6})$ iii. $(\sqrt{3},\frac{5\pi}{3})$

Polar Coordinates

The rectangular coordinate system is best suited for graphs and regions that are naturally considered over a rectangular grid. The polar coordinate system is an alternative that offers good options for functions and domains that have more circular characteristics. A point P in rectangular coordinates that is described by an ordered pair (x,y), where x is the displacement from P to the y-axis and y is the displacement from P to the x-axis, as seen in Preview Activity 11.5.1, can also be described with polar coordinates (r,θ), where r is the distance from P to the origin and θ is the angle formed by the line segment $\overline{OP}$ and the positive

x-axis, as shown at left in Figure 11.5.1.

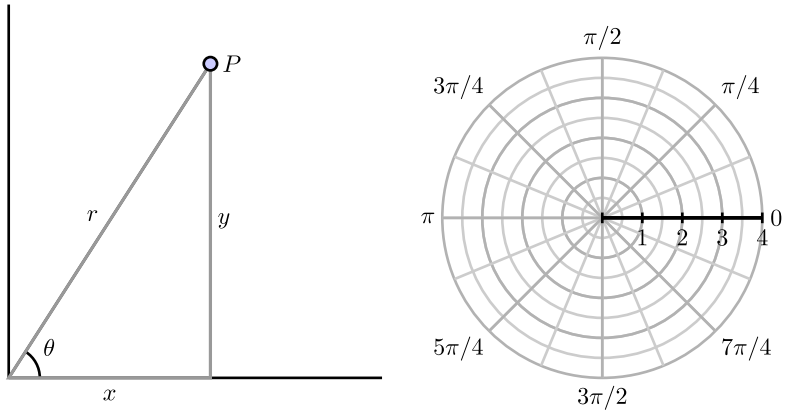


Figure 11.5.1. The polar coordinates of a point and the polar coordinate grid.

Trigonometry and the Pythagorean Theorem allow for straightforward conversion from rectangular to polar, and vice versa.

Converting between rectangular and polar coordinates

- If we are given the rectangular coordinates (x, y) of a point P, then the polar coordinates (r, θ) of P satisfy
$$r = \sqrt{x^2 + y^2} \qquad \text{and} \qquad \tan(\theta) = \frac{y}{x}, \text{ assuming } x \neq 0.$$
- If we are given the polar coordinates (r, θ) of a point P, then the rectangular coordinates (x, y) of P satisfy
$$x = r\cos(\theta) \qquad \text{and} \qquad y = r\sin(\theta).$$

Note: The angle θ in the polar coordinates of a point is not unique. We could replace θ with $\theta + 2\pi$ and still be at the same terminal point. In addition, the sign of $\tan(\theta)$ does not uniquely determine the quadrant in which θ lies, so we have to determine the value of θ from the location of the point. In other words, more care has to be paid when using polar coordinates than rectangular coordinates.

We can draw graphs of curves in polar coordinates in a similar way to how we do in rectangular coordinates. However, when plotting in polar coordinates, we use a grid that considers changes in angles and changes in distance from the origin. In particular, the angles θ and distances r partition the plane into small wedges as shown at right in Figure 11.5.1.

Activity 11.5.2

Most polar graphing devices can plot curves in polar coordinates of the form $r = f(\theta)$. Use such a device to complete this activity.

1. Before plotting the polar curve $r = 1$ (where θ can have any value), think about what shape it should have, in light of how r is connected to x and y. Then use appropriate technology to draw the graph and test your intuition.
2. The equation $\theta = 1$ does not define r as a function of θ, so we can't graph this equation on many polar plotters. What do you think the graph of the polar curve $\theta = 1$ looks like? Why?
3. Before plotting the polar curve $r = \theta$, what do you think the graph looks like? Why? Use technology to plot the curve and compare your intuition.
4. What does the region defined by $1 \leq r \leq 3$ (where θ can have any value) look like? (Hint: Compare to your response from part (a).)
5. What does the region defined by $1 \leq r \leq 3$ and $\pi/4 \leq \theta \leq \pi/2$ look like?
6. Consider the curve $r = \sin(\theta)$. For some values of θ we will have $r < 0$. In these situations, we plot the point (r, θ) as $(|r|, \theta + \pi)$ (in other words, when $r < 0$, we reflect the point through the origin). With that in mind, what do you think the graph of $r = \sin(\theta)$ looks like? Plot this curve using technology and compare to your intuition.

Integrating in Polar Coordinates

Consider the double integral

$$\iint_D e^{x^2+y^2}\,dA,$$

where D is the unit disk. While we cannot directly evaluate this integral in rectangular coordinates, a change to polar coordinates will convert it to one we can easily evaluate.

We have seen how to evaluate a double integral $\iint_D f(x,y)\ dA$ as an iterated integral of the form

$$\int_a^b \int_{g_1(x)}^{g_2(x)} f(x,y)\,dy\,dx$$

in rectangular coordinates, because we know that $dA = dy\ dx$ in rectangular coordinates. To make the change to polar coordinates, we not only need to represent the variables x and y in polar coordinates, but we also must understand how to write the area element, dA, in polar coordinates. That is, we must determine how the area element dA can be written in terms of dr and $d\theta$ in the context of polar coordinates. We address this question in the following activity.

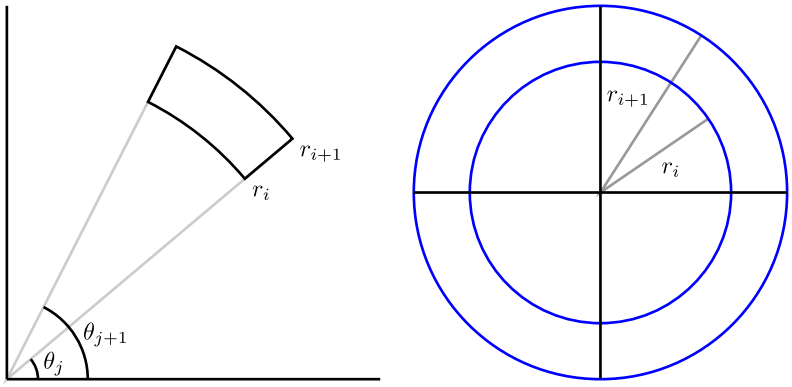


Figure 11.5.2. Left: A polar rectangle. Right: An annulus.

Activity 11.5.3

Consider a polar rectangle R, with r between r_i and r_{i+1} and θ between θ_j and θ_{j+1} as shown at left in Figure 11.5.2. Let $\Delta r = r_{i+1} - r_i$ and $\Delta\theta = \theta_{j+1} - \theta_j$. Let ΔA be the area of this region.

1. Explain why the area ΔA in polar coordinates is not $\Delta r\ \Delta\theta$.
2. Now find ΔA by the following steps:
   1. Find the area of the annulus (the washer-like region) between r_i and r_{i+1}, as shown at right in Figure 11.5.2. This area will be in terms of r_i and r_{i+1}.
   2. Observe that the region R is only a portion of the annulus, so the area ΔA of R is only a fraction of the area of the annulus. For instance, if $\theta_{i+1} - \theta_i$ were $\frac{\pi}{4}$, then the resulting wedge would be
   $$\frac{\frac{\pi}{4}}{2\pi} = \frac{1}{8}$$
   of the entire annulus. In this more general context, using the wedge between the two noted angles, what fraction of the area of the annulus is the area ΔA?
   3. Write an expression for ΔA in terms of r_i, r_{i+1}, θ_j, and θ_{j+1}.
   4. Finally, write the area ΔA in terms of r_i, r_{i+1}, Δr, and $\Delta\theta$, where each quantity appears only once in the expression. (Hint: Think about how to factor a difference of squares.)

3. As we take the limit as Δr and $\Delta\theta$ go to 0, Δr becomes dr, $\Delta\theta$ becomes $d\theta$, and ΔA becomes dA, the area element. Using your work in (iv), write dA in terms of r, dr, and $d\theta$.

From the result of Activity 11.5.3, we see when we convert an integral from rectangular coordinates to polar coordinates, we must not only convert x and y to being in terms of r and θ, but we also have to change the area element to $dA = r\ dr\ d\theta$ in polar coordinates. As we saw in Activity 11.5.3, the reason the additional factor of r in the polar area element is due to the fact that in polar coordinates, the cross sectional area element increases as r increases, while the cross sectional area element in rectangular coordinates is constant. So, given a double integral $\iint_D f(x,y)\ dA$ in rectangular coordinates, to write a corresponding iterated integral in polar coordinates, we replace x with $r\cos(\theta)$, y with $r\sin(\theta)$ and dA with $r\ dr\ d\theta$. Of course, we need to describe the region D in polar coordinates as well. To summarize:

Double integrals in polar coordinates

The double integral $\iint_D f(x,y)\ dA$ in rectangular coordinates can be converted to a double integral in polar coordinates as $\iint_D f(r\cos(\theta), r\sin(\theta))\ r\ dr\ d\theta$.

Example 11.5.3

Let $f(x,y) = e^{x^2+y^2}$ on the disk $D = \{(x,y) : x^2 + y^2 \le 1\}$. We will evaluate $\iint_D f(x,y)\ dA$.

In rectangular coordinates the double integral $\iint_D f(x,y)\ dA$ can be written as the iterated integral

$$\iint_D f(x,y)\,dA = \int_{x=-1}^{x=1}\int_{y=-\sqrt{1-x^2}}^{y=\sqrt{1-x^2}} e^{x^2+y^2}\,dy\,dx.$$

We cannot evaluate this iterated integral, because $e^{x^2+y^2}$ does not have an elementary antiderivative with respect to either x or y. However, since $r^2 = x^2 + y^2$ and the region D is circular, it is natural to wonder whether converting to polar coordinates will allow us to evaluate the new integral. To do so, we replace x with $r\cos(\theta)$, y with $r\sin(\theta)$, and $dy\ dx$ with $r\ dr\ d\theta$ to obtain

$$\iint_D f(x,y)\,dA = \iint_D e^{r^2}\,r\,dr\,d\theta.$$

The disc D is described in polar coordinates by the constraints $0 \le r \le 1$ and $0 \le \theta \le 2\pi$. Therefore, it follows that

$$\iint_D e^{r^2}\,r\,dr\,d\theta = \int_{\theta=0}^{\theta=2\pi}\int_{r=0}^{r=1} e^{r^2}\,r\,dr\,d\theta.$$

We can evaluate the resulting iterated polar integral as follows:

$$\begin{aligned}
\int_{\theta=0}^{\theta=2\pi}\int_{r=0}^{r=1} e^{r^2}\,r\,dr\,d\theta &= \int_{\theta=0}^{2\pi}\left(\frac{1}{2}e^{r^2}\Big|_{r=0}^{r=1}\right)d\theta \\
&= \frac{1}{2}\int_{\theta=0}^{\theta=2\pi}(e-1)\ d\theta \\
&= \frac{1}{2}(e-1)\int_{\theta=0}^{\theta=2\pi} d\theta \\
&= \frac{1}{2}(e-1)\,[\theta]\Big|_{\theta=0}^{\theta=2\pi} \\
&= \pi(e-1).
\end{aligned}$$

While there is no firm rule for when polar coordinates can or should be used, they are a natural alternative anytime the domain of integration may be expressed simply in polar form, and/or when the integrand involves expressions such as $\sqrt{x^2+y^2}$.

Activity 11.5.4

Let $f(x,y) = x + y$ and $D = \{(x,y) : x^2 + y^2 \leq 4\}$.

1. Sketch the region D and then write the double integral of f over D as an iterated integral in rectangular coordinates.
2. Write the double integral of f over D as an iterated integral in polar coordinates.
3. Evaluate one of the iterated integrals. Why is the final value you found not surprising?

Activity 11.5.5

Consider the circle given by $x^2 + (y-1)^2 = 1$ as shown in Figure 11.5.4.

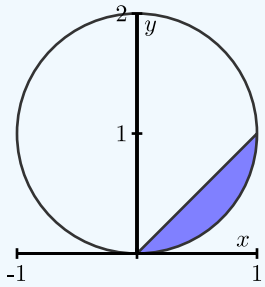


Figure 11.5.4. The graphs of $y = x$ and $x^2 + (y-1)^2 = 1$, for use in Activity 11.5.5.

1. Determine a polar curve in the form $r = f(\theta)$ that traces out the circle $x^2 + (y-1)^2 = 1$. (Hint: Recall that a circle centered at the origin of radius r can be described by the equations $x = r\cos(\theta)$ and $y = r\sin(\theta)$.)
2. Find the exact average value of $g(x,y) = \sqrt{x^2 + y^2}$ over the interior of the circle $x^2 + (y-1)^2 = 1$.
3. Find the volume under the surface $h(x,y) = x$ over the region D, where D is the region bounded above by the line $y = x$ and below by the circle (this is the shaded region in Figure 11.5.4).
4. Explain why in both (b) and (c) it is advantageous to use polar coordinates.

Summary

- The polar representation of a point P is the ordered pair (r,θ) where r is the distance from the origin to P and θ is the angle the ray through the origin and P makes with the positive x-axis.
- The polar coordinates r and θ of a point (x,y) in rectangular coordinates satisfy

$$r = \sqrt{x^2 + y^2} \qquad \text{and} \qquad \tan(\theta) = \frac{y}{x};$$

  the rectangular coordinates x and y of a point (r,θ) in polar coordinates satisfy

$$x = r\cos(\theta) \qquad \text{and} \qquad y = r\sin(\theta).$$

- The area element dA in polar coordinates is determined by the area of a slice of an annulus and is given by

$$dA = r\,dr\,d\theta.$$

- To convert the double integral $\iint_D f(x,y)\,dA$ to an iterated integral in polar coordinates, we substitute $r\cos(\theta)$ for x, $r\sin(\theta)$ for y, and $r\ dr\ d\theta$ for dA to obtain the iterated integral

$$\iint_D f(r\cos(\theta), r\sin(\theta))\, r\,dr\,d\theta.$$

This page titled 11.5: Double Integrals in Polar Coordinates is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.5: Double Integrals in Polar Coordinates by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

11.6: Surfaces Defined Parametrically and Surface Area

Motivating Questions

- What is a parameterization of a surface?
- How do we find the surface area of a parametrically defined surface?

