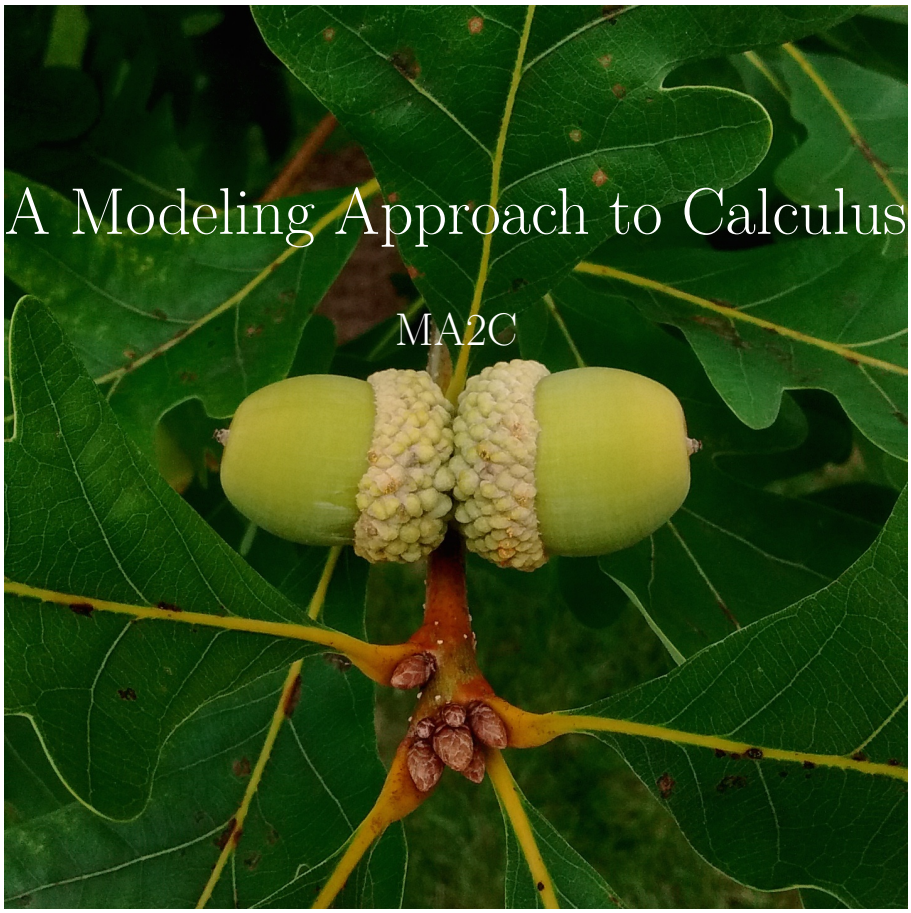
A Modeling Approach to Calculus

MA2C



A Modeling Approach to Calculus

MA2C

D. Brian Walton

James Madison University
Harrisonburg, Virginia, USA

October 15, 2019

Preface

Calculus is a branch of mathematics that studies functions through the processes of limits, derivatives, and integrals. This text introduces the ideas of calculus through the context of mathematical models. This text is organized to review precalculus concepts necessary to be successful in learning calculus. The goal is not, however, to introduce these concepts to a beginner but to leverage some familiarity toward a deeper understanding.

The text begins with a basic review of numbers and variables with an aim of connecting these foundational concepts to the goals of modeling relationships between physically measured quantities. It then turns to the development of functions as the mathematical tool used for predictive relationships between variables. Intuitive ideas of continuity are introduced in relationship to making piecewise functions connected, with limits introduced as a way to characterize the behavior of a function at the break points.

Sequences are introduced early as a discrete illustration of many of the concepts relating to modeling with functions. Patterns found in sequences are familiar to students and these patterns can often be described both explicitly and recursively. This will serve as a prelude to the idea that functions can be defined explicitly with a formula or indirectly through differential equations. In addition, sequences foreshadow the ideas of limits, derivatives, and integrals. Describing the monotonicity and concavity of a sequence then provides a direct correspondence to describing the behavior of a function in terms of its first and second derivatives.

Summation rules and formulas naturally occur as part of the discussion of sequences. This motivates an early introduction of the definite integral as the generalization of the accumulation of increments of change. The rate of change is introduced in the context of accumulation and the reader is explicitly told to look forward to the Fundamental Theorem of Calculus as the formal connection between the rate of accumulation and the instantaneous rate of change as being equivalent. Properties of the definite integral as well as elementary formulas are introduced. The behavior of functions in terms of the first and second derivative are introduced using the integral representation of functions.

Next, we develop a more thorough investigation of limits and continuity of functions. Properties of limits are developed in the context of the limits of sequences. The epsilon-delta definition of a limit is given in an optional section. Continuity of functions is formalized and the major theorems for continuous functions are presented, namely the Intermediate Value Theorem and the Extreme Value Theorem.

This is followed by the formal development of the derivative and the rules of differentiation, including the first part of the Fundamental Theorem of Calculus. Applications of derivatives and antiderivatives are included.

You can download a PDF copy of the book with the links below. There are

differently formatted options, depending on how you intend to use the PDF.

- http://educ.jmu.edu/~waltondb/MA2C/model-calculus.pdf (Margins formatted to print 2-sided)
- http://educ.jmu.edu/~waltondb/MA2C/model-calculus-1page.pdf (Margins formatted to print 1-sided)

In order to keep track of changes during the semester, the following list describes the changes that have taken place in the text since August 2018.

- n/a

Contents

Part I

Chapter 1

Foundational Principles

1.1 Learning Mathematics

I assume that you want to be successful in learning mathematics. It might be true, however, that your past experiences make you doubt this as a possibility. Before we start talking about actual mathematics, let's have a discussion about what it means to learn mathematics and how to approach this more successfully.

1.1.1 How People Learn

This section is based on my reading, understanding and interpretation of a report issued by the National Research Council in 2000 called *How People Learn: Brain, Mind, Experience and School* and a follow-up report in 2005 focused more directly on mathematics, *How Students Learn: Mathematics in the Classroom*. Although there is a lot to digest, I would encourage you to read some of this if you want more background. I hope to convey some key points that I hope will be especially helpful in focusing our efforts in learning.

Let us start with some questions. What does it mean to learn? Can we become more effective at learning?

Studies of brain activity reveal that learning physically changes the structure of the brain. The brain consist of neurons (brain cells) that form networks. Any particular neuron has dendrites which receive signals from other neurons. When the total signal received by a neuron reaches a critical threshold, the neuron will produce its own signal which it transmits along its axon to synaptic connections with the dendrites of other neurons. External stimuli trigger patterns of neuron firing responses which are ultimately translated into memories and actions. Learning consists of transforming these networks so that the firing patterns change. Feedback (both positive and negative) is necessary to weaken connections that are not desired and to create and strengthen connections that are desired.