We have now studied at length how curves in space can be defined parametrically by functions of the form $\mathbf{r}(t) = \langle x(t), y(t), z(t) \rangle$, and surfaces can be represented by functions $z = f(x, y)$. In what follows, we will see how we can also define surfaces parametrically. A one-dimensional curve in space results from a vector function that relies upon one parameter, so a two-dimensional surface naturally involves the use of two parameters. If $x = x(s, t)$, $y = y(s, t)$, and $z = z(s, t)$ are functions of independent parameters s and t, then the terminal points of all vectors of the form

$$\mathbf{r}(s, t) = x(s, t)\mathbf{i} + y(s, t)\mathbf{j} + z(s, t)\mathbf{k}$$

form a surface in space. The equations $x = x(s, t)$, $y = y(s, t)$, and $z = z(s, t)$ are the *parametric equations* for the surface, or a *parametrization* of the surface. In Preview Activity $11.6.1$ we investigate how to parameterize a cylinder and a cone.

Preview Activity $11.6.1$

Recall the standard parameterization of the unit circle that is given by

$$x(t) = \cos(t) \quad \text{and} \quad y(t) = \sin(t),$$

where $0 \leq t \leq 2\pi$.

1. Determine a parameterization of the circle of radius 1 in $\mathbb{R}^3$ that has its center at $(0, 0, 1)$ and lies in the plane $z = 1$.
2. Determine a parameterization of the circle of radius 1 in 3-space that has its center at $(0, 0, -1)$ and lies in the plane $z = -1$.
3. Determine a parameterization of the circle of radius 1 in 3-space that has its center at $(0, 0, 5)$ and lies in the plane $z = 5$.
4. Taking into account your responses in (a), (b), and (c), describe the graph that results from the set of parametric equations

   $$x(s, t) = \cos(t), \quad y(s, t) = \sin(t), \quad \text{and} \quad z(s, t) = s,$$

   where $0 \leq t \leq 2\pi$ and $-5 \leq s \leq 5$. Explain your thinking.

5. Just as a cylinder can be viewed as a "stack" of circles of constant radius, a cone can be viewed as a stack of circles with varying radius. Modify the parametrizations of the circles above in order to construct the parameterization of a cone whose vertex lies at the origin, whose base radius is 4, and whose height is 3, where the base of the cone lies in the plane $z = 3$. Use appropriate technology to plot the parametric equations you develop. (Hint: The cross sections parallel to the xy-plane are circles, with the radii varying linearly as z increases.)

Parametric Surfaces

In a single-variable setting, any function may have its graph expressed parametrically. For instance, given $y = g(x)$, by considering the parameterization $\langle t, g(t) \rangle$ (where t belongs to the domain of g), we generate the same curve. What is more important is that certain curves that are not functions may be represented parametrically; for instance, the circle (which cannot be represented by a single function) can be parameterized by $\langle \cos(t), \sin(t) \rangle$, where $0 \leq t \leq 2\pi$.

In the same way, in a two-variable setting, the surface $z = f(x, y)$ may be expressed parametrically by considering

$$\langle x(s, t), y(s, t), z(s, t) \rangle = \langle s, t, f(s, t) \rangle,$$

where (s, t) varies over the entire domain of f. Therefore, any familiar surface that we have studied so far can be generated as a parametric surface. But what is more powerful is that there are surfaces that cannot be generated by a single function $z = f(x, y)$ (such as the unit sphere), but that can be represented parametrically. We now consider an important example.

Example 11.6.1

Consider the torus (or doughnut) shown in Figure 11.6.1.

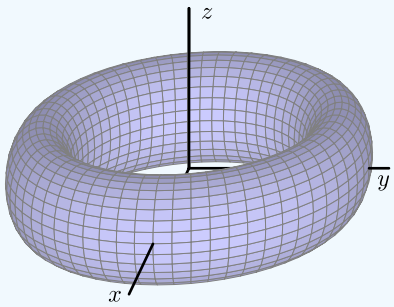


Figure 11.6.1. A torus.

To find a parametrization of this torus, we recall our work in Preview Activity 11.6.1. There, we saw that a circle of radius r that has its center at the point $(0, 0, z_0)$ and is contained in the horizontal plane $z = z_0$, as shown in Figure 11.6.2, can be parametrized using the vector-valued function $\mathbf{r}$ defined by

$$\mathbf{r}(t) = r\cos(t)\mathbf{i} + r\sin(t)\mathbf{j} + z_0\mathbf{k}$$

where $0 \le t \le 2\pi$.

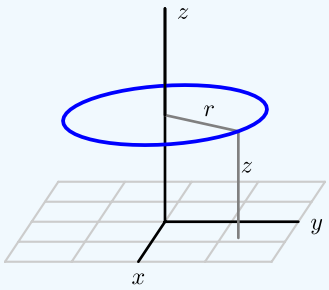


Figure 11.6.2. A circle in a horizontal plane centered at $(0, 0, z_0)$.

To obtain the torus in Figure 11.6.1, we begin with a circle of radius a in the xz-plane centered at $(b, 0)$, as shown on the left of Figure 11.6.4. We may parametrize the points on this circle, using the parameter s, by using the equations

$$x(s) = b + a\cos(s) \text{ and } z(s) = a\sin(s),$$

where $0 \le s \le 2\pi$.

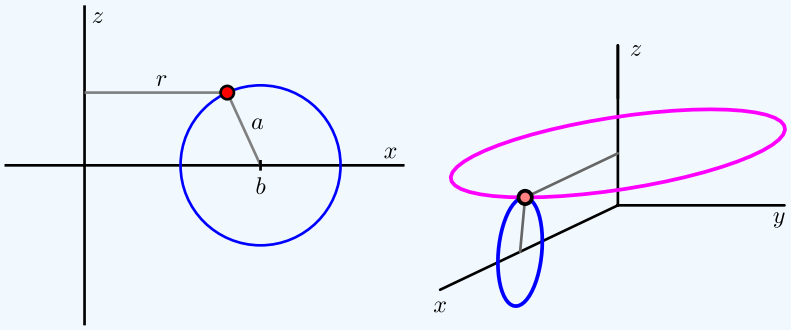


Figure 11.6.3. Revolving a circle to obtain a torus.

Let's focus our attention on one point on this circle, such as the indicated point, which has coordinates $(x(s), 0, z(s))$ for a fixed value of the parameter s. When this point is revolved about the z-axis, we obtain a circle contained in a horizontal plane centered at $(0, 0, z(s))$ and having radius $x(s)$, as shown on the right of Figure 11.6.3. If we let t be the new parameter that generates the circle for the rotation about the z-axis, this circle may be parametrized by

$$\mathbf{r}(s,t) = x(s)\cos(t)\mathbf{i} + x(s)\sin(t)\mathbf{j} + z(s)\mathbf{k}.$$

Now using our earlier parametric equations for $x(s)$ and $z(s)$ for the original smaller circle, we have an overall parameterization of the torus given by

$$\mathbf{r}(s,t) = (b + a\cos(s))\cos(t)\mathbf{i} + (b + a\cos(s))\sin(t)\mathbf{j} + a\sin(s)\mathbf{k}.$$

To trace out the entire torus, we require that the parameters vary through the values $0 \le s \le 2\pi$ and $0 \le t \le 2\pi$.

Activity 11.6.2

In this activity, we seek a parametrization of the sphere of radius R centered at the origin, as shown on the left in Figure 11.6.4. Notice that this sphere may be obtained by revolving a half-circle contained in the xz-plane about the z-axis, as shown on the right.

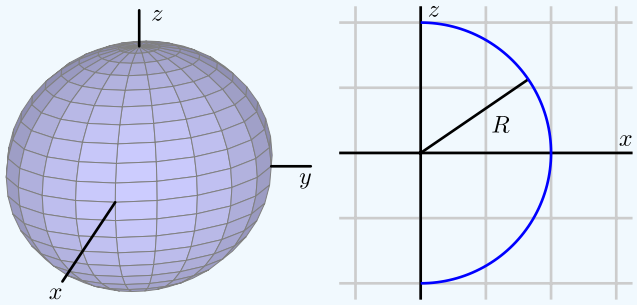


Figure 11.6.4. A sphere obtained by revolving a half-circle.

1. Begin by writing a parametrization of this half-circle using the parameter s:

   $$x(s) = \ldots \qquad z(s) = \ldots .$$

   Be sure to state the domain of the parameter s.

2. By revolving the points on this half-circle about the z-axis, obtain a parametrization $\mathbf{r}(s,t)$ of the points on the sphere of radius R. Be sure to include the domain of both parameters s and t. (Hint: What is the radius of the circle obtained when revolving a point on the half-circle around the z axis?)

3. Draw the surface defined by your parameterization with appropriate technology.

The Surface Area of Parametrically Defined Surfaces

Recall that a differentiable function is locally linear — that is, if we zoom in on the surface around a point, the surface looks like its tangent plane. We now exploit this idea in order to determine the surface area generated by a parametrization $\langle x(s,t), y(s,t), z(s,t)\rangle$. The basic idea is a familiar one: we will subdivide the surface into small pieces, in the approximate shape of small parallelograms, and thus estimate the entire the surface area by adding the areas of these approximation parallelograms. Ultimately, we use an integral to sum these approximations and determine the exact surface area.

Let

$$\mathbf{r}(s,t) = x(s,t)\mathbf{i} + y(s,t)\mathbf{j} + z(s,t)\mathbf{k}$$

define a surface over a rectangular domain $a \le s \le b$ and $c \le t \le d$. As a function of two variables, s and t, it is natural to consider the two partial derivatives of the vector-valued function $\mathbf{r}$, which we define by

$$\mathbf{r}_s(s,t) = x_s(s,t)\mathbf{i} + y_s(s,t)\mathbf{j} + z_s(s,t)\mathbf{k}$$
$$\mathbf{r}_t(s,t) = x_t(s,t)\mathbf{i} + y_t(s,t)\mathbf{j} + z_t(s,t)\mathbf{k}.$$

In the usual way, we slice the domain into small rectangles. In particular, we partition the interval $[a,b]$ into m subintervals of length $\Delta s = \frac{b-a}{n}$ and let $s_0, s_1, \ldots, s_m$ be the endpoints of these subintervals, where $a = s_0 < s_1 < s_2 < \cdots < s_m = b$. Also partition the interval $[c,d]$ into n subintervals of equal length $\Delta t = \frac{d-c}{n}$ and let $t_0, t_1, \ldots, t_n$ be the endpoints of these subintervals, where $c = t_0 < t_1 < t_2 < \cdots < t_n = d$. These two partitions create a partition of the rectangle $R = [a,b] \times [c,d]$ in st-coordinates into mn sub-rectangles R_{ij} with opposite vertices (s_{i-1}, t_{j-1}) and (s_i, t_j) for i between 1 and m and j between 1 and n. These rectangles all have equal area $\Delta A = \Delta s \cdot \Delta t$.

Now we want to think about the small piece of area on the surface itself that lies above one of these small rectangles in the domain. Observe that if we increase s by a small amount Δs from the point (s_{i-1}, t_{j-1}) in the domain, then $\mathbf{r}$ changes by approximately $\mathbf{r}_s(s_{i-1}, t_{j-1})\Delta s$. Similarly, if we increase t by a small amount Δt from the point (s_{i-1}, t_{j-1}), then $\mathbf{r}$ changes by approximately $\mathbf{r}_t(s_{i-1}, t_{j-1})\Delta t$. So we can approximate the surface defined by $\mathbf{r}$ on the st-rectangle $[s_{i-1}, s_i] \times [t_{j-1}, t_j]$ with the parallelogram determined by the vectors $\mathbf{r}_s(s_{i-1}, t_{j-1})\Delta s$ and $\mathbf{r}_t(s_{i-1}, t_{j-1})\Delta t$, as seen in Figure 11.6.5.

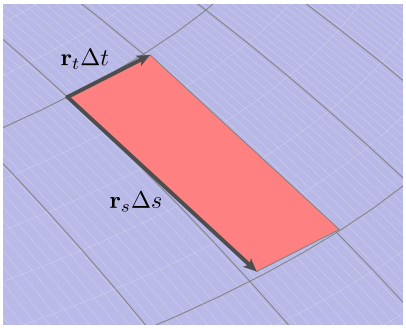


Figure 11.6.5. Approximation surface area with a parallelogram.

Say that the small parallelogram has area S_{ij}. If we can find its area, then all that remains is to sum the areas of all of the generated parallelograms and take a limit. Recall from our earlier work in the course that given two vectors $\mathbf{u}$ and $\mathbf{v}$, the area of the parallelogram spanned by $\mathbf{u}$ and $\mathbf{v}$ is given by the magnitude of their cross product, $|\mathbf{u} \times \mathbf{v}|$. In the present context, it follows that the area, S_{ij}, of the parallelogram determined by the vectors $\mathbf{r}_s(s_{i-1}, t_{j-1})\Delta s$ and $\mathbf{r}_t(s_{i-1}, t_{j-1})\Delta t$ is

$$\begin{aligned} S_{ij} &= |(\mathbf{r}_s(s_{i-1}, t_{j-1})\Delta s) \times (\mathbf{r}_t(s_{i-1}, t_{j-1})\Delta t)| \\ &= |\mathbf{r}_s(s_{i-1}, t_{j-1}) \times \mathbf{r}_t(s_{i-1}, t_{j-1})|\Delta s \Delta t, \end{aligned} \tag{11.6.1}$$

where the latter equality holds from standard properties of the cross product and length.

We sum the surface area approximations from Equation 11.6.1 over all sub-rectangles to obtain an estimate for the total surface

area, S, given by

$$S \approx \sum_{i=1}^{m} \sum_{j=1}^{n} |\mathbf{r}_s(s_{i-1}, t_{j-1}) \times \mathbf{r}_t(s_{i-1}, t_{j-1})| \Delta s \Delta t.$$

Taking the limit as $m, n \to \infty$ shows that the surface area of the surface defined by $\mathbf{r}$ over the domain D is given as follows.