Early attempts to measure learning (which pre-date the understanding of the brain), especially when studying animal models like rats in maze, considered learning as behavior that is reinforced by a stimulus. An example is given of a cat trapped in box with an exit that opens when a particular string is pulled. By (frantic) trial and error, the cat eventually opens the box. When put back in the box, the cat will not immediately repeat the necessary action to escape; it takes a number of repetitions for the action to be reinforced by the reward before the cat learns to attempt the particular action immediately in order to escape. Using modern understanding of the brain's neural networks, the negative feedback (from being trapped) and the positive feedback (from being released and rewarded) established new neural connections that triggered pulling the release when the cat found itself in the trapped environment.

Memorization of facts using flashcards or online drills can be interpreted in the context of this model of learning. A presented statement (the clue or question) provides a stimulus. Recalling the desired response (the answer) is the behavior desired from that response. Success or failure and the resulting emotional responses during training provide the feedback for the new network connections to be formed. Unfortunately, memorization of facts is not an effective learning approach when dealing with more abstract problem solving scenarios.

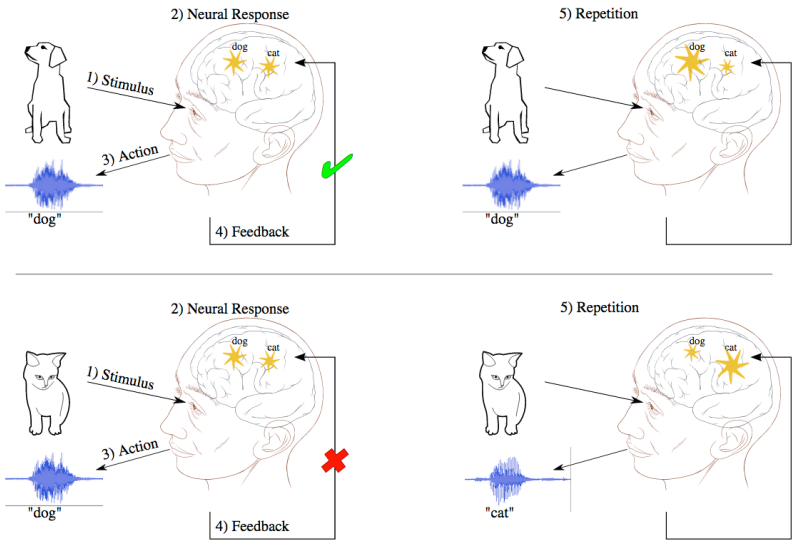


Figure 1.1.1 A model for learning as feedback-driven rewiring of neural pathways.

Views of learning have expanded to consider learning with understanding wherein learning is not just a reinforced behavior but a cognitive effort (thought) built on a framework (understanding). Learning therefore is not just a filing system of individual memories; it is structured according to some organizational process that itself involves neural feedback. One of the key distinctions that has been identified between novices and experts in a problem-solving area is that an expert has a rich body of knowledge that is organized around core concepts as opposed to a list of facts or strategies.

Thus, we see that learning requires more than just the memorization of facts, processes and strategies; it requires developing a mental framework by which information is organized. These steps require that we undertake processes by which neural network connections in our brain are rewired, breaking connections that correspond to misconceptions and poor organization and developing new connections that correspond to more effective knowledge and organization. This requires a significant level of interaction with the material that we wish to learn in order to effect such a change.

1.1.2 Best Practices for Learning

How People Learn identified three basic principles to guide effective learning.

1. Because students approach learning with preconceptions, these understandings must be engaged or else learning will be superficial or thwarted.

2. Developing competence requires (a) a deep foundation of factual knowledge, (b) a conceptual framework by which these facts and ideas are understood, and (c) an organized memory system that facilitates retrieval and application of that knowledge.

3. Students should take control of their own learning by defining learning goals and monitoring their progress in achieving them.

We should attempt to frame our learning experiences so that these three principles are implemented.

One key aspect of this framing has to do with our **mindset**. Mindset refers to our own ideas about how we learn. We have a fixed mindset when we believe that our intelligence is predetermined and our ability to learn is limited. We have a growth mindset when we believe that our intelligence can grow and we can learn anything with enough effort.

Jo Boaler, a mathematics education expert on mindset, has a fascinating project on mathematical mindset and mathematical learning, youcubed.org, in which she emphasizes the following three key points about our attitude and approaches to learning.

1. Anyone can learn to high levels.
2. Mistakes and struggle are good for brain growth.
3. Visualization of mathematics helps develop our brain connections.

I highly recommend watching some of the videos that she has developed.

1.1.2.1 Engaging Preconceptions

Preconceptions represent the sum of all knowledge accumulated as well as the manner in which that knowledge is organized and interpreted. Valid preconceptions provide the necessary foundation on which additional knowledge accumulates. Unfortunately, it is often the case that preconceptions might also be an obstacle for learning. This may be due to something learned incorrectly. But it can just as often be something learned correctly but organized in a way that obscures generalizations necessary to advance in learning.

Engaging our preconceptions involves recognizing exactly what our preconceptions may be. For valid preconceptions, we integrate new knowledge into our system of understanding and it is held more tightly than if we did not connect it to our existing knowledge. For invalid or obstructive preconceptions, we face an uncomfortable **cognitive dissonance** that may require dismantling and reconstructing our framework of understanding so that future learning can proceed.

1.1.2.2 Developing and Organizing a Deep Foundation of Knowledge

The second principle focuses on the knowledge itself and how we organize our thinking about that knowledge. We start with the need for a deep foundation of facts. However, we should notice that the factual knowledge alone is really just one component of this learning principle. At first glance, I thought the second and third components—a conceptual framework and an organized memory system—were the same thing. Then I saw that although they could be related, they emphasize two different aspects.

The conceptual framework is about how the facts and ideas are understood. This is how we make sense of the facts, how we relate them with one another, and how we interpret them. Effective learning requires developing this framework as we add knowledge to our memory.

The organized memory system refers to our methods and strategies for recall. Consider how we can organize files on a computer drive. We could adopt a flat filing system where every file is in a single location, distinguished only by name; or we could adopt a hierarchical filing system with folders or directories organized in a way that related files are grouped together. More modern storage strategies include tagging files (e.g., hashtags), which may be easier to relate to memory. Imagine that our memories work in a similar way,

that we can establish *tags* that go along with the knowledge. By thinking about the relevant tags as a stimulus for our mind, we can trigger our memory to recall that desired knowledge.

1.1.2.3 Metacognition and Taking Control of Your Learning

The third principle states that students should take control of their own learning. Each student should establish clear learning goals and monitor their progress. This requires thinking about their thinking, reflecting on their own understanding and effectiveness in organizing their knowledge. This process is called **metacognition**. Metacognition allows us assess our own progress and identify our strengths and weaknesses. It allows us to recognize when we need additional help. It is essential to make these assessments while our knowledge and skill set are still forming. When we recognize our weaknesses, we can adapt our learning methods and accelerate our progress. We shouldn't wait for a class exam to decide we don't understand.