Surface area

Let $\mathbf{r}(s,t) = \langle x(s,t), y(s,t), z(s,t) \rangle$ be a parameterization of a smooth surface over a domain D. The *area of the surface* defined by $\mathbf{r}$ on D is given by

$$S = \iint_D |\mathbf{r}_s \times \mathbf{r}_t| \, dA. \tag{11.6.2}$$

Activity 11.6.3

Consider the cylinder with radius a and height h defined parametrically by

$$\mathbf{r}(s,t) = a\cos(s)\mathbf{i} + a\sin(s)\mathbf{j} + t\mathbf{k}$$

for $0 \le s \le 2\pi$ and $0 \le t \le h$, as shown in Figure 11.6.6.

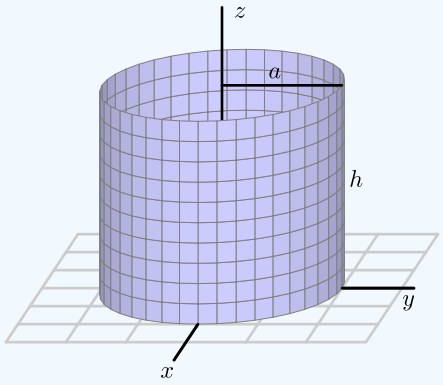


Figure 11.6.6. A cylinder.

1. Set up an iterated integral to determine the surface area of this cylinder.
2. Evaluate the iterated integral.
3. Recall that one way to think about the surface area of a cylinder is to cut the cylinder horizontally and find the perimeter of the resulting cross sectional circle, then multiply by the height. Calculate the surface area of the given cylinder using this alternate approach, and compare your work in (b).

As we noted earlier, we can take any surface $z = f(x,y)$ and generate a corresponding parameterization for the surface by writing $\langle s, t, f(s,t) \rangle$. Hence, we can use our recent work with parametrically defined surfaces to find the surface area that is generated by a function $f = f(x,y)$ over a given domain.

Activity 11.6.4

Let $z = f(x,y)$ define a smooth surface, and consider the corresponding parameterization $\mathbf{r}(s,t) = \langle s, t, f(s,t) \rangle$.

1. Let D be a region in the domain of f. Using Equation 11.6.2, show that the area, S, of the surface defined by the graph of f over D is
$$S = \iint_D \sqrt{(f_x(x,y))^2 + (f_y(x,y))^2 + 1}\ dA.$$
2. Use the formula developed in (a) to calculate the area of the surface defined by $f(x,y) = \sqrt{4-x^2}$ over the rectangle $D = [-2,2] \times [0,3]$.
3. Observe that the surface of the solid describe in (b) is half of a circular cylinder. Use the standard formula for the surface area of a cylinder to calculate the surface area in a different way, and compare your result from (b).

Summary

- A parameterization of a curve describes the coordinates of a point on the curve in terms of a single parameter t, while a parameterization of a surface describes the coordinates of points on the surface in terms of two independent parameters.
- If $\mathbf{r}(s,t) = \langle x(s,t), y(s,t), z(s,t)\rangle$ describes a smooth surface in 3-space on a domain D, then the area, S, of that surface is given by
$$S = \iint_D |\mathbf{r}_s \times \mathbf{r}_t|\ dA.$$

This page titled 11.6: Surfaces Defined Parametrically and Surface Area is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.6: Surfaces Defined Parametrically and Surface Area by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

11.7: Triple Integrals

6Motivating Questions

- How are a triple Riemann sum and the corresponding triple integral of a continuous function $f = f(x, y, z)$ defined?
- What are two things the triple integral of a function can tell us?

We have now learned that we define the double integral of a continuous function $f = f(x, y)$ over a rectangle $R = [a, b] \times [c, d]$ as a limit of a double Riemann sum, and that these ideas parallel the single-variable integral of a function $g = g(x)$ on an interval $[a, b]$. Moreover, this double integral has natural interpretations and applications, and can even be considered over non-rectangular regions, D. For instance, given a continuous function f over a region D, the average value of f, $f_{\text{AVG}(D)}$, is given by

$$f_{\text{AVG}(D)} = \frac{1}{A(D)} \iint_D f(x, y)\, dA,$$

where $A(D)$ is the area of D. Likewise, if $\delta(x, y)$ describes a mass density function on a lamina over D, the mass, M, of the lamina is given by

$$M = \iint_D \delta(x, y)\, dA.$$

It is natural to wonder if it is possible to extend these ideas of double Riemann sums and double integrals for functions of two variables to triple Riemann sums and then triple integrals for functions of three variables. We begin investigating in Preview Activity 11.7.1.

Preview Activity 11.7.1

Consider a solid piece of granite in the shape of a box $B = \{(x, y, z) : 0 \le x \le 4, 0 \le y \le 6, 0 \le z \le 8\}$, whose density varies from point to point. Let $\delta(x, y, z)$ represent the mass density of the piece of granite at point (x, y, z) in kilograms per cubic meter (so we are measuring x, y, and z in meters). Our goal is to find the mass of this solid.

Recall that if the density was constant, we could find the mass by multiplying the density and volume; since the density varies from point to point, we will use the approach we did with two-variable lamina problems, and slice the solid into small pieces on which the density is roughly constant.

Partition the interval $[0, 4]$ into 2 subintervals of equal length, the interval $[0, 6]$ into 3 subintervals of equal length, and the interval $[0, 8]$ into 2 subintervals of equal length. This partitions the box B into sub-boxes as shown in Figure 11.7.1.

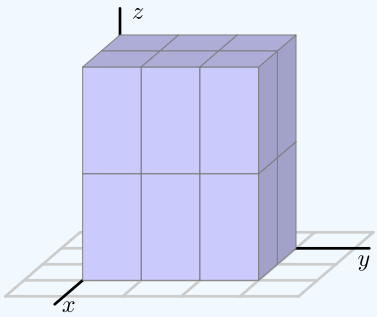


Figure 11.7.1. A partitioned three-dimensional domain.

1. Let $0 = x_0 < x_1 < x_2 = 4$ be the endpoints of the subintervals of $[0, 4]$ after partitioning. Draw a picture of Figure 11.7.1 and label these endpoints on your drawing. Do likewise with $0 = y_0 < y_1 < y_2 < y_3 = 6$ and $0 = z_0 < z_1 < z_2 = 8$ What is the length Δx of each subinterval $[x_{i-1}, x_i]$ for i from 1 to 2? the length of Δy? of Δz?
2. The partitions of the intervals $[0, 4]$, $[0, 6]$ and $[0, 8]$ partition the box B into sub-boxes. How many sub-boxes are there? What is volume ΔV of each sub-box?
3. Let B_{ijk} denote the sub-box $[x_{i-1}, x_i] \times [y_{j-1}, y_j] \times [z_{k-1}, z_k]$. Say that we choose a point $(x^*_{ijk}, y^*_{ijk}, z^*_{ijk})$ in the i, j, kth sub-box for each possible combination of i, j, k. What is the meaning of $\delta(x^*_{ijk}, y^*_{ijk}, z^*_{ijk})$? What physical quantity will $\delta(x^*_{ijk}, y^*_{ijk}, z^*_{ijk})\Delta V$ approximate?
4. What final step(s) would it take to determine the exact mass of the piece of granite?

Triple Riemann Sums and Triple Integrals

Through the application of a mass density distribution over a three-dimensional solid, Preview Activity 11.7.1 suggests that the generalization from double Riemann sums of functions of two variables to triple Riemann sums of functions of three variables is natural. In the same way, so is the generalization from double integrals to triple integrals. By simply adding a z-coordinate to our earlier work, we can define both a triple Riemann sum and the corresponding triple integral.

Definition 11.7.1

Let $f = f(x, y, z)$ be a continuous function on a box $B = [a, b] \times [c, d] \times [r, s]$. The ***triple Riemann sum*** *of f over B* is created as follows.

- Partition the interval $[a, b]$ into m subintervals of equal length $\Delta x = \frac{b-a}{m}$. Let x_0, x_1, ..., x_m be the endpoints of these subintervals, where $a = x_0 < x_1 < x_2 < \cdots < x_m = b$. Do likewise with the interval $[c, d]$ using n subintervals of equal length $\Delta y = \frac{d-c}{n}$ to generate $c = y_0 < y_1 < y_2 < \cdots < y_n = d$, and with the interval $[r, s]$ using ℓ subintervals of equal length $\Delta z = \frac{s-r}{\ell}$ to have $r = z_0 < z_1 < z_2 < \cdots < z_l = s$.
- Let B_{ijk} be the sub-box of B with opposite vertices $(x_{i-1}, y_{j-1}, z_{k-1})$ and (x_i, y_j, z_k) for i between 1 and m, j between 1 and n, and k between 1 and ℓ. The volume of each B_{ijk} is $\Delta V = \Delta x \cdot \Delta y \cdot \Delta z$.
- Let $(x^*_{ijk}, y^*_{ijk}, z^*_{ijk})$ be a point in box B_{ijk} for each i, j, and k. The resulting triple Riemann sum for f on B is

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{\ell} f(x^*_{ijk}, y^*_{ijk}, z^*_{ijk}) \cdot \Delta V.$$

If $f(x, y, z)$ represents the mass density of the box B, then, as we saw in Preview Activity 11.7.1, the triple Riemann sum approximates the total mass of the box B. In order to find the exact mass of the box, we need to let the number of sub-boxes increase without bound (in other words, let m, n, and ℓ go to infinity); in this case, the finite sum of the mass approximations becomes the actual mass of the solid B. More generally, we have the following definition of the triple integral.

Definition 11.7.2

With following notation defined as in a triple Riemann sum, the *triple integral of f over B* is

$$\iiint_B f(x, y, z)\, dV = \lim_{m,n,\ell\to\infty} \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{\ell} f(x^*_{ijk}, y^*_{ijk}, z^*_{ijk}) \cdot \Delta V.$$

As we noted earlier, if $f(x, y, z)$ represents the density of the solid B at each point (x, y, z), then

$$M = \iiint_B f(x, y, z)\, dV$$

is the mass of B. Even more importantly, for any continuous function f over the solid B, we can use a triple integral to determine the average value of f over B, $f_{\text{AVG}(B)}$. We note this generalization of our work with functions of two variables along with several others in the following important boxed information. Note that each of these quantities may actually be considered over a general

domain S in $\mathbb{R}^3$, not simply a box, B.

- The triple integral
$$V(S) = \iiint_S 1\, dV$$
represents the *volume* of the solid S.

- The *average value* of the function $f = f(x, y, x)$ over a solid domain S is given by
$$f_{\text{AVG}(S)} = \left(\frac{1}{V(S)}\right) \iiint_S f(x, y, z)\, dV,$$
where $V(S)$ is the volume of the solid S.

- The *center of mass* of the solid S with density $\delta = \delta(x, y, z)$ is $(\overline{x}, \overline{y}, \overline{z})$, where
$$\overline{x} = \frac{\iiint_S x\ \delta(x, y, z)\, dV}{M},$$
$$\overline{y} = \frac{\iiint_S y\ \delta(x, y, z)\, dV}{M},$$
$$\overline{z} = \frac{\iiint_S z\ \delta(x, y, z)\, dV}{M},$$
and $M = \iiint_S \delta(x, y, z)\ dV$ is the mass of the solid S.

In the Cartesian coordinate system, the volume element dV is $dz\ dy\ dx$, and, as a consequence, a triple integral of a function f over a box $B = [a, b] \times [c, d] \times [r, s]$ in Cartesian coordinates can be evaluated as an iterated integral of the form
$$\iiint_B f(x, y, z)\, dV = \int_a^b \int_c^d \int_r^s f(x, y, z)\, dz\, dy\, dx.$$

If we want to evaluate a triple integral as an iterated integral over a solid S that is not a box, then we need to describe the solid in terms of variable limits.

Activity 11.7.2

a. Set up and evaluate the triple integral of $f(x, y, z) = x - y + 2z$ over the box $B = [-2, 3] \times [1, 4] \times [0, 2]$.

b. Let S be the solid cone bounded by $z = \sqrt{x^2 + y^2}$ and $z = 3$. A picture of S is shown at right in Figure 11.7.2. Our goal in what follows is to set up an iterated integral of the form
$$\int_{x=?}^{x=?} \int_{y=?}^{y=?} \int_{z=?}^{z=?} \delta(x, y, z)\, dz\, dy\, dx \tag{11.7.1}$$
to represent the mass of S in the setting where $\delta(x, y, z)$ tells us the density of S at the point (x, y, z). Our particular task is to find the limits on each of the three integrals.

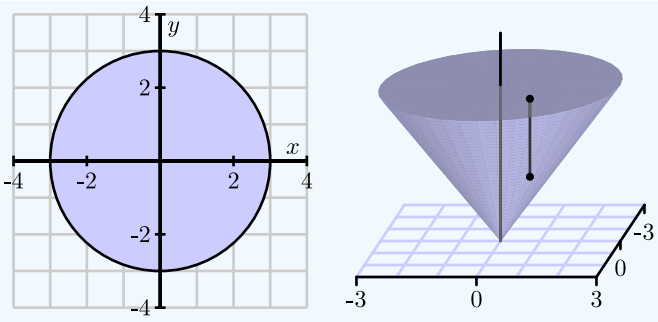


Figure 11.7.2. Left: The cone. Right: Its projection.

i. If we think about slicing up the solid, we can consider slicing the domain of the solid's projection onto the xy-plane (just as we would slice a two-dimensional region in $\mathbb{R}^2$), and then slice in the z-direction as well. The projection of the solid onto the xy-plane is shown at left in Figure 11.7.4. If we decide to first slice the domain of the solid's projection perpendicular to the x-axis, over what range of constant x-values would we have to slice?

ii. If we continue with slicing the domain, what are the limits on y on a typical slice? How do these depend on x? What, therefore, are the limits on the middle integral?

iii. Finally, now that we have thought about slicing up the two-dimensional domain that is the projection of the cone, what are the limits on z in the innermost integral? Note that over any point (x, y) in the plane, a vertical slice in the z direction will involve a range of values from the cone itself to its flat top. In particular, observe that at least one of these limits is not constant but depends on x and y.

iv. In conclusion, write an iterated integral of the form 11.7.1 that represents the mass of the cone S.

Note well: When setting up iterated integrals, the limits on a given variable can be *only* in terms of the remaining variables. In addition, there are multiple different ways we can choose to set up such an integral. For example, two possibilities for iterated integrals that represent a triple integral $\iiint_S f(x,y,z)\ dV$ over a solid S are

- $\displaystyle\int_a^b \int_{g_1(x)}^{g_2(x)} \int_{h_1(x,y)}^{h_2(x,y)} f(x,y,z)\ dz\ dy\ dx$
- $\displaystyle\int_r^s \int_{p_1(z)}^{p_2(z)} \int_{q_1(x,z)}^{q_2(x,z)} f(x,y,z)\ dy\ dx\ dz$

where g_1, g_2, h_1, h_2, p_1, p_2, q_1, and q_2 are functions of the indicated variables. There are four other options beyond the two stated here, since the variables x, y, and z can (theoretically) be arranged in any order. Of course, in many circumstances, an insightful choice of variable order will make it easier to set up an iterated integral, just as was the case when we worked with double integrals.

Example 11.7.1

Find the mass of the tetrahedron in the first octant bounded by the coordinate planes and the plane $x + 2y + 3z = 6$ if the density at point (x, y, z) is given by $\delta(x, y, z) = x + y + z$. A picture of the solid tetrahedron is shown at left in Figure 11.7.3.

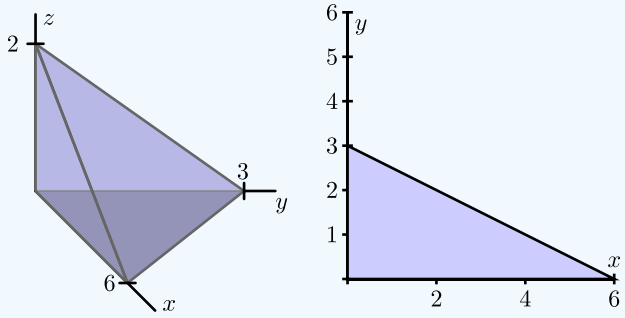


Figure 11.7.3. Left: The tetrahedron. Right: Its projection.

We find the mass, M, of the tetrahedron by the triple integral

$$M = \iiint_S \delta(x, y, z)\, dV,$$

where S is the solid tetrahedron described above. In this example, we choose to integrate with respect to z first for the innermost integral. The top of the tetrahedron is given by the equation

$$x + 2y + 3z = 6;$$

solving for z then yields

$$z = \frac{1}{3}(6 - x - 2y).$$

The bottom of the tetrahedron is the xy-plane, so the limits on z in the iterated integral will be $0 \le z \le \frac{1}{3}(6 - x - 2y)$.

To find the bounds on x and y we project the tetrahedron onto the xy-plane; this corresponds to setting $z = 0$ in the equation $z = \frac{1}{3}(6 - x - 2y)$. The resulting relation between x and y is

$$x + 2y = 6.$$

The right image in Figure 11.7.3 shows the projection of the tetrahedron onto the xy-plane.

If we choose to integrate with respect to y for the middle integral in the iterated integral, then the lower limit on y is the x -axis and the upper limit is the hypotenuse of the triangle. Note that the hypotenuse joins the points $(6, 0)$ and $(0, 3)$ and so has equation $y = 3 - \frac{1}{2}x$. Thus, the bounds on y are $0 \le y \le 3 - \frac{1}{2}x$. Finally, the x values run from 0 to 6, so the iterated integral that gives the mass of the tetrahedron is

$$M = \int_0^6 \int_0^{3-(1/2)x} \int_0^{(1/3)(6-x-2y)} x + y + z\, dz\, dy\, dx. \tag{11.7.2}$$

Evaluating the triple integral gives us

$$
\begin{aligned}
M &= \int_0^6 \int_0^{3-(1/2)x} \int_0^{(1/3)(6-x-2y)} x+y+z\,dz\,dy\,dx \\
&= \int_0^6 \int_0^{3-(1/2)x} \left[xz+yz+\frac{z}{2}\right]\Big|_0^{(1/3)(6-x-2y)} dy\,dx \\
&= \int_0^6 \int_0^{3-(1/2)x} \frac{4}{3}x-\frac{5}{18}x^2-\frac{1}{9}xy+\frac{2}{3}y-\frac{4}{9}y^2+2\,dy\,dx \\
&= \int_0^6 \left[\frac{4}{3}xy-\frac{5}{18}x^2y-\frac{7}{18}xy^2+\frac{1}{3}y^2-\frac{4}{27}y^3+2y\right]\Big|_0^{3-(1/2)x} dx \\
&= \int_0^6 5+\frac{1}{2}x-\frac{7}{12}x^2+\frac{13}{216}x^3\,dx \\
&= \left[5x+\frac{1}{4}x^2-\frac{7}{36}x^3+\frac{13}{864}x^4\right]\Big|_0^6 \\
&= \frac{33}{2}.
\end{aligned}
$$

Setting up limits on iterated integrals can require considerable geometric intuition. It is important to not only create carefully labeled figures, but also to think about how we wish to slice the solid. Further, note that when we say "we will integrate first with respect to x," by "first" we are referring to the innermost integral in the iterated integral. The next activity explores several different ways we might set up the integral in the preceding example.

Activity 11.7.3

There are several other ways we could have set up the integral to give the mass of the tetrahedron in Example 11.7.1.

1. How many different iterated integrals could be set up that are equal to the integral in Equation 11.7.2?
2. Set up an iterated integral, integrating first with respect to z, then x, then y that is equivalent to the integral in Equation 11.7.2. Before you write down the integral, think about Figure 11.7.6, and draw an appropriate two-dimensional image of an important projection.
3. Set up an iterated integral, integrating first with respect to y, then z, then x that is equivalent to the integral in Equation 11.7.2. As in (b), think carefully about the geometry first.
4. Set up an iterated integral, integrating first with respect to x, then y, then z that is equivalent to the integral in Equation 11.7.2.

Now that we have begun to understand how to set up iterated triple integrals, we can apply them to determine important quantities, such as those found in the next activity.

Activity 11.7.4

A solid S is bounded below by the square $z=0$, $-1\le x\le 1$, $-1\le y\le 1$ and above by the surface $z=2-x^2-y^2$. A picture of the solid is shown in Figure 11.7.4.

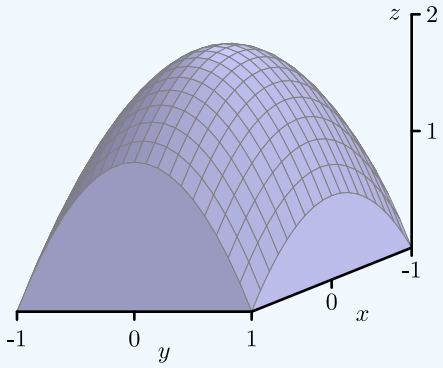


Figure 11.7.4. The solid bounded by the surface $z = 2 - x^2 - y^2$.

1. First, set up an iterated double integral to find the volume of the solid S as a double integral of a solid under a surface. Then set up an iterated triple integral that gives the volume of the solid S. You do not need to evaluate either integral. Compare the two approaches.
2. Set up (but do not evaluate) iterated integral expressions that will tell us the center of mass of S, if the density at point (x, y, z) is $\delta(x, y, z) = x^2 + 1$.
3. Set up (but do not evaluate) an iterated integral to find the average density on S using the density function from part (b).
4. Use technology appropriately to evaluate the iterated integrals you determined in (a), (b), and (c); does the location you determined for the center of mass make sense?

Summary

- Let $f = f(x, y, z)$ be a continuous function on a box $B = [a, b] \times [c, d] \times [r, s]$. The triple integral of f over B is defined as

$$\iiint_B f(x, y, z)\, dV = \lim_{\Delta V \to 0} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} f(x^*_{ijk}, y^*_{ijk}, z^*_{ijk}) \cdot \Delta V,$$

  where the triple Riemann sum is defined in the usual way. The definition of the triple integral naturally extends to non-rectangular solid regions S.

- The triple integral $\iiint_S f(x, y, z)\ dV$ can tell us
  - the volume of the solid S if $f(x, y, z) = 1$,
  - the mass of the solid S if f represents the density of S at the point (x, y, z).

  Moreover,

$$f_{\text{AVG}(S)} = \frac{1}{V(S)} \iiint_S f(x, y, z)\, dV,$$

  is the average value of f over S.

This page titled 11.7: Triple Integrals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.7: Triple Integrals by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

11.8: Triple Integrals in Cylindrical and Spherical Coordinates

Motivating Questions

- What are the cylindrical coordinates of a point, and how are they related to Cartesian coordinates?
- What is the volume element in cylindrical coordinates? How does this inform us about evaluating a triple integral as an iterated integral in cylindrical coordinates?
- What are the spherical coordinates of a point, and how are they related to Cartesian coordinates?
- What is the volume element in spherical coordinates? How does this inform us about evaluating a triple integral as an iterated integral in spherical coordinates?

We have encountered two different coordinate systems in $\mathbb{R}^2$ — the rectangular and polar coordinates systems — and seen how in certain situations, polar coordinates form a convenient alternative. In a similar way, there are two additional natural coordinate systems in $\mathbb{R}^3$. Given that we are already familiar with the Cartesian coordinate system for $\mathbb{R}^3$, we next investigate the cylindrical and spherical coordinate systems (each of which builds upon polar coordinates in $\mathbb{R}^2$). In what follows, we will see how to convert among the different coordinate systems, how to evaluate triple integrals using them, and some situations in which these other coordinate systems prove advantageous.

Preview Activity 11.8.1

In the following questions, we investigate the two new coordinate systems that are the subject of this section: cylindrical and spherical coordinates. Our goal is to consider some examples of how to convert from rectangular coordinates to each of these systems, and vice versa. Triangles and trigonometry prove to be particularly important.

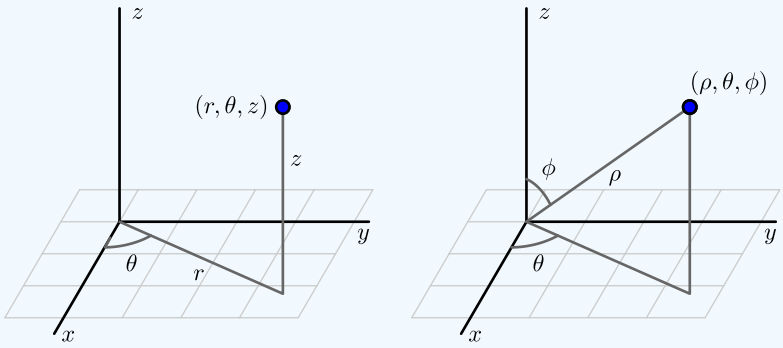


Figure 11.8.1. The cylindrical (left) and spherical (right) coordinates of a point.

1. The cylindrical coordinates of a point in $\mathbb{R}^3$ are given by (r, θ, z) where r and θ are the polar coordinates of the point (x, y) and z is the same z coordinate as in Cartesian coordinates. An illustration is given at left in Figure 11.8.1.
   1. Find cylindrical coordinates for the point whose Cartesian coordinates are $(-1, \sqrt{3}, 3)$. Draw a labeled picture illustrating all of the coordinates.
   2. Find the Cartesian coordinates of the point whose cylindrical coordinates are $\left(2, \frac{5\pi}{4}, 1\right)$. Draw a labeled picture illustrating all of the coordinates.
2. The spherical coordinates of a point in $\mathbb{R}^3$ are ρ (rho), θ, and ϕ (phi), where ρ is the distance from the point to the origin, θ has the same interpretation it does in polar coordinates, and ϕ is the angle between the positive z axis and the vector from the origin to the point, as illustrated at right in Figure 11.8.1. You should convince yourself that any point in $\mathbb{R}^3$ can be represented in spherical coordinates with $\rho \geq 0$, $0 \leq \theta < 2\pi$, and $0 \leq \phi \leq \pi$.

For the following questions, consider the point P whose Cartesian coordinates are $(-2, 2, \sqrt{8})$.

1. What is the distance from P to the origin? Your result is the value of ρ in the spherical coordinates of P.
2. Determine the point that is the projection of P onto the xy-plane. Then, use this projection to find the value of θ in the polar coordinates of the projection of P that lies in the plane. Your result is also the value of θ for the spherical coordinates of the point.
3. Based on the illustration in Figure 11.8.1, how is the angle ϕ determined by ρ and the z coordinate of P? Use a well-chosen right triangle to find the value of ϕ, which is the final component in the spherical coordinates of P. Draw a carefully labeled picture that clearly illustrates the values of ρ, θ, and ϕ in this example, along with the original rectangular coordinates of P.
4. Based on your responses to i., ii., and iii., if we are given the Cartesian coordinates (x, y, z) of a point Q, how are the values of ρ, θ, and ϕ in the spherical coordinates of Q determined by x, y, and z?

Cylindrical Coordinates

As we stated in Preview Activity 11.8.1, the cylindrical coordinates of a point are (r, θ, z), where r and θ are the polar coordinates of the point (x, y), and z is the same z coordinate as in Cartesian coordinates. The general situation is illustrated Figure 11.8.1.

Since we already know how to convert between rectangular and polar coordinates in the plane, and the z coordinate is identical in both Cartesian and cylindrical coordinates, the conversion equations between the two systems in $\mathbb{R}^3$ are essentially those we found for polar coordinates.