1.1.3 What's Class Got to Do With It?

As the understanding of how we learn has grown, experts recommend that our educational settings and environments be designed to facilitate effective learning. Four design characteristics summarize what we want to create effective learning. The environment should be (1) learner-centered, (2) knowledge-centered, (3) assessment-centered, and (4) community-centered.

In a learner-centered classroom, the educational experiences give attention to students' ideas, knowledge, skills, and attitudes. There is an awareness that existing ideas can lead to misconceptions as well as a path to new understanding. Student experiences as well as the ways students reflect and understand these experiences will be different for different individuals. Consequently, students must be individually engaged in the educational experience. This is strongly related to the idea that brain growth occurs during the struggle of learning new things.

In a knowledge-centered classroom, the educational experiences provide clear guidance on what is intended for learning, and those experiences are designed to develop understanding of that knowledge. Effective knowledge involves both the content matter and an understanding of the context, relationship, and application of that matter. Knowledge-centered learning helps students create an effective mental organization by identifying core connected ideas.

Assessment-centered education emphasizes that both the learner and the educator need to assess the progress of learning and understanding. Assessment helps make learning and understanding visible to both the student and the teacher during the learning process and not just through formal evaluation. Instruction that helps students become aware of their own progress through informal assessments provides them with opportunities to revise and improve their thinking. In-class activities where students engage in material and evaluate their own progress are examples of such informal assessment opportunities. In response to these assessment opportunities, students can develop their metacognitive abilities and learn to evaluate the effectiveness of their strategies in learning.

A community-centered classroom establishes norms of behavior and connections to the world that support core learning values. In our classes, we encourage the processes of development, questioning, and progress. Because our minds grow more when we make mistakes, we welcome mistakes. They

do not measure inadequacy but are evidence of a healthy struggle to learn. We also encourage taking emotional risks that are part of asking questions or suggesting alternative approaches. At the same time, the classroom never has room for comments or behaviors that degrade, belittle, or hurt others. In summary, a community-centered classroom seeks to establish an environment that encourages a growth mindset.

1.2 Numbers, Measurements and Relations

1.2.1 Overview

Numbers play a fundamental role in science because they allow us to quantify observations. That is, instead of saying things are big or small, we can assign numbers to measurements. Science in large part has progressed because of our ability to determine mathematical relationships between different measurements that make prediction possible.

Mathematics admittedly views numbers themselves as objects worthy of study. A number is an exact entity—other numbers, regardless of how close they might be, are different. Consequently, a mathematical result has an exact value. Sometimes we can only approximate the value, but we should give an exact expression for that value when possible. Mathematicians classify numbers according to the complexity of their definition. Historical conventions often suggest ways in which we can simplify an expression, often so that the classification can be more easily recognized.

Measurements allow us to assign numerical values to physical attributes, such as length, temperature, mass, and speed. Instruments need to be designed so that repeated measurements will result in the same values using standard units. Measurement error results from the limited accuracy intrinsic in reading a measurement from an instrument's scale. Most often, we measure more than one attribute of an object or system under a given condition. The collection of all such measurements is called the **state** of the system. We generally wish to understand the relationships between these different quantities.

1.2.2 Numbers in Mathematics

In mathematics, numbers have precise meanings and classifications. Here, we review the basic **sets** of numbers. The **natural numbers** are the positive integers

$$\mathbb{N} = \{1, 2, 3, \ldots\}.$$

Including the number zero gives us all **counting numbers**

$$\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}.$$

The set of all **integers** is written

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}.$$

The **rational numbers** are all numbers that can be represented as a ratio of integers

$$\mathbb{Q} = \{\frac{p}{q} : p \in \mathbb{Z}, q \in \mathbb{N}\}.$$

This **set-builder notation** indicates that $\mathbb{Q}$ is a set of numbers that can be written in the form p/q where p is some integer and q is some natural number.

We often visualize numbers geometrically using a number line. First, the origin of the line is specified with a value of zero. The integers are then equally spaced by a unit length counting from zero. (See Figure A.1.1.) Subdividing the unit length into a whole number of equal parts generates additional points that are rational numbers. (See Figure A.1.2.) However, even when all rational numbers are included, there are infinitely many points on the line that are never covered. These are the **irrational numbers**, which include algebraic numbers like $\sqrt{2}$ or $\sqrt{3}$, as well as transcendental numbers like π and e. The set of **real numbers** is written $\mathbb{R}$ and consists of both rational and irrational numbers.

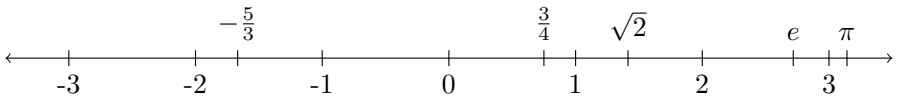


Figure 1.2.1 The number line graphically represents real numbers, both rational and irrational.

Every mathematical value represents a single point on the number line. Two values are equal only when they refer to the very same point on the line. Calculators give decimal approximations for numbers using a limited number of digits and so they can actually only represent a finite collection of the infinitely many possible numbers. In particular, irrational numbers can never be represented exactly using decimals. Thus, we usually represent mathematical values by their mathematical expression rather than decimals. When a decimal approximation is useful, we should indicate we are making an approximation using the approximation symbol ($\approx$) rather than an equals sign ($=$).

Simplification of numbers corresponds to finding a new representation of a number in a reduced form. For example, a rational number has many different representations of the form p/q with $p \in \mathbb{Z}$ and $q \in \mathbb{N}$. But there is only one representation where p and q have no common factors. Canceling any common factors to find this representation would be simplification. Other examples of simplification include simplifying a root or rationalizing a denominator.

Example 1.2.2 The fraction $\frac{126}{24}$ is not simplified. If we find the prime factorization of the numerator and denominator, we find

$$126 = 6 \cdot 21 = 2 \cdot 3^2 \cdot 7,$$
$$24 = 3 \cdot 8 = 2^3 \cdot 3.$$

The fraction simplifies by canceling all common factors:

$$\frac{126}{24} = \frac{2 \cdot 3^2 \cdot 7}{2^3 \cdot 3} = \frac{3 \cdot 7}{2^2} = \frac{21}{4}.$$

In practice, we don't always have to find the prime factorization. Instead, we can find one common factor at a time until no common factors remain. For example, since 126 and 24 are both even, we could write

$$\frac{126}{24} = \frac{63}{12}.$$

Then, we look at 63 and 12 and recognize that they are both divisible by 3, allowing us to rewrite the fraction as

$$\frac{126}{24} = \frac{63}{12} = \frac{21}{4}.$$

Because $21 = 3 \cdot 7$ and $4 = 2^2$ do not have common factors, we know this is simplified. □

A square root is not simplified if there is a factor inside the root that is a perfect square. Similarly, a cube root is not simplified if there is a factor inside that is a perfect cube. We use the factors of the value inside a root to determine if we can simplify it.