Converting between Cartesian and cylindrical coordinates

- If we are given the Cartesian coordinates (x, y, z) of a point P, then the cylindrical coordinates (r, θ, z) of P satisfy

$$x = r\cos(\theta) \qquad y = r\sin(\theta) \qquad \text{and} \qquad z = z.$$

- If we are given the cylindrical coordinates (r, θ, z) of a point P, then the Cartesian coordinates (x, y, z) of P satisfy

$$r^2 = x^2 + y^2 \qquad \tan(\theta) = \frac{y}{x} \qquad \text{and} \qquad z = z,$$

assuming $x \neq 0$.

Just as with rectangular coordinates, where we usually write z as a function of x and y to plot the resulting surface, in cylindrical coordinates, we often express z as a function of r and θ. In the following activity, we explore several basic equations in cylindrical coordinates and the corresponding surface each generates.

Activity 11.8.2

In this activity, we graph some surfaces using cylindrical coordinates. To improve your intuition and test your understanding, you should first think about what each graph should look like before you plot it using appropriate technology.

1. What familiar surface is described by the points in cylindrical coordinates with $r = 2$, $0 \leq \theta \leq 2\pi$, and $0 \leq z \leq 2$? How does this example suggest that we call these coordinates *cylindrical coordinates*? How does your answer change if we restrict θ to $0 \leq \theta \leq \pi$?
2. What familiar surface is described by the points in cylindrical coordinates with $\theta = 2$, $0 \leq r \leq 2$, and $0 \leq z \leq 2$?
3. What familiar surface is described by the points in cylindrical coordinates with $z = 2$, $0 \leq \theta \leq 2\pi$, and $0 \leq r \leq 2$?
4. Plot the graph of the cylindrical equation $z = r$, where $0 \leq \theta \leq 2\pi$ and $0 \leq r \leq 2$. What familiar surface results?
5. Plot the graph of the cylindrical equation $z = \theta$ for $0 \leq \theta \leq 4\pi$. What does this surface look like?

As the name and Activity 11.8.2 suggests, cylindrical coordinates are useful for describing surfaces that are cylindrical in nature.

Triple Integrals in Cylindrical Coordinates

To evaluate a triple integral $\iiint_S f(x,y,z)\ dV$ as an iterated integral in Cartesian coordinates, we use the fact that the volume element dV is equal to $dz\ dy\ dx$ (which corresponds to the volume of a small box). To evaluate a triple integral in cylindrical coordinates, we similarly must understand the volume element dV in cylindrical coordinates.

Activity 11.8.3

A picture of a cylindrical box, $B = \{(r,\theta,z) : r_1 \le r \le r_2, \theta_1 \le \theta \le \theta_2, z_1 \le z \le z_2\}$, is shown in Figure 11.8.2. Let $\Delta r = r_2 - r_1$, $\Delta\theta = \theta_2 - \theta_1$, and $\Delta z = z_2 - z_1$. We want to determine the volume ΔV of B in terms of Δr, $\Delta\theta$, Δz, r, θ, and z.

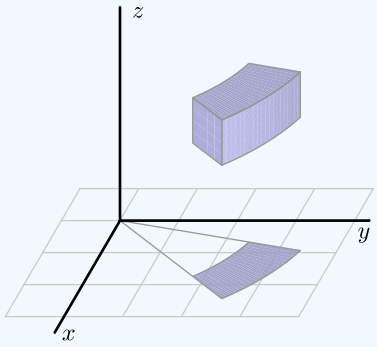


Figure 11.8.2. A cylindrical box.

1. Appropriately label Δr, $\Delta\theta$, and Δz in Figure 11.8.2.
2. Let ΔA be the area of the projection of the box, B, onto the xy-plane, which is shaded blue in Figure 11.8.2. Recall that we previously determined the area ΔA in polar coordinates in terms of r, Δr, and $\Delta\theta$. In light of the fact that we know ΔA and that z is the standard z coordinate from Cartesian coordinates, what is the volume ΔV in cylindrical coordinates?

Activity 11.8.3 demonstrates that the volume element dV in cylindrical coordinates is given by $dV = r\ dz\ dr\ d\theta$, and hence the following rule holds in general.

Triple integrals in cylindrical coordinates

Given a continuous function $f = f(x,y,z)$ over a region S in $\mathbb{R}^3$,

$$\iiint_S f(x,y,z)\,dV = \iiint_S f(r\cos(\theta), r\sin(\theta), z)\, r\,dz\,dr\,d\theta.$$

The latter expression is an *iterated integral in cylindrical coordinates*.

Of course, to complete the task of writing an iterated integral in cylindrical coordinates, we need to determine the limits on the three integrals: θ, r, and z. In the following activity, we explore how to do this in several situations where cylindrical coordinates are natural and advantageous.

Activity 11.8.4

In this activity we work with triple integrals in cylindrical coordinates.

1. Let S be the solid bounded above by the graph of $z = x^2 + y^2$ and below by $z = 0$ on the unit disk in the xy-plane.
    1. The projection of the solid S onto the xy-plane is a disk. Describe this disk using polar coordinates.

2. Now describe the surfaces bounding the solid S using cylindrical coordinates.
3. Determine an iterated triple integral expression in cylindrical coordinates that gives the volume of S. You do not need to evaluate this integral.

2. Suppose the density of the cone defined by $r = 1 - z$, with $z \geq 0$, is given by $\delta(r, \theta, z) = z$. A picture of the cone is shown at left in Figure 11.8.3, and the projection of the cone onto the xy-plane in given at right in Figure 11.8.3. Set up an iterated integral in cylindrical coordinates that gives the mass of the cone. You do not need to evaluate this integral.

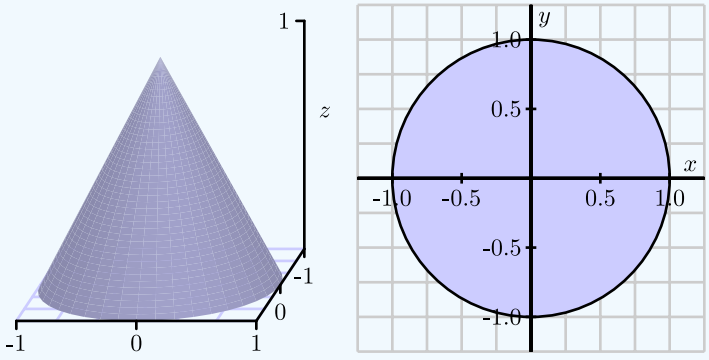


Figure 11.8.3. The cylindrical cone $r = 1 - z$ and its projection onto the xy-plane.

3. Determine an iterated integral expression in cylindrical coordinates whose value is the volume of the solid bounded below by the cone $z = \sqrt{x^2 + y^2}$ and above by the cone $z = 4 - \sqrt{x^2 + y^2}$. A picture is shown in Figure 11.8.4. You do not need to evaluate this integral.

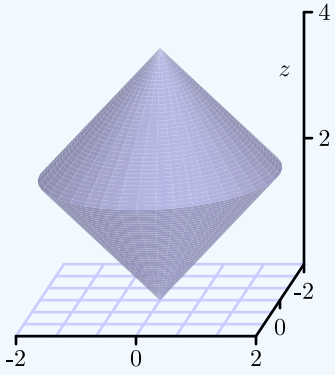


Figure 11.8.4. A solid bounded by the cones $z = \sqrt{x^2 + y^2}$ and $z = 4 - \sqrt{x^2 + y^2}$.

Spherical Coordinates

As we saw in Preview Activity 11.8.1, the spherical coordinates of a point in 3-space have the form (ρ, θ, ϕ), where ρ is the distance from the point to the origin, θ has the same meaning as in polar coordinates, and ϕ is the angle between the positive z axis and the vector from the origin to the point. The overall situation is illustrated at right in Figure 11.8.1.

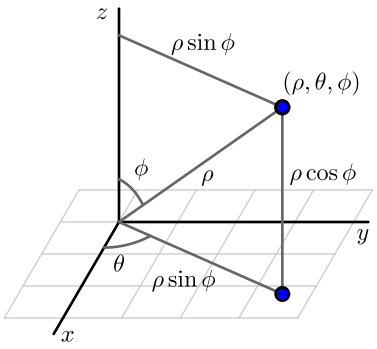


Figure 11.8.5. Converting from spherical to Cartesian coordinates.

The example in Preview Activity 11.8.1 and Figure 11.8.5 suggest how to convert between Cartesian and spherical coordinates.

Coverting between Cartesian and spherical coordinates

- If we are given the Cartesian coordinates (x, y, z) of a point P, then the spherical coordinates (ρ, θ, ϕ) of P satisfy
$$\rho = \sqrt{x^2 + y^2 + z^2} \qquad \tan(\theta) = \frac{y}{x} \qquad \text{and} \qquad \cos(\phi) = \frac{z}{\rho},$$
where in the latter two equations, we require $x \neq 0$ and $\rho \neq 0$.
- If we are given the spherical coordinates (ρ, θ, ϕ) of a point P, then the Cartesian coordinates (x, y, z) of P satisfy
$$x = \rho \sin(\phi) \cos(\theta) \qquad y = \rho \sin(\phi) \sin(\theta) \qquad \text{and} \qquad z = \rho \cos(\phi).$$

When it comes to thinking about particular surfaces in spherical coordinates, similar to our work with cylindrical and Cartesian coordinates, we usually write ρ as a function of θ and ϕ; this is a natural analog to polar coordinates, where we often think of our distance from the origin in the plane as being a function of θ. In spherical coordinates, we likewise often view ρ as a function of θ and ϕ, thus viewing distance from the origin as a function of two key angles.

In the following activity, we explore several basic equations in spherical coordinates and the surfaces they generate.

Activity 11.8.5

In this activity, we graph some surfaces using spherical coordinates. To improve your intuition and test your understanding, you should first think about what each graph should look like before you plot it using appropriate technology.

1. What familiar surface is described by the points in spherical coordinates with $\rho = 1$, $0 \leq \theta \leq 2\pi$, and $0 \leq \phi \leq \pi$? How does this particular example demonstrate the reason for the name of this coordinate system? What if we restrict ϕ to $0 \leq \phi \leq \frac{\pi}{2}$?
2. What familiar surface is described by the points in spherical coordinates with $\phi = \frac{\pi}{3}$, $0 \leq \rho \leq 1$, and $0 \leq \theta \leq 2\pi$?
3. What familiar shape is described by the points in spherical coordinates with $\theta = \frac{\pi}{6}$, $0 \leq \rho \leq 1$, and $0 \leq \phi \leq \pi$?
4. Plot the graph of $\rho = \theta$, for $0 \leq \phi \leq \pi$ and $0 \leq \theta \leq 2\pi$. How does the resulting surface appear?

As the name and Activity 11.8.5 indicate, spherical coordinates are particularly useful for describing surfaces that are spherical in nature; they are also convenient for working with certain conical surfaces.

Triple Integrals in Spherical Coordinates

As with rectangular and cylindrical coordinates, a triple integral $\iiint_S f(x, y, z)\ dV$ in spherical coordinates can be evaluated as an iterated integral once we understand the volume element dV.

Activity 11.8.6

To find the volume element dV in spherical coordinates, we need to understand how to determine the volume of a spherical box of the form $\rho_1 \le \rho \le \rho_2$ (with $\Delta\rho = \rho_2 - \rho_1$), $\phi_1 \le \phi \le \phi_2$ (with $\Delta\phi = \phi_2 - \phi_1$), and $\theta_1 \le \theta \le \theta_2$ (with $\Delta\theta = \theta_2 - \theta_1$). An illustration of such a box is given at left in Figure 11.8.6. This spherical box is a bit more complicated than the cylindrical box we encountered earlier. In this situation, it is easier to approximate the volume ΔV than to compute it directly. Here we can approximate the volume ΔV of this spherical box with the volume of a Cartesian box whose sides have the lengths of the sides of this spherical box. In other words,

$$\Delta V \approx |PS|\ |\widehat{PR}|\ |\widehat{PQ}|,$$

where $|\widehat{PR}|$ denotes the length of the circular arc from P to R.

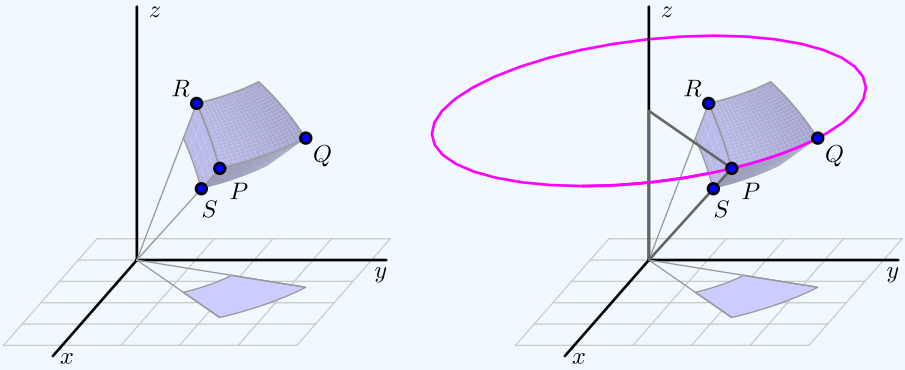


Figure 11.8.6. Left: A spherical box. Right: A spherical volume element.

1. What is the length $|PS|$ in terms of ρ?
2. What is the length of the arc $\widehat{PR}$? (Hint: The arc $\widehat{PR}$ is an arc of a circle of radius ρ_2, and arc length along a circle is the product of the angle measure (in radians) and the circle's radius.)
3. What is the length of the arc $\widehat{PQ}$? (Hint: The arc $\widehat{PQ}$ lies on a horizontal circle as illustrated at right in Figure 11.8.6. What is the radius of this circle?)
4. Use your work in (a), (b), and (c) to determine an approximation for ΔV in spherical coordinates.

Letting $\Delta\rho$, $\Delta\theta$, and $\Delta\phi$ go to 0, it follows from the final result in Activity 11.8.6 that $dV = \rho^2\ \sin(\phi)\ d\rho\ d\theta\ d\phi$ in spherical coordinates, and thus allows us to state the following general rule.

Triple integrals in spherical coordinates

Given a continuous function $f = f(x, y, z)$ over a region S in $\mathbb{R}^3$, the triple integral $\iiint_S f(x, y, z)\ dV$ is converted to the integral

$$\iiint_S f(\rho\sin(\phi)\cos(\theta), \rho\sin(\phi)\sin(\theta), \rho\cos(\phi))\ \rho^2\sin(\phi)\, d\rho\, d\theta\, d\phi$$

in spherical coordinates.

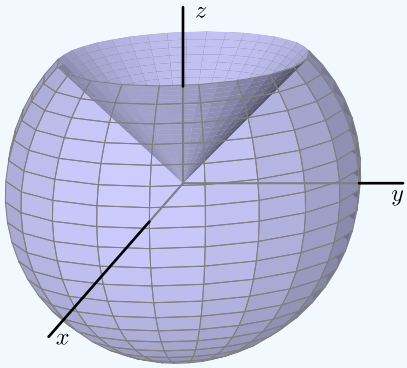
The latter expression is an *iterated integral in spherical coordinates*.

Finally, in order to actually evaluate an iterated integral in spherical coordinates, we must of course determine the limits of integration in ϕ, θ, and ρ. The process is similar to our earlier work in the other two coordinate systems.

Activity 11.8.7

We can use spherical coordinates to help us more easily understand some natural geometric objects.

1. Recall that the sphere of radius a has spherical equation $\rho = a$. Set up and evaluate an iterated integral in spherical coordinates to determine the volume of a sphere of radius a.
2. Set up, but do not evaluate, an iterated integral expression in spherical coordinates whose value is the mass of the solid obtained by removing the cone $\phi = \frac{\pi}{4}$ from the sphere $\rho = 2$ if the density δ at the point (x, y, z) is $\delta(x, y, z) = \sqrt{x^2 + y^2 + z^2}$. An illustration of the solid is shown in Figure 11.8.7.



1.

Figure 11.8.7. The solid cut from the sphere $\rho = 2$ by the cone $\phi = \frac{\pi}{4}$.

Summary

- The cylindrical coordinates of a point P are (r, θ, z) where r is the distance from the origin to the projection of P onto the xy -plane, θ is the angle that the projection of P onto the xy-plane makes with the positive x-axis, and z is the vertical distance from P to the projection of P onto the xy-plane. When P has rectangular coordinates (x, y, z), it follows that its cylindrical coordinates are given by

$$r^2 = x^2 + y^2, \qquad \tan(\theta) = \frac{y}{x}, \qquad z = z.$$

  When P has given cylindrical coordinates (r, θ, z), its rectangular coordinates are

$$x = r\cos(\theta), \qquad y = r\sin(\theta), \qquad z = z.$$

- The volume element dV in cylindrical coordinates is $dV = r\ dz\ dr\ d\theta$. Hence, a triple integral $\iiint_S f(x, y, z)\ dA$ can be evaluated as the iterated integral

$$\iiint_S f(r\cos(\theta), r\sin(\theta), z)\, r\, dz\, dr\, d\theta.$$

- The spherical coordinates of a point P in 3-space are ρ (rho), θ, and ϕ (phi), where ρ is the distance from P to the origin, θ is the angle that the projection of P onto the xy-plane makes with the positive x-axis, and ϕ is the angle between the positive z axis and the vector from the origin to P. When P has Cartesian coordinates (x, y, z), the spherical coordinates are given by

$$\rho^2 = x^2 + y^2 + z^2, \qquad \tan(\theta) = \frac{y}{x}, \qquad \cos(\phi) = \frac{z}{\rho}.$$

  Given the point P in spherical coordinates (ρ, ϕ, θ), its rectangular coordinates are

$$x = \rho\sin(\phi)\cos(\theta), \qquad y = \rho\sin(\phi)\sin(\theta), \qquad z = \rho\cos(\phi).$$

- The volume element dV in spherical coordinates is $dV = \rho^2\sin(\phi)\ d\rho\ d\theta\ d\phi$. Thus, a triple integral $\iiint_S f(x, y, z)\ dA$ can be evaluated as the iterated integral

$$\iiint_S f(\rho\sin(\phi)\cos(\theta), \rho\sin(\phi)\sin(\theta), \rho\cos(\phi))\, \rho^2\sin(\phi)\, d\rho\, d\theta\, d\phi.$$

This page titled 11.8: Triple Integrals in Cylindrical and Spherical Coordinates is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.8: Triple Integrals in Cylindrical and Spherical Coordinates by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

11.9: Change of Variables

Motivating Questions

- What is a change of variables?
- What is the Jacobian, and how is it related to a change of variables?

In single variable calculus, we encountered the idea of a change of variable in a definite integral through the method of substitution. For example, given the definite integral

$$\int_0^2 2x(x^2+1)^3\,dx,$$

we naturally consider the change of variable $u = x^2 + 1$. From this substitution, it follows that $du = 2x\ dx$, and since $x = 0$ implies $u = 1$ and $x = 2$ implies $u = 5$, we have transformed the original integral in x into a new integral in u. In particular,

$$\int_0^2 2x(x^2+1)^3\,dx = \int_1^5 u^3\,du.$$

The latter integral, of course, is far easier to evaluate.

Through our work with polar, cylindrical, and spherical coordinates, we have already implicitly seen some of the issues that arise in using a change of variables with two or three variables present. In what follows, we seek to understand the general ideas behind any change of variables in a multiple integral.

Preview Activity 11.9.1

Consider the double integral

$$I = \iint_D x^2 + y^2\,dA, \tag{11.9.1}$$

where D is the upper half of the unit disk.

1. 1. Write the double integral I given in Equation 11.9.1 as an iterated integral in rectangular coordinates.
   2. Write the double integral I given in Equation 11.9.1 as an iterated integral in polar coordinates.
2. When we write the double integral 11.9.1 as an iterated integral in polar coordinates we make a change of variables, namely

   $$x = r\cos(\theta) \qquad \text{and} \qquad y = r\sin(\theta). \tag{11.9.2}$$

   We also then have to change dA to $r\ dr\ d\theta$. This process also identifies a "polar rectangle" $[r_1, r_2] \times [\theta_1, \theta_2]$ with the original Cartesian rectangle, under the transformation[1] in Equation 11.9.2. The vertices of the polar rectangle are transformed into the vertices of a closed and bounded region in rectangular coordinates.

   To work with a numerical example, let's now consider the polar rectangle P given by $[1, 2] \times [\frac{\pi}{6}, \frac{\pi}{4}]$, so that $r_1 = 1$, $r_2 = 2$, $\theta_1 = \frac{\pi}{6}$, and $\theta_2 = \frac{\pi}{4}$.

   1. Use the transformation determined by the equations in 11.9.2 to find the rectangular vertices that correspond to the polar vertices in the polar rectangle P. In other words, by substituting appropriate values of r and θ into the two equations in 11.9.2, find the values of the corresponding x and y coordinates for the vertices of the polar rectangle P. Label the point that corresponds to the polar vertex (r_1, θ_1) as (x_1, y_1), the point corresponding to the polar vertex (r_2, θ_1) as (x_2, y_2), the point corresponding to the polar vertex (r_1, θ_2) as (x_3, y_3), and the point corresponding to the polar vertex (r_2, θ_2) as (x_4, y_4).

2. Draw a picture of the figure in rectangular coordinates that has the points (x_1, y_1), (x_2, y_2), (x_3, y_3), and (x_4, y_4) as vertices. (Note carefully that because of the trigonometric functions in the transformation, this region will not look like a Cartesian rectangle.) What is the area of this region in rectangular coordinates? How does this area compare to the area of the original polar rectangle?

A *transformation* is another name for function: here, the equations $x = r\cos(\theta)$ and $y = r\sin(\theta)$ define a function T by $T(r, \theta) = (r\cos(\theta), r\sin(\theta))$ so that T is a function (transformation) from $\mathbb{R}^2$ to $\mathbb{R}^2$. We view this transformation as mapping a version of the xy-plane where the axes are viewed as representing r and θ (the $r\theta$-plane) to the familiar xy-plane.

Change of Variables in Polar Coordinates

The general idea behind a change of variables is suggested by Preview Activity 11.9.1. There, we saw that in a change of variables from rectangular coordinates to polar coordinates, a polar rectangle $[r_1, r_2] \times [\theta_1, \theta_2]$ gets mapped to a Cartesian rectangle under the transformation

$$x = r\cos(\theta) \qquad \text{and} \qquad y = r\sin(\theta).$$

The vertices of the polar rectangle P are transformed into the vertices of a closed and bounded region P' in rectangular coordinates. If we view the standard coordinate system as having the horizontal axis represent r and the vertical axis represent θ, then the polar rectangle P appears to us at left in Figure 11.9.1. The image P' of the polar rectangle P under the transformation given by 11.9.2 is shown at right in Figure 11.9.1. We thus see that there is a correspondence between a simple region (a traditional, right-angled rectangle) and a more complicated region (a fraction of an annulus) under the function T given by $T(r, \theta) = (r\cos(\theta), r\sin(\theta))$.

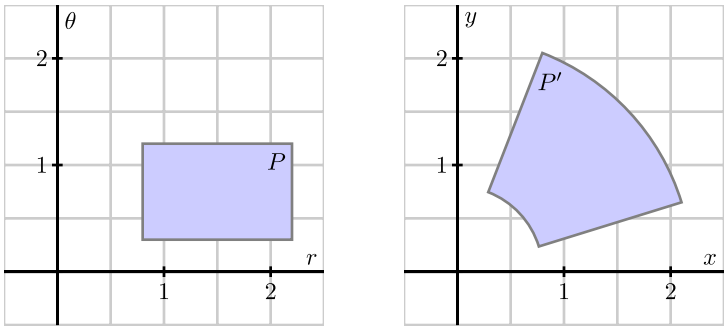


Figure 11.9.1. A rectangle P and its image P'.

Furthermore, as Preview Activity 11.9.1 suggests, it follows generally that for an original polar rectangle $P = [r_1, r_2] \times [\theta_1, \theta_2]$, the area of the transformed rectangle P' is given by $\frac{r_2+r_1}{2}\Delta r\Delta\theta$. Therefore, as Δr and $\Delta\theta$ go to 0 this area becomes the familiar area element $dA = r\ dr\ d\theta$ in polar coordinates. When we proceed to working with other transformations for different changes in coordinates, we have to understand how the transformation affects area so that we may use the correct area element in the new system of variables.

General Change of Coordinates

We first focus on double integrals. As with single integrals, we may be able to simplify a double integral of the form

$$\iint_D f(x, y)\, dA$$

by making a change of variables (that is, a substitution) of the form

$$x = x(s, t) \qquad \text{and} \qquad y = y(s, t)$$

where x and y are functions of new variables s and t. This transformation introduces a correspondence between a problem in the xy-plane and one in the the st-plane. The equations $x = x(s, t)$ and $y = y(s, t)$ convert s and t to x and y; we call these formulas

the *change of variable* formulas. To complete the change to the new s, t variables, we need to understand the area element, dA, in this new system. The following activity helps to illustrate the idea.

Activity 11.9.2

Consider the change of variables

$$x = s + 2t \qquad \text{and} \qquad y = 2s + \sqrt{t}.$$

Let's see what happens to the rectangle $T = [0, 1] \times [1, 4]$ in the st-plane under this change of variable.

1. Draw a labeled picture of T in the st-plane.
2. Find the image of the st-vertex $(0, 1)$ in the xy-plane. Likewise, find the respective images of the other three vertices of the rectangle T: $(0, 4)$, $(1, 1)$, and $(1, 4)$.
3. In the xy-plane, draw a labeled picture of the image, T', of the original st-rectangle T. What appears to be the shape of the image, T'?
4. To transform an integral with a change of variables, we need to determine the area element dA for image of the transformed rectangle. Note that T' is not exactly a parallelogram since the equations that define the transformation are not linear. But we can approximate the area of T' with the area of a parallelogram. How would we find the area of a parallelogram that approximates the area of the xy-figure T'? (Hint: Remember what the cross product of two vectors tells us.)

Activity 11.9.2 presents the general idea of how a change of variables works. We partition a rectangular domain in the st system into subrectangles. Let $T = [a, b] \times [a + \Delta s, b + \Delta t]$ be one of these subrectangles. Then we transform this into a region T' in the standard xy Cartesian coordinate system. The region T' is called the *image* of T; the region T is the *pre-image* of T'. Although the sides of this xy region T' aren't necessarily straight (linear), we will approximate the element of area dA for this region with the area of the parallelogram whose sides are given by the vectors $\mathbf{v}$ and $\mathbf{w}$, where $\mathbf{v}$ is the vector from $(x(a, b), y(a, b))$ to $(x(a + \Delta s, b), y(a + \Delta s, b))$, and $\mathbf{w}$ is the vector from $(x(a, b), y(a, b))$ to $(x(a, b + \Delta t), y(a, b + \Delta t))$.

An example of an image T' in the xy-plane that results from a transformation of a rectangle T in the st-plane is shown in Figure 11.9.2.

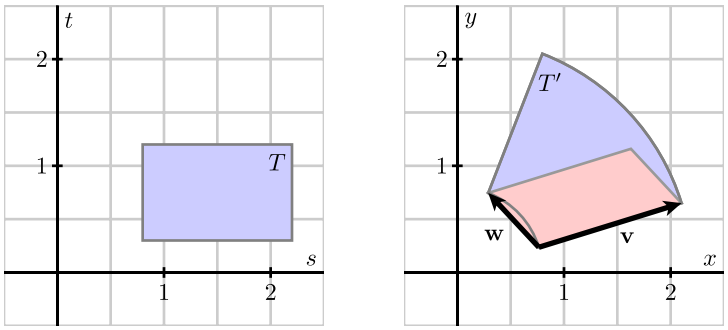


Figure 11.9.2. Approximating an area of an image resulting from a transformation.