Example 1.2.3 The square root $\sqrt{126}$ is not simplified. A square root is the inverse operation of squaring numbers (for non-negative numbers) so that

$\sqrt{3^2} = 3$. Because $\sqrt{a \cdot b} = \sqrt{a} \cdot \sqrt{b}$ (for $a, b \geq 0$), we can simplify as

$$\sqrt{126} = \sqrt{2 \cdot 3^2 \cdot 7} = \sqrt{9} \cdot \sqrt{14} = 3\sqrt{14}.$$

□

Example 1.2.4 The cube root $\sqrt[3]{48}$ is not simplified. We start by factoring:

$$48 = 2 \cdot 24 = 2 \cdot 3 \cdot 8 = 2^4 \cdot 3.$$

A cube root is the inverse operation of cubing numbers so that for every perfect cube, we can simplify $\sqrt[3]{a^3} = a$. We have

$$\sqrt[3]{48} = \sqrt[3]{2^4 \cdot 3} = \sqrt[3]{2^3 \cdot 2 \cdot 3} = 2\sqrt[3]{6}.$$

□

Some additional rules of simplification you may have learned were created before fast calculators and computers were available. These rules were taught so that scientists and engineers could express an answer that would be in a form where it would be faster to use the tables and slide rules available at the time. We no longer need such rules for efficiency, but they often illustrate important algebra rules.

One example of such a rule is the simplification of fractions with square roots, called rationalizing a fraction. It was much more costly to use a table or slide rule if the root was in the denominator. The practice was developed to rewrite such an answer so that the root was in the numerator. This could be accomplished by multiplying the fraction on top and bottom by a factor that would eliminate the undesired root.

Example 1.2.5 Simplify $\dfrac{4}{3\sqrt{2}}$ by rationalizing the denominator.

Solution. A square root can simplify if there is a perfect square inside. The square root in this denominator $\sqrt{2}$ would need another 2 inside to have a perfect square. Multiply numerator and denominator by the extra $\sqrt{2}$ to get a square in the denominator.

$$\frac{4}{3\sqrt{2}} = \frac{4\sqrt{2}}{3\sqrt{2}\sqrt{2}} = \frac{4\sqrt{2}}{3 \cdot 2}$$

Now we can finish simplifying the fraction by canceling common factors:

$$\frac{4}{3\sqrt{2}} = \frac{4\sqrt{2}}{6} = \frac{2\sqrt{2}}{3}.$$

□

In the previous example, the two expressions $\dfrac{4}{3\sqrt{2}}$ and $\dfrac{2\sqrt{2}}{3}$ are equally simplified. The first expression has a rationalized numerator. The second expression has a rationalized denominator. You should ask your instructor whether they expect a preferred simplified form.

Although we usually use simplification for aesthetic reasons, having a standard way to write numbers can be useful to prove mathematical results. Thanks to Pythagorus, the ancient Greeks knew that $\sqrt{2}$ was a number that represented the hypotenuse of an isosceles right triangle with legs of unit length. The Greeks also originally thought that all numbers would ultimately be rational numbers. Realizing that $\sqrt{2}$ was irrational was so shocking that, according to legend, the discoverer of this fact was drowned at sea.

Example 1.2.6 The proof of the irrationality of $\sqrt{2}$ uses the idea that rational numbers have a simplified form. The basic argument is to consider a rational number that *might* represent $\sqrt{2}$ and then proceed to show that such a representation doesn't make sense. The detailed argument is shown below.

Solution. Suppose that $\sqrt{2}$ is a rational number. Then it can be written as the ratio of two integers $\sqrt{2} = \frac{p}{q}$ in reduced form, meaning p and q do not have common factors. By definition of square roots, we must have $\frac{p^2}{q^2} = 2$ which implies

$$p^2 = 2q^2$$

so that p^2 is an even number. The only way that p^2 can be even is if p itself is even, since the product of two odd numbers is always odd.

Once we know p is even, we can factor out 2 and write $p = 2k$ where k is also an integer. Now $p^2 = 4k^2$ which implies $4k^2 = 2q^2$ or

$$q^2 = 2k^2.$$

This means q would also be an even number. This is where the contradiction occurs—since p and q were to have had no common factors, they couldn't both be even. This means that $\sqrt{2}$ can not be written as a reduced fraction, which in turn means that $\sqrt{2}$ is not a rational number. □

1.2.3 Numbers as Measurements

In science, numbers often arise from measurements. When counting objects, measurements use integers and are exact. Most measurements, however, are not exact and require the use of a scale. An instrument for measurement provides a physical tool that allows us to identify a number of units associated with the physical quantity.

The most elementary physical measurement of this type is a measurement of length. The instrument of measurement, a ruler, uses a constructed number line such that the spacing between numbers on the ruler represents distance. The standard unit for the ruler, such as an inch or centimeter, sets the spacing between integer distances. The designer of a ruler also chooses the number of subdivisions per unit. For example, many rulers with inches use either 8 or 16 subdivisions per inch while metric rulers use 10 subdivisions per centimeter. You should see a similarity between the construction of the ruler and the development of rational numbers, except that the set of rational numbers allows for all possible integer number of subdivisions of the unit.

Measured quantities generally occur between two values that an instrument can measure. Given a ruler, the observer must choose a length based on the existing rulings. Even a length that appears to be exactly on a ruling might be found to be slightly off when examined under magnification. The observer might also make judgment errors in reading off a measurement. Consequently, a measurement represents an approximation of the value of a quantity. The difference between the true and measured values of a quantity is called an **error**.

Definition 1.2.7 Given any quantity with an actual value Q and a measured value $\widehat{Q}$, the **error** or **residual** is a value, say E, that measures the difference between the actual and measured values,

$$E = Q - \widehat{Q}.$$

Equivalently, E is that quantity such that the actual value is equal to the

measured value plus the error,

$$Q = \widehat{Q} + E.$$

$\Diamond$

Definition 1.2.8 Given any quantity with an actual value Q and a measured value $\widehat{Q}$, the **absolute error** measures the absolute value of the error:

$$|E| = |Q - \widehat{Q}|.$$

$\Diamond$

Note 1.2.9 Symbols that represent variables correspond to the entire symbol being used. In the previous definitions, Q and $\widehat{Q}$ are different symbols and represent different numbers even though they both use the letter "Q". Similarly, the case of a letter matters so that R and r are different symbols. Avoid the trap of thinking that the letter's name is the symbol.

In science, the error in a measurement is not known precisely but is represented by a *bound*. There are a variety of techniques used to indicate the bound for an error. We will briefly discuss how the use of significant digits or a margin of error represent error bounds.

One way that the accuracy of measurements are described is using a number of **significant digits**. The idea is that the last digit reported represents the smallest subdivision the instrument can distinguish.

Example 1.2.10 Imagine that an object has a length of 15.2772 cm, measured to the nearest micron (micrometer). (We never really know exact lengths of physical objects.) How would the length be reported with different numbers of significant digits?

If that object was measured using a ruler showing only centimeters, we would see that the length was between 15 and 16 cm but closer to 15. Our measurement would be written as 15 cm, or $\widehat{Q} = 15$, and we would have two significant digits. However, if we did not know about details of the original measurement and only saw the recorded value of 15 cm, then we would have to assume that the true length was somewhere between 14.5 cm and 15.5 cm,

$$14.5 \leq Q \leq 15.5$$

Subtracting $\widehat{Q} = 15$ from each term, we find

$$-0.5 \leq Q - \widehat{Q} \leq 0.5.$$

The absolute error is therefore *bounded* by 0.5 cm.

If the ruler showed millimeters, then our measurement would be 15.3 cm with three significant digits. Knowing only the measurement $\widehat{Q} = 15.3$ and that there are three significant digits, we can infer

$$15.25 \leq Q \leq 15.35$$

so that the error is bounded between

$$-0.05 \leq Q - \widehat{Q} \leq 0.05.$$

The absolute error based on the measurement is bounded by 0.05 cm. $\square$

An alternative to using significant digits is to state explicitly a **margin of error**. The margin of error is equivalent to providing a bound for the error. We saw that a measurement 15.3 cm with three significant digits corresponds

to an inequality

$$15.25 \leq Q \leq 15.35$$

such that the absolute error is bounded by 0.05,

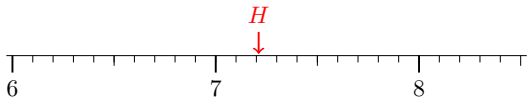
$$|Q - \widehat{Q}| \leq 0.05.$$

Using a margin of error, we write $Q = 15.3 \pm 0.05$ cm. The margin of error ± 0.05 is interpreted as $-0.05 \leq Q - \widehat{Q} \leq 0.05$.

A margin of error is more precise than significant digits. For example, if we wanted to say that a measurement was somewhere between 15.2 cm and 15.4 cm, then we would write 15.3 ± 0.1 cm. The value 15.3 was used as a central value and the margin of error gives a distance in either direction to reach the extreme values. The true value must be between the extremes.

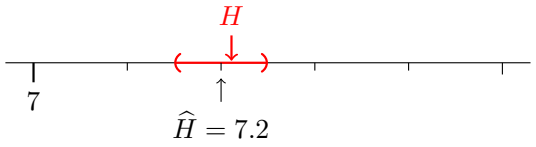
Example 1.2.11 The length of the hypotenuse of a right triangle with legs of lengths 4 cm and 6 cm is $H = \sqrt{4^2 + 6^2} = \sqrt{52} = 2\sqrt{13}$ cm. A calculator shows the decimal approximation is $H \approx 7.211103$ cm.

Now, suppose we use a ruler using centimeters but showing millimeters to measure the length. Different ways of describing the measurement with a margin of error give different information about the length.



- Write H using a margin of error to state that the measurement to the nearest millimeter is 7.2 cm.
- Write H using a margin of error to state that the measurement is between 7.2 and 7.3 cm.
- Write H using a margin of error to state that the measurement is between 7.2 and 7.25 cm.

Solution. First, we consider the nearest tick mark on the ruler. The measurement $\widehat{H} = 7.2$ cm will be the *nearest* value for any actual length satisfying $7.15 \leq H \leq 7.25$.



The spacing from $\widehat{H}$ and the edge of this interval is

$$\epsilon = |7.25 - 7.2| = |7.15 - 7.2| = 0.05.$$

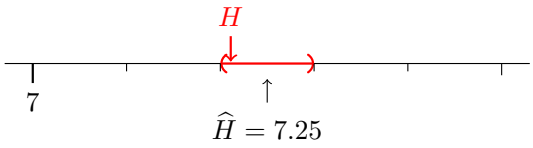
This value ϵ is the largest margin of error so that

$$|H - 7.2| \leq 0.05.$$

We write $H = 7.2 \pm 0.05$ cm.

Next, we work with the range $7.2 \leq H \leq 7.3$. We find the mid-point of this interval as our recorded measurement

$$\widehat{H} = \frac{7.2 + 7.3}{2} = 7.25.$$



Then we measure the distance from the center to the edge,

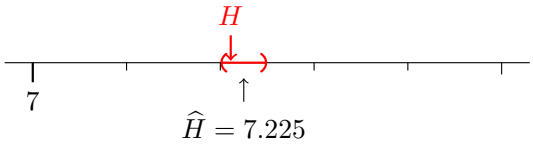
$$\epsilon = |7.3 - 7.25| = |7.2 - 7.25| = 0.05$$

to find the margin of error 0.05 cm. We can then express our measurement with a margin of error as $H = 7.25 \pm 0.05$ cm corresponding to a bounded error

$$|H - 7.25| \leq 0.05.$$

Finally, we repeat this process to indicate that H is in the range $7.2 \leq H \leq 7.25$. The mid-point gives

$$\widehat{H} = \frac{7.2 + 7.25}{2} = 7.225.$$



The margin of error is 0.025 so that our new approximate measurement with a margin of error is written $H = 7.25 \pm 0.025$ cm. Using an inequality involving absolute values, we could write

$$|H - 7.225| \leq 0.025.$$

□

In general, a margin of error establishes a symmetric interval of possible values around the measurement. If we symbolically represent the margin of error by $\epsilon > 0$ (the Greek letter epsilon), then the statement $Q = \widehat{Q} \pm \epsilon$ is a statement that the absolute error is bounded by ϵ,

$$|Q - \widehat{Q}| \leq \epsilon.$$

In other words, we know from the measurement that the true value Q is in the interval

$$\widehat{Q} - \epsilon \leq Q \leq \widehat{Q} + \epsilon.$$

1.2.4 Summary

- In mathematics, numbers represent specific points on the number line. Real numbers ($\mathbb{R}$) can be classified as natural numbers ($\mathbb{N}$), integers ($\mathbb{Z}$), rational numbers ($\mathbb{Q}$), and irrational numbers.
- To simplify an expression is to find an expression representing the same value in a simpler form. For fractions, there should be no common factors. For roots, the power of prime factors inside should be less than the root.
- In a physical context, numbers represent measurements that have limited precision. This precision might be characterized by significant digits or by a margin of error.

- The error of approximation E for a quantity Q and an approximation $\widehat{Q}$ is defined by
$$E = Q - \widehat{Q}.$$
A symmetrical error bound $-\epsilon \leq E \leq \epsilon$ corresponds to the absolute value inequality $|Q - \widehat{Q}| \leq \epsilon$ for a range of values
$$\widehat{Q} - \epsilon \leq Q \leq \widehat{Q} + \epsilon.$$

1.2.5 Exercises

Simplify the following values.

1. $\frac{42}{12}$

2. $\frac{210}{28}$

3. $\sqrt{75}$

4. $\sqrt{160}$

5. $\sqrt[3]{160}$

6. $\sqrt[4]{160}$

7. $\frac{\sqrt{72}}{4}$

8. $\frac{\sqrt{864}}{15}$

Simplify the following values by rationalizing the denominator.

9. $\dfrac{6}{5\sqrt{3}}$

10. $\dfrac{10}{\sqrt[3]{2}}$

11. $\dfrac{4\sqrt{2}}{\sqrt{3}}$

Simplify the following values by rationalizing the numerator.