The components of the vector $\mathbf{v}$ are

$$\mathbf{v} = \langle x(a + \Delta s, b) - x(a, b), y(a + \Delta s, b) - y(a, b), 0 \rangle$$

and similarly those for $\mathbf{w}$ are

$$\mathbf{w} = \langle x(a, b + \Delta t) - x(a, b), y(a, b + \Delta s) - y(a, b), 0 \rangle .$$

Slightly rewriting $\mathbf{v}$ and $\mathbf{w}$, we have

$$\mathbf{v} = \left\langle \frac{x(a+\Delta s, b) - x(a,b)}{\Delta s}, \frac{y(a+\Delta s, b) - y(a,b)}{\Delta s}, 0 \right\rangle \Delta s, \text{ and}$$
$$\mathbf{w} = \left\langle \frac{x(a, b+\Delta t) - x(a,b)}{\Delta t}, \frac{y(a, b+\Delta s) - y(a,b)}{\Delta t}, 0 \right\rangle \Delta t.$$

For small Δs and Δt, the definition of the partial derivative tells us that

$$\mathbf{v} \approx \left\langle \frac{\partial x}{\partial s}(a,b), \frac{\partial y}{\partial s}(a,b), 0 \right\rangle \Delta s \qquad \text{and} \qquad \mathbf{w} \approx \left\langle \frac{\partial x}{\partial t}(a,b), \frac{\partial y}{\partial t}(a,b), 0 \right\rangle \Delta t.$$

Recall that the area of the parallelogram with sides $\mathbf{v}$ and $\mathbf{w}$ is the length of the cross product of the two vectors, $|\mathbf{v} \times \mathbf{w}|$. From this, we observe that

$$\begin{aligned} \mathbf{v} \times \mathbf{w} &\approx \left\langle \frac{\partial x}{\partial s}(a,b), \frac{\partial y}{\partial s}(a,b), 0 \right\rangle \Delta s \times \left\langle \frac{\partial x}{\partial t}(a,b), \frac{\partial y}{\partial t}(a,b), 0 \right\rangle \Delta t \\ &= \left\langle 0,\ 0,\ \frac{\partial x}{\partial s}(a,b)\frac{\partial y}{\partial t}(a,b) - \frac{\partial x}{\partial t}(a,b)\frac{\partial y}{\partial s}(a,b) \right\rangle \Delta s\, \Delta t. \end{aligned}$$

Finally, by computing the magnitude of the cross product, we see that

$$\begin{aligned} |\mathbf{v} \times \mathbf{w}| &\approx \left| \left\langle 0, 0, \frac{\partial x}{\partial s}(a,b)\frac{\partial y}{\partial t}(a,b) - \frac{\partial x}{\partial t}(a,b)\frac{\partial y}{\partial s}(a,b) \right\rangle \Delta s\, \Delta t \right| \\ &= \left| \frac{\partial x}{\partial s}(a,b)\frac{\partial y}{\partial t}(a,b) - \frac{\partial x}{\partial t}(a,b)\frac{\partial y}{\partial s}(a,b) \right| \Delta s\, \Delta t. \end{aligned}$$

Therefore, as the number of subdivisions increases without bound in each direction, Δs and Δt both go to zero, and we have

$$dA = \left| \frac{\partial x}{\partial s}\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\frac{\partial y}{\partial s} \right| ds\, dt. \tag{11.9.3}$$

Equation 11.9.3 hence determines the general change of variable formula in a double integral, and we can now say that

$$\iint_T f(x,y)\, dy\, dx = \iint_{T'} f(x(s,t), y(s,t)) \left| \frac{\partial x}{\partial s}\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\frac{\partial y}{\partial s} \right| ds\, dt.$$

The quantity

$$\frac{\partial x}{\partial s}\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\frac{\partial y}{\partial s}$$

is called the *Jacobian*, and we denote the Jacobian using the shorthand notation

$$\frac{\partial(x,y)}{\partial(s,t)} = \frac{\partial x}{\partial s}\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\frac{\partial y}{\partial s}.$$

Recall from Section 9.4 that we can also write this Jacobian as the determinant of the 2×2 matrix $\begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{bmatrix}$. Note that, as discussed in Section 9.4, the absolute value of the determinant of $\begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{bmatrix}$ is the area of the parallelogram determined by the vectors $\mathbf{v}$ and $\mathbf{w}$, and so the area element dA in xy-coordinates is also represented by the area element $\left|\frac{\partial(x,y)}{\partial(s,t)}\right| \; ds \; dt$ in st-coordinates, and $\left|\frac{\partial(x,y)}{\partial(s,t)}\right|$ is the factor by which the transformation magnifies area.

To summarize, the preceding change of variable formula that we have derived now follows.

Change of Variables in a Double Integral

Suppose a change of variables $x = x(s,t)$ and $y = y(s,t)$ transforms a closed and bounded region R in the st-plane into a closed and bounded region R' in the xy-plane. Under modest conditions (that are studied in advanced calculus), it follows that

$$\iint_{R'} f(x,y)\,dA = \iint_{R} f(x(s,t),y(s,t))\left|\frac{\partial(x,y)}{\partial(s,t)}\right| ds\,dt.$$

Activity 11.9.3

Find the Jacobian when changing from rectangular to polar coordinates. That is, for the transformation given by $x = r\cos(\theta)$, $y = r\sin(\theta)$, determine a simplified expression for the quantity

$$\frac{\partial x}{\partial r}\frac{\partial y}{\partial \theta} - \frac{\partial x}{\partial \theta}\frac{\partial y}{\partial r}.$$

What do you observe about your result? How is this connected to our earlier work with double integrals in polar coordinates?

Activity 11.9.4

Let D' be the region in the xy-plane bounded by the lines $y = 0$, $x = 0$, and $x + y = 1$. We will evaluate the double integral

$$\iint_{D'} \sqrt{x+y}(x-y)^2\,dA \tag{11.9.4}$$

with a change of variables.

1. Sketch the region D' in the xy-plane.
2. We would like to make a substitution that makes the integrand easier to antidifferentiate. Let $s = x + y$ and $t = x - y$. Explain why this should make antidifferentiation easier by making the corresponding substitutions and writing the new integrand in terms of s and t.
3. Solve the equations $s = x + y$ and $t = x - y$ for x and y. (Doing so determines the standard form of the transformation, since we will have x as a function of s and t, and y as a function of s and t.)
4. To actually execute this change of variables, we need to know the st-region D that corresponds to the xy-region D'.
   1. What st equation corresponds to the xy equation $x + y = 1$?
   2. What st equation corresponds to the xy equation $x = 0$?
   3. What st equation corresponds to the xy equation $y = 0$?
   4. Sketch the st region D that corresponds to the xy domain D'.
5. Make the change of variables indicated by $s = x + y$ and $t = x - y$ in the double integral (11.9.4) and set up an iterated integral in st variables whose value is the original given double integral. Finally, evaluate the iterated integral.

Change of Variables in a Triple Integral

The argument for the change of variable formula for triple integrals is complicated, and we will not go into the details. The general process, though, is the same as the two-dimensional case. Given a solid S' in the xyz-coordinate system in $\mathbb{R}^3$, a change of variables transformation $x = x(s,t,u)$, $y = y(s,t,u)$, and $z = z(s,t,u)$ transforms S' into a region S in stu-coordinates. Any function $f = f(x,y,z)$ defined on S' can be considered as a function $f = f(x(s,t,u),y(s,y,u),z(s,t,u))$ in stu-coordinates defined on S. The volume element dV in xyz-coordinates cooresponds to a scaled volume element in stu-coordinates, where the scale factor is given by the absolute value of the Jacobian, $\frac{\partial(x,y,z)}{\partial(s,t,u)}$, which is the determinant of the 3×3 matrix

$$\begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} & \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} & \frac{\partial y}{\partial u} \\ \frac{\partial z}{\partial s} & \frac{\partial z}{\partial t} & \frac{\partial z}{\partial u} \end{bmatrix}.$$

(Recall that this determinant was introduced in Section 9.4.) That is, $\frac{\partial(x,y,z)}{\partial(s,t,u)}$ is given by

$$\frac{\partial x}{\partial s}\left[\frac{\partial y}{\partial t}\frac{\partial z}{\partial u} - \frac{\partial y}{\partial u}\frac{\partial z}{\partial t}\right] - \frac{\partial x}{\partial t}\left[\frac{\partial y}{\partial s}\frac{\partial z}{\partial u} - \frac{\partial y}{\partial u}\frac{\partial z}{\partial s}\right] + \frac{\partial x}{\partial u}\left[\frac{\partial y}{\partial s}\frac{\partial z}{\partial t} - \frac{\partial y}{\partial t}\frac{\partial z}{\partial s}\right].$$

To summarize,

Change of Variables in a Triple Integral

Suppose a change of variables $x = x(s,t,u)$, $y = y(s,t,u)$, and $z = z(s,t,u)$ transforms a closed and bounded region S in stu-coordinates into a closed and bounded region S' in xyz-coordinates. Under modest conditions (that are studied in advanced calculus), the triple integral $\iiint_{S'} f(x,y,z)\ dV$ is equal to

$$\iiint_S f(x(s,t,u), y(s,t,u), z(s,t,u)) \left|\frac{\partial(x,y,z)}{\partial(s,t,u)}\right| ds\, dt\, du.$$

Activity 11.9.5

Find the Jacobian when changing from Cartesian to cylindrical coordinates. That is, for the transformation given by $x = r\cos(\theta)$, $y = r\sin(\theta)$, and $z = z$, determine a simplified expression for the quantity

$$\frac{\partial(x,y,z)}{\partial(r,\theta,z)}.$$

What do you observe about your result? How is this connected to our earlier work with triple integrals in cylindrial coordinates?

Activity 11.9.6

Consider the solid S' defined by the inequalities $0 \le x \le 2$, $\frac{x}{2} \le y \le \frac{x}{2} + 1$, and $0 \le z \le 6$. Consider the transformation defined by $s = \frac{x}{2}$, $t = \frac{x-2y}{2}$, and $u = \frac{z}{3}$. Let $f(x,y,x) = x - 2y + z$.

1. The transformation turns the solid S' in xyz-coordinates into a box S in stu-coordinates. Apply the transformation to the boundries of the solid S' to find stu-coordinatte descriptions of the box S.
2. Find the Jacobian $\frac{\partial(x,y,z)}{\partial(s,t,u)}$.
3. Use the transformation to perform a change of variables and evaluate $\iiint_{S'} f(x,y,z)\ dV$ by evaluating

$$\iiint_S f(x(s,t,u), y(s,t,u), z(s,t,u)) \left|\frac{\partial(x,y,z)}{\partial(s,t,u)}\right| ds\, dt\, du.$$

Summary

- If an integral is described in terms of one set of variables, we may write that set of variables in terms of another set of the same number of variables. If the new variables are chosen appropriately, the transformed integral may be easier to evaluate.
- The Jacobian is a scalar function that relates the area or volume element in one coordinate system to the corresponding element in a new system determined by a change of variables.

This page titled 11.9: Change of Variables is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.9: Change of Variables by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/ACM.html.

12: APPENDICES

12.1: A- Short Table of Integrals
12.2: B- Answers to Activities
12.3: C- Answers to Selected Exercises

This page titled 12: Appendices is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University).

- 12: Appendices by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0.

12.1: A- Short Table of Integrals

Appendix A A Short Table of Integrals

a.

$$\int \frac{du}{a^2+u^2} = \frac{1}{a}\arctan\left(\frac{u}{a}\right) + C \tag{12.1.1}$$

b.

$$\int \frac{du}{\sqrt{u^2 \pm a^2}} = \ln\left|u + \sqrt{u^2 \pm a^2}\right| + C \tag{12.1.2}$$

c.

$$\int \sqrt{u^2 \pm a^2}du = \frac{u}{2}\sqrt{u^2 \pm a^2} \pm \frac{a^2}{2}\ln\left|u + \sqrt{u^2 \pm a^2}\right| + C \tag{12.1.3}$$

d.

$$\int \frac{u^2 du}{\sqrt{u^2 \pm a^2}} = \frac{u}{2}\sqrt{u^2 \pm a^2} \mp \frac{a^2}{2}\ln\left|u + \sqrt{u^2 \pm a^2}\right| + C \tag{12.1.4}$$

e.

$$\int \frac{du}{u\sqrt{u^2+a^2}} = -\frac{1}{a}\ln\left|\frac{a+\sqrt{u^2+a^2}}{u}\right| + C \tag{12.1.5}$$

f.

$$\int \frac{du}{u\sqrt{u^2-a^2}} = \frac{1}{a}\text{arcsec}\left(\frac{u}{a}\right) + C \tag{12.1.6}$$

g.

$$\int \frac{du}{\sqrt{a^2-u^2}} = \arcsin\left(\frac{u}{a}\right) + C \tag{12.1.7}$$

h.

$$\int \sqrt{a^2-u^2}du = \frac{u}{2}\sqrt{a^2-u^2} + \frac{a^2}{2}\arcsin\left(\frac{u}{a}\right) + C \tag{12.1.8}$$

i.

$$\int \frac{u^2}{\sqrt{a^2-u^2}}du = -\frac{u}{2}\sqrt{a^2-u^2} + \frac{a^2}{2}\arcsin\left(\frac{u}{a}\right) + C \tag{12.1.9}$$

j.

$$\int \frac{du}{u\sqrt{a^2-u^2}} = -\frac{1}{a}\ln\left|\frac{a+\sqrt{a^2-u^2}}{u}\right| + C \tag{12.1.10}$$

k.

$$\int \frac{du}{u^2\sqrt{a^2-u^2}} = -\frac{\sqrt{a^2-u^2}}{a^2 u} + C \tag{12.1.11}$$

This page titled 12.1: A- Short Table of Integrals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University).

12.2: B- Answers to Activities

This page titled 12.2: B- Answers to Activities is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University).

12.3: C- Answers to Selected Exercises

Solutions C Answers to Selected Exercises

This appendix contains answers to all non-WeBWorK exercises in the text. For WeBWorK exercises, please use the HTML version of the text for access to answers and solutions.

Chapter 1 Understanding the Derivative

Section 1.1 How do we measure velocity?