12. $\dfrac{4\sqrt{2}}{\sqrt{3}}$

13. $\dfrac{5\sqrt[3]{4}}{\sqrt[3]{3}}$

14. One of your colleagues has recorded the mass of a specimen in your lab's notebook. The recording is given as 35.8 g. How much uncertainty is in this measurement? What are the possible actual masses that might correspond to that measurement?

15. One of your colleagues has recorded the mass of a specimen in your lab's notebook. The recording is given as 35.8 g. Suppose you also know that the lab instrument that was used always rounds measurements to the nearest 0.2 g. How should the recording have been written to indicate this additional information? What are the possible actual masses that might correspond to that measurement?

16. A thermometer's scale shows every 5 degrees. You observe the current temperature registers on the thermometer as being between 75 and 80 but clearly closer to 75 degrees. How would you report the temperature in order to reflect both your measurement and your uncertainty?

17. A right triangle is formed with legs measuring 5 cm and 8 cm. Express the length of the hypotenuse H to the nearest tenth of a centimeter, stating the margin of error based on a ruler showing millimeters. Rewrite your statement about margin of error as an inequality involving absolute values.

18. A right triangle is formed with one leg measuring 10 cm and the hypotenuse measuring 18 cm. Express the length of the other leg L to the nearest tenth of a centimeter, stating the margin of error based on a ruler showing millimeters. Rewrite your statement about margin of error as an inequality involving absolute values.

1.3 Graphs and Relations between Variables

1.3.1 Overview

In physical settings, we usually consider measurements of multiple quantities at the same time. We do this because we are interested in the relationships between these different quantities. We think of each quantity of interest as a **state variable**; the collection of all such variables under consideration is called the **system**. At any instant, the variables of the system will each have a particular value and the collection of those values at that instant is called the **state** of the system. Graphs are often used to reveal relationships between different variables.

In this section, we explore graphs of data through scatter plots and graphs of equations. An equation involving multiple variables represents a relation between those variables. We consider the role of solving equations in the context of these relations.

1.3.2 Systems, States and Variables

In the course of an experiment, or even just in observation, many different quantities typically are covarying, or changing with one another. For example, an object in motion has changing position, changing velocity, and changing forces. In the course of a chemical reaction, there are changing concentrations of the different reactants and products. Other quantities might also change, such as temperature, pH, and volume. While observing a changing population, there could be changing population numbers, total biomass, birth and death rates, consumption of resources, and production of products and waste.

Mathematically, the **system** consists of all possible observable quantities associated with the experiment or observed physical system. The **state** of the system refers to the collection of instantaneous values of all such quantities at a particular instant or configuration of the system. A **state variable**, or more simply a variable, represents a single quantity that is or could be observed in the system. Quantities that can be calculated in terms of state variables are mathematically **dependent variables** and are also examples of state variables, even if they can not be directly measured.

Example 1.3.1 Consider the following data about the population, births and deaths in the United States. To conserve space, the data are given using scientific notation expressed in the standard machine form where the power of 10 follows the letter E, so that 2.521×10^8 would be written 2.521E8.

Year	Population	Births	Deaths		Year	Population	Births	Deaths
1991	2.530E8	4.111E6	2.170E6		2001	2.850E8	4.026E6	2.416E6
1992	2.565E8	4.065E6	2.176E6		2002	2.876E8	4.022E6	2.443E6
1993	2.599E8	4.000E6	2.269E6		2003	2.901E8	4.090E6	2.448E6
1994	2.631E8	3.953E6	2.279E6		2004	2.928E8	4.112E6	2.397E6
1995	2.663E8	3.900E6	2.312E6		2005	2.955E8	4.138E6	2.448E6
1996	2.694E8	3.891E6	2.315E6		2006	2.984E8	4.266E6	2.426E6
1997	2.727E8	3.881E6	2.314E6		2007	3.012E8	4.316E6	2.424E6
1998	2.758E8	3.942E6	2.337E6		2008	3.041E8	4.248E6	2.472E6
1999	2.790E8	3.959E6	2.391E6		2009	3.068E8	4.131E6	2.437E6
2000	2.822E8	4.059E6	2.403E6		2010	3.094E8	3.999E6	2.468E6

Each row (corresponding to the population in a given year) represents a distinct state of the system. The observed values in the state are the **vari-**

ables: the year, the total population at the beginning of the year, the total number of births in the year, and the total number of deaths in the year. The year is included as one of the variables—an independent variable—in order to distinguish the different states with respect to time. □

We often represent a variable by a symbol—a letter, a Greek letter, an abbreviation, or even a word. That symbol becomes a name for the variable to be used in sentences, expressions, and equations. Uppercase and lowercase letters are different symbols and should not be interchanged with one another. The choice of symbol should generally be related to the meaning of the variable. An important part of communication in modeling is in stating clearly the variables of a system and identifying the symbols that are chosen to represent them.

Example 1.3.2 In the previous example, there were four variables. A common strategy is to use the first letter of a word describing each variable. The population variable might be represented by the symbol P. Births and deaths might be represented by the symbols B and D, respectively. The year might be represented by the symbol Y.

Note that the symbols p, b, d and y are *not* the same as the symbols above, even though they have the same letter names. They should not be used for this problem. □

The next example illustrates how we might write a short explanation of a system and the variables associated with it. Note how the physical explanation of the system is described first, followed by an introduction of the measurements taken and the symbols used to represent those variables. Any time you have data and refer to the data by variables, you need a few sentences that introduce the meaning of each variable along with the units of measurement.

Example 1.3.3 In biology, scientists run electrophoresis gels to determine the size of polymers, such as proteins or DNA strands. The gel provides a porous structure for the polymers to travel through while an electric potential (voltage) creates a force that pulls the polymers through the gel. Different size polymers travel at different speeds. The experiment is setup with all polymers starting at one end of the gel, the voltage is turned on for a certain amount of time and then disconnected. Clusters of similarly sized polymers are identified visually as bands on the gel, with smaller polymers traveling a greater distance.

The image below represents an electrophoresis gel run on a standardized collection of DNA of fixed sizes. Because the image does not show a length scale, the distances traveled by the different lengths are measured in image pixels and recorded in the table below. The variables for the experiment are the length of DNA segments and the distance traveled through the gel. Let L represent the length of the segment (in nucleotides) and let D represent the distance traveled (in pixels), measured from the center of the starting well to the center of the corresponding band in the image. Each row represents a single state (L, D) of the system.

L (nts)	D (px)
100	342
200	327
300	312
400	299
500	288
600	278
700	270
800	263
900	256
1000	249

□

1.3.3 Scatter Plots and Relationships Between Variables

The primary motivation for collecting data regarding different variables in the state of a system is to determine relationships between those variables. One of the ways that we look for relationships is using a **scatter plot**. A scatter plot is a graph showing the relationship between two variables. Suppose the two variables use symbols x and y. For each state of the system, there will have been observed values for both x and y. The graph will include points for each pair (x, y).

Spreadsheets (like Microsoft Excel, Apple Numbers or Google Sheets) are a common tool to generate scatter plots. The data are first put in a table. The first column of data will correspond to the variable used for the horizontal axis (x), and the second column of data will correspond to the variable for the vertical axis (y). Select the two columns at the same time and add a chart to your spreadsheet, choosing the scatter plot style of graph. You should become familiar with how to create a scatter plot. Always be sure that you label your axes, using the variables of the system rather than the generic names of x and y.

The following figure shows two different scatter plots for the electrophoresis gel data above. One plot is based on the pairs (L, D) whereas the other is based on the pairs (D, L). These graphs contains the same information but viewed from a reverse perspective. When we switch the order of the variables, we call the relationships **inverse relations**.

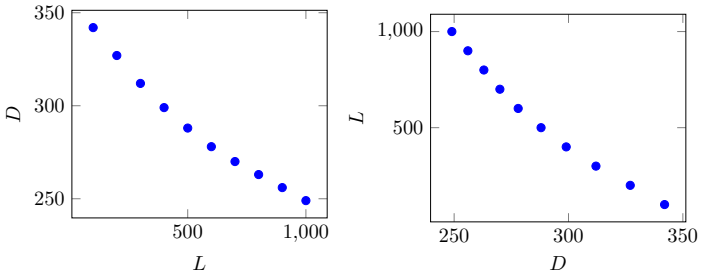


Figure 1.3.4 Scatter plot of electrophoresis data, displacement vs size and the inverse relation size vs displacement.

When a system has a state defined by more than two variables, scatter plots can be defined for each pair of state variables. For example, the population data has four state variables, (Y, P, B, D). Three scatter plots can be formed by plotting the population, the total births and the total deaths versus the year, giving graphs of points (Y, P), (Y, B), (Y, D). Because the births and deaths

are on the same scale, we can combine the plots as one. The inverse relations (P, Y), (B, Y) and (D, Y) contain the same information from a different view and are not shown.

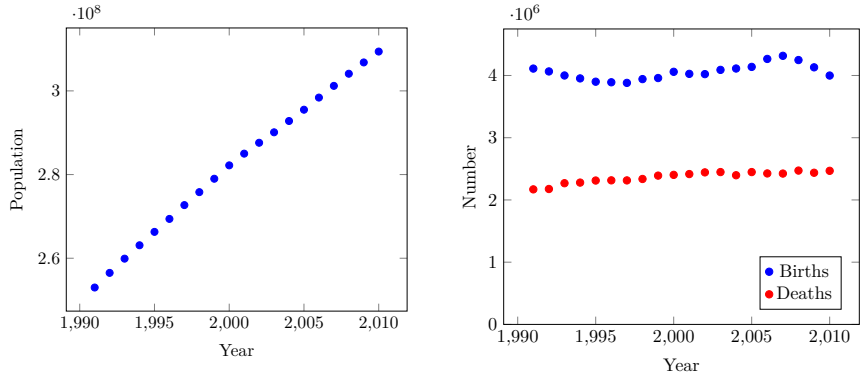


Figure 1.3.5 Scatter plot of population, births and deaths with respect to time.

We can also look at relationships between other pairs of variables. For example, we can look at how the number of births or deaths relate to the population, plotting (P, B) and (P, D), or how the number of births relate to the number of deaths with (B, D). The graph showing the relation between births and deaths to time (above) is very similar to the graph showing the relation between births and deaths to population (below). However, the relation between the births and deaths illustrates that sometimes variables do not show a clear relation.

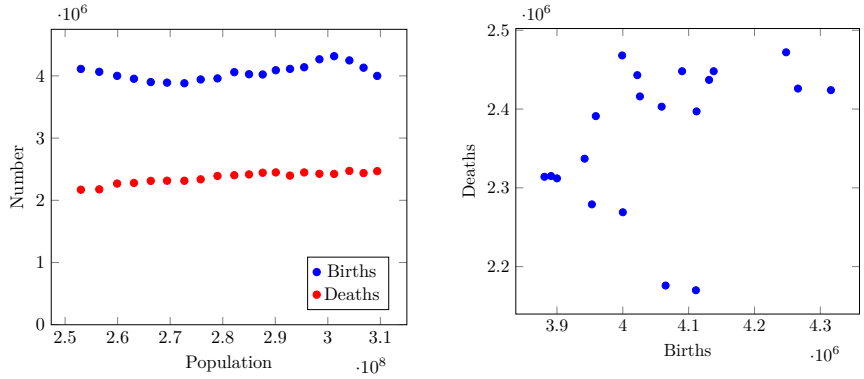


Figure 1.3.6 Graphs showing the relations (P, B), (P, D), and (B, D).

1.3.4 Graphs of Equations

An **equation** gives an abstract representation of a relationship between variables by stating that two **expressions** are equal in value. Just as the state of an experimental system is defined by the value of the variables defining the state, an equation can be considered as a mathematical way to define relationships between variables of an abstract system. A **solution** to the equation is a state for the variables such that the equation is true. The graph of an equation generalizes a scatter plot by including all solutions of the equation. If we choose an ordering for the variables (e.g., alphabetical), the values for the variables can be conveniently listed as an ordered list. When two variables are involved in an equation, the ordered list is called an ordered pair, or point, like (x, y), and the graph of the equation is typically a curve in the plane.

Example 1.3.7 The equation

$$2x + 3y = 12$$

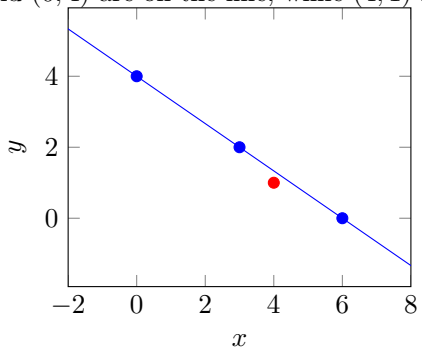
involves two variables, x and y, and is the equation of a line. The expressions in the equation are $2x+3y$ and 12. The values $x = 3$ and $y = 2$, corresponding to the ordered pair $(x, y) = (3, 2)$, provide one solution because for those values,

$$2x + 3y = 2(3) + 3(2) = 12,$$

so that the equation is true. On the other hand, $(x, y) = (4, 1)$ is not a solution because for that state,

$$2x + 3y = 2(4) + 3(1) = 11$$

and $11 \neq 12$. Some other solutions include the points $(6, 0)$ and $(0, 4)$. The line corresponding to this equation represents the set of all such solutions. The points $(3, 2)$, $(6, 0)$ and $(0, 4)$ are on the line, while $(4, 1)$ is not.