Exercises

Answer

Answer

Answer

Section 1.2 The notion of limit

Exercises

Answer

Answer

Answer

Answer

Section 1.3 The derivative of a function at a point

Exercises

Answer

Answer

Answer

Answer

Section 1.4 The derivative function

Exercises

Answer

Answer

Answer

Answer

Section 1.5 Interpreting, estimating, and using the derivative

Exercises

Answer

Answer

Answer

Answer

Section 1.6 The second derivative

Exercises

Answer

Answer

Answer

Answer

Section 1.7 Limits, Continuity, and Differentiability

Exercises

Answer

Answer

Answer

Answer

Section 1.8 The Tangent Line Approximation

Exercises

Answer

Answer

Answer

Answer

Chapter 2 Computing Derivatives

Section 2.1 Elementary derivative rules

Exercises

Answer

Answer

Answer

Answer

Section 2.2 The sine and cosine functions

Exercises

Answer

Answer

Answer

Section 2.3 The product and quotient rules

Exercises

Answer

Answer

Answer

Answer

Answer

Section 2.4 Derivatives of other trigonometric functions

Exercises

Answer

Answer

Answer

Section 2.5 The chain rule

Exercises

Answer

Answer

Answer

Answer

Section 2.6 Derivatives of Inverse Functions

Exercises

Answer

Answer

Answer

Answer

Section 2.7 Derivatives of Functions Given Implicitly

Exercises

Answer

Answer

Answer

Section 2.8 Using Derivatives to Evaluate Limits

Exercises

Answer

Answer

Answer

Answer

Chapter 3 Using Derivatives

Section 3.1 Using derivatives to identify extreme values

Exercises

Answer

Answer

Answer

Answer

Section 3.2 Using derivatives to describe families of functions

Exercises

Answer

Answer

Answer

Section 3.3 Global Optimization

Exercises

Answer

Answer

Answer

Answer

Section 3.4 Applied Optimization

Exercises

Answer

Answer

Answer

Answer

Section 3.5 Related Rates

Exercises

Answer

Answer

Answer

Answer

Chapter 4 The Definite Integral

Section 4.1 Determining distance traveled from velocity

Exercises

Answer

Answer

Answer

Answer

Section 4.2 Riemann Sums

Exercises

Answer

Answer

Answer

Answer

Section 4.3 The Definite Integral

Exercises

Answer

Answer

Answer

Answer

Section 4.4 The Fundamental Theorem of Calculus

Exercises

Answer

Answer

Answer

Answer

Answer

Chapter 5 Evaluating Integrals

Section 5.1 Constructing Accurate Graphs of Antiderivatives

Exercises

Answer

Answer

Answer

Section 5.2 The Second Fundamental Theorem of Calculus

Exercises

Answer

Answer

Answer

Section 5.3 Integration by Substitution

Exercises

Answer

Answer

Answer

Answer

Section 5.4 Integration by Parts

Exercises

Answer

Answer

Answer

Section 5.5 Other Options for Finding Algebraic Antiderivatives

Exercises

Answer

Answer

Answer

Section 5.6 Numerical Integration

Exercises

Answer

Answer

Answer

Chapter 6 Using Definite Integrals

Section 6.1 Using Definite Integrals to Find Area and Length

Exercises

Answer

Answer

Answer

Section 6.2 Using Definite Integrals to Find Volume

Exercises

Answer

Answer

Answer

Section 6.3 Density, Mass, and Center of Mass

Exercises

Answer

Answer

Answer

Section 6.4 Physics Applications: Work, Force, and Pressure

Exercises

Answer

Answer

Section 6.5 Improper Integrals

Exercises

Answer

Answer

Chapter 7 Differential Equations

Section 7.1 An Introduction to Differential Equations

Exercises

Answer

Answer

Answer

Section 7.2 Qualitative behavior of solutions to DEs

Exercises

Answer

Answer

Answer

Answer

Section 7.3 Euler's method

Exercises

Answer

Answer

Answer

Section 7.4 Separable differential equations

Exercises

Answer

Answer

Answer

Answer

Section 7.5 Modeling with differential equations

Exercises

Answer

Answer

Answer

Answer

Section 7.6 Population Growth and the Logistic Equation

Exercises

Answer

Answer

Answer

Chapter 8 Sequences and Series

Section 8.1 Sequences

Exercises

Answer

Answer

Answer

Answer

Section 8.2 Geometric Series

Exercises

Answer

Answer

Answer

Answer

Answer

Section 8.3 Series of Real Numbers

Exercises

Answer

Answer

Answer

Answer

Section 8.4 Alternating Series

Exercises

Answer

Answer

Answer

Section 8.5 Taylor Polynomials and Taylor Series

Exercises

Answer

Answer

Answer

Section 8.6 Power Series

Exercises

Answer

Answer

Answer

This page titled 12.3: C- Answers to Selected Exercises is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin, and Steven Schlicker (ScholarWorks @Grand Valley State University).

- 12.3: C- Answers to Selected Exercises by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0.

INDEX

A

autonumheader:yes2

1: Understanding the Derivative
10: Derivatives of Multivariable Functions
11: Multiple Integrals
2: Computing Derivatives
3: Using Derivatives
4: The Definite Integral
5: Finding Antiderivatives and Evaluating Integrals
6: Using Definite Integrals
7: Differential Equations
8: Sequences and Series
9: Multivariable and Vector Functions

This page titled Index is shared under a license and was authored, remixed, and/or curated by LibreTexts.

GLOSSARY

S

Sample Word 1 | Sample Definition 1

This page titled Glossary is shared under a license and was authored, remixed, and/or curated by LibreTexts.

DETAILED LICENSING

Overview

Title: 5: Active Calculus (Boelkins et al.)

Webpages: 106

All licenses found:

- CC BY-SA 4.0: 91.5% (97 pages)
- Undeclared: 8.5% (9 pages)

By Page

- 5: Active Calculus (Boelkins et al.) - *CC BY-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Understanding the Derivative - *CC BY-SA 4.0*
    - 1.1: How do we Measure Velocity? - *CC BY-SA 4.0*
    - 1.2: The Notion of Limit - *CC BY-SA 4.0*
    - 1.3: The Derivative of a Function at a Point - *CC BY-SA 4.0*
    - 1.4: The Derivative Function - *CC BY-SA 4.0*
    - 1.5: Interpretating, Estimating, and Using the Derivative - *CC BY-SA 4.0*
    - 1.6: The Second Derivative - *CC BY-SA 4.0*
    - 1.7: Limits, Continuity, and Differentiability - *CC BY-SA 4.0*
    - 1.8: The Tangent Line Approximation - *CC BY-SA 4.0*
    - 1.9: Understanding the Derivative (Exercises) - *CC BY-SA 4.0*
  - 2: Computing Derivatives - *CC BY-SA 4.0*
    - 2.1: Elementary Derivative Rules - *CC BY-SA 4.0*
    - 2.2: The Sine and Cosine Function - *CC BY-SA 4.0*
    - 2.3: The Product and Quotient Rules - *CC BY-SA 4.0*
    - 2.4: Derivatives of Other Trigonometric Functions - *CC BY-SA 4.0*
    - 2.5: The Chain Rule - *CC BY-SA 4.0*
    - 2.6: Derivatives of Inverse Functions - *CC BY-SA 4.0*
    - 2.7: Derivatives of Functions Given Implicitly - *CC BY-SA 4.0*
    - 2.8: Using Derivatives to Evaluate Limits - *CC BY-SA 4.0*
    - 2.9: Computing Derivatives (Exercises) - *CC BY-SA 4.0*
  - 3: Using Derivatives - *CC BY-SA 4.0*
    - 3.1: Using Derivatives to Identify Extreme Values - *CC BY-SA 4.0*
    - 3.2: Using Derivatives to Describe Families of Functions - *CC BY-SA 4.0*
    - 3.3: Global Optimization - *CC BY-SA 4.0*
    - 3.4: Applied Optimization - *CC BY-SA 4.0*
    - 3.5: Related Rates - *CC BY-SA 4.0*
    - 3.6: Using Derivatives (Exercises) - *CC BY-SA 4.0*
  - 4: The Definite Integral - *CC BY-SA 4.0*
    - 4.1: Determining Distance Traveled from Velocity - *CC BY-SA 4.0*
    - 4.2: Riemann Sums - *CC BY-SA 4.0*
    - 4.3: The Definite Integral - *CC BY-SA 4.0*
    - 4.4: The Fundamental Theorem of Calculus - *CC BY-SA 4.0*
    - 4.5: The Definite Integral (Exercises) - *CC BY-SA 4.0*
  - 5: Finding Antiderivatives and Evaluating Integrals - *CC BY-SA 4.0*
    - 5.1: Construction Accurate Graphs of Antiderivatives - *CC BY-SA 4.0*
    - 5.2: The Second Fundamental Theorem of Calculus - *CC BY-SA 4.0*
    - 5.3: Integration by Substitution - *CC BY-SA 4.0*
    - 5.4: Integration by Parts - *CC BY-SA 4.0*
    - 5.5: Other Options for Finding Algebraic Derivatives - *CC BY-SA 4.0*
    - 5.6: Numerical Integration - *CC BY-SA 4.0*
    - 5.7: Finding Antiderivatives and Evaluating Integrals (Exercises) - *CC BY-SA 4.0*
  - 6: Using Definite Integrals - *CC BY-SA 4.0*
    - 6.1: Using Definite Integrals to Find Area and Length - *CC BY-SA 4.0*
    - 6.2: Using Definite Integrals to Find Volume - *CC BY-SA 4.0*
    - 6.3: Density, Mass, and Center of Mass - *CC BY-SA 4.0*
    - 6.4: Physics Applications - Work, Force, and Pressure - *CC BY-SA 4.0*

- 6.5: Improper Integrals - *CC BY-SA 4.0*
- 6.6: Using Definite Integrals (Exercises) - *CC BY-SA 4.0*

◦ 7: Differential Equations - *CC BY-SA 4.0*
- 7.1: An Introduction to Differential Equations - *CC BY-SA 4.0*
- 7.2: Qualitative Behavior of Solutions to Differential Equations - *CC BY-SA 4.0*
- 7.3: Euler's Method - *CC BY-SA 4.0*
- 7.4: Separable Differential Equations - *CC BY-SA 4.0*
- 7.5: Modeling with Differential Equations - *CC BY-SA 4.0*
- 7.6: Population Growth and the Logistic Equation - *CC BY-SA 4.0*
- 7.7: Differential Equations (Exercises) - *CC BY-SA 4.0*

◦ 8: Sequences and Series - *CC BY-SA 4.0*
- 8.1: Sequences - *CC BY-SA 4.0*
- 8.2: Geometric Series - *CC BY-SA 4.0*
- 8.3: Series of Real Numbers - *CC BY-SA 4.0*
- 8.4: Alternating Series - *CC BY-SA 4.0*
- 8.5: Taylor Polynomials and Taylor Series - *CC BY-SA 4.0*
- 8.6: Power Series - *CC BY-SA 4.0*
- 8.7: Sequences and Series (Exercises) - *CC BY-SA 4.0*

◦ 9: Multivariable and Vector Functions - *CC BY-SA 4.0*
- 9.1: Functions of Several Variables and Three Dimensional Space - *CC BY-SA 4.0*
- 9.2: Vectors - *CC BY-SA 4.0*
- 9.3: Dot Product - *CC BY-SA 4.0*
- 9.4: The Cross Product - *CC BY-SA 4.0*
- 9.5: Lines and Planes in Space - *CC BY-SA 4.0*
- 9.6: Vector-Valued Functions - *CC BY-SA 4.0*
- 9.7: Derivatives and Integrals of Vector-Valued Functions - *CC BY-SA 4.0*
- 9.8: Arc Length and Curvature - *CC BY-SA 4.0*

◦ 10: Derivatives of Multivariable Functions - *CC BY-SA 4.0*
- 10.1: Limits - *CC BY-SA 4.0*
- 10.2: First-Order Partial Derivatives - *CC BY-SA 4.0*
- 10.3: Second-Order Partial Derivatives - *CC BY-SA 4.0*
- 10.4: Linearization- Tangent Planes and Differentials - *CC BY-SA 4.0*
- 10.5: The Chain Rule - *CC BY-SA 4.0*
- 10.6: Directional Derivatives and the Gradient - *CC BY-SA 4.0*
- 10.7: Optimization - *CC BY-SA 4.0*
- 10.8: Constrained Optimization - Lagrange Multipliers - *CC BY-SA 4.0*

◦ 11: Multiple Integrals - *CC BY-SA 4.0*
- 11.1: Double Riemann Sums and Double Integrals over Rectangles - *CC BY-SA 4.0*
- 11.2: Iterated Integrals - *CC BY-SA 4.0*
- 11.3: Double Integrals over General Regions - *CC BY-SA 4.0*
- 11.4: Applications of Double Integrals - *CC BY-SA 4.0*
- 11.5: Double Integrals in Polar Coordinates - *CC BY-SA 4.0*
- 11.6: Surfaces Defined Parametrically and Surface Area - *CC BY-SA 4.0*
- 11.7: Triple Integrals - *CC BY-SA 4.0*
- 11.8: Triple Integrals in Cylindrical and Spherical Coordinates - *CC BY-SA 4.0*
- 11.9: Change of Variables - *CC BY-SA 4.0*

◦ 12: Appendices - *CC BY-SA 4.0*
- 12.1: A- Short Table of Integrals - *CC BY-SA 4.0*
- 12.2: B- Answers to Activities - *CC BY-SA 4.0*
- 12.3: C- Answers to Selected Exercises - *CC BY-SA 4.0*

◦ Back Matter - *Undeclared*
- Index - *Undeclared*
- Glossary - *Undeclared*
- Detailed Licensing - *Undeclared*

This page titled Active Calculus (Boelkins et al.) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew Boelkins, David Austin & Steven Schlicker via source content that was edited to the style and standards of the LibreTexts platform.

- Active Calculus (Boelkins et al.) by Matthew Boelkins, David Austin, and Steven Schlicker is licensed CC BY-SA 4.0. Original source: https://activecalculus.org/single.