□

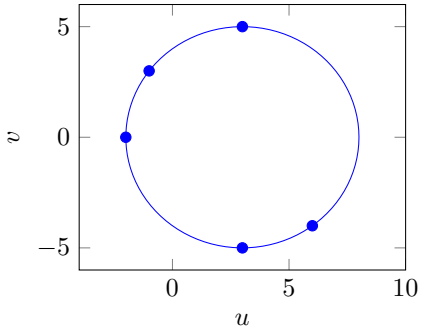
Example 1.3.8 The equation

$$u^2 + v^2 = 16 + 6u$$

also involves two variables, u and v. The expressions in the equation are $u^2 + v^2$ and $16 + 6u$. Using ordered pairs (u, v), the points $(3, 5)$ and $(3, -5)$ are solutions. That is, if $(u, v) = (3, 5)$, the expressions have the same value:

$$u^2 + v^2 = 3^2 + 5^2 = 9 + 25 = 34,$$
$$16 + 6u = 16 + 6(3) = 16 + 18 = 34.$$

It is possible to show that the graph of solutions for this equation is a circle centered at $(3, 0)$ with radius 5. Other points on this circle include such points as $(-2, 0)$ and $(6, -4)$. You should verify that these are also solutions, at least for one or two points to reinforce the idea that a solution makes the statement of the equation true.



□

It is usually difficult to know how to sketch the graph of an arbitrary equation. Computer utilities that support implicit plots can be used. For example, the online graphing calculator at desmos.com allows you to enter an equation involving variables x and y. We also could use computational systems, such as SageMath shown below, to create an implicit plot.

```
# Declare the variables
var("u,v")
# Create a graphics object from an implicit plot
myplot = implicit_plot(u^2+v^2==16+6*u, (u,-5,10), (v,-5,5))
# Show the graph with axis labels
show(myplot, axes_labels=['u','v'])
```

When an equation is written as a dependent variable being equal to an expression involving an independent variable, we can easily generate points that are in the solution set using a table with the independent variable in the first column and the dependent variable in the second column. We choose convenient values for the independent variable, compute the value of the expression that depends on that variable, and then use that resulting value for the dependent variable. All such points will be solutions to the equation. This is precisely how a graphing calculator works internally; it computes many such points very quickly and connects the points with line segments.

Example 1.3.9 Rewrite the equation $2x + 3y = 12$ so that y is the dependent variable. Use the new equation to find four points in the solution set.

Solution. We need to isolate the variable y using balanced operations.

$$\begin{aligned} 2x + 3y &= 12 \\ 3y &= -2x + 12 \\ y &= \frac{1}{3}(-2x + 12) \\ y &= -\frac{2}{3}x + 4 \end{aligned}$$

The final equation $y = -\frac{2}{3}x + 4$ should be recognized as a slope-intercept equation of a line. The slope is $m = -\frac{2}{3}$ while the y-intercept value is $b = 4$. Having solved for y, we can finish the task by using four different values for x to find corresponding values for y. We do this in a table.

x	$y = -\frac{2}{3}x + 4$	(x, y)
0	$-\frac{2}{3}(0) + 4 = 4$	$(0, 4)$
1	$-\frac{2}{3}(1) + 4 = \frac{10}{3}$	$(1, \frac{10}{3})$
2	$-\frac{2}{3}(2) + 4 = \frac{8}{3}$	$(2, \frac{8}{3})$
3	$-\frac{2}{3}(3) + 4 = 2$	$(3, 2)$

□

1.3.5 Parametrized Models and Regression Curves

Suppose we have data that appear to show a relation between two variables in a scatter plot. We would like to extend the relation to data that are not in the table of known values. If we had a mathematical equation that described our relation, we could use that equation to find the solution that would match the desired values. As a practitioner, we choose a **parametrized model** and

then use a computational tool to select the best model given our data. The most common computational strategy is called **regression**.

A **parametrized model** is an equation relating state variables that includes additional variables representing **model parameters**. The model is identified by choosing particular values for each of the parameters. Once the parameters are known, the equation establishes a relation for the state variables. A given parametrized model describes an entire family of different relations, one relation for each choice of parameters.

The most common example in algebra of a parametrized model is a linear equation

$$y = mx + b.$$

The symbols m and b are the model parameters, and x and y are the state variables. The particular equation $y = 2x - 5$ is in this family of relations based on the parameter values $m = 2$ and $b = -5$.

Another example of a parametrized model,

$$y = ax^2 + bx + c,$$

which has three parameters a, b, and c, can be used to create relations whose graphs are parabolas. The simplest parabola, $y = x^2$, corresponds to the parameter values $a = 1$, $b = 0$, and $c = 0$. Curiously, linear models are contained in this family as well by choosing $a = 0$. Our earlier example $y = 2x - 5$ could have been obtained from this model using $a = 0$, $b = 2$, and $c = -5$.

Notice that the symbols used for the parameters do not have universal meaning. In the linear parametrized models, we had chosen b to represent the y-intercept value. In the quadratic models, the parameter b was used for the coefficient of x.

Regression is a strategy to select parameters for a parametrized model in such a way that it "best" matches data for a given relation. Mathematical equations are exact. Real data exhibit uncertainty and randomness. Consequently, there usually aren't parameter values that will match all of the data simultaneously. The most common regression algorithms seek to the sum of the squared errors and are called **least-squares regression**. Spreadsheets and graphing calculators that find a trend line for data use this type of regression. We revisit finding parametrized model to match exact data in a later section.

A trend line or a trend curve resulting from regression provides a model that allows us to predict values where there are not observed data. When the prediction occurs between observed data, such prediction is called **interpolation**. If the prediction is occurring beyond the extremes of the data, such prediction is called **extrapolation**. We can use the value for one variable and the regression equation to solve for the predicted value of the related variable. Often, a formula may not describe all of the data but provides a good approximation for a certain range of values. Interpolation is usually safer than extrapolation.

Example 1.3.10 Consider the population example with the scatter plot of the number of deaths plotted with respect to the total population size. Find the linear regression model for these data and predict the number of deaths in a year if the population were 300 million.

The easiest tool to find a regression model seems to be at the website desmos.com/calculator. The site desmos.com does not support scientific notation for data entry, we can make a modified model. Let $\widetilde{P} = P/10^8$ be the population in units of 100 million and let $\widetilde{D} = D/10^6$ be the annual death rate

in units of 1 million. We are going to enter the data shown in the table below.

$\widetilde{P}$	$\widetilde{B}$		$\widetilde{P}$	$\widetilde{B}$		$\widetilde{P}$	$\widetilde{B}$		$\widetilde{P}$	$\widetilde{B}$
2.530	2.170		2.694	2.315		2.850	2.416		2.984	2.426
2.565	2.176		2.727	2.314		2.876	2.443		3.012	2.424
2.599	2.269		2.758	2.337		2.901	2.448		3.041	2.472
2.631	2.279		2.790	2.391		2.928	2.397		3.068	2.437
2.663	2.312		2.822	2.403		2.955	2.448		3.094	2.468

1. Create a table to enter the data. Either click the + menu and select `table` or type `table` in the formula field.

2. Enter the population values $\widetilde{P}$ in the column x_1 and the corresponding death rate values $\widetilde{D}$ in the column y_1. The data are now plotted and you should see they look roughly linear.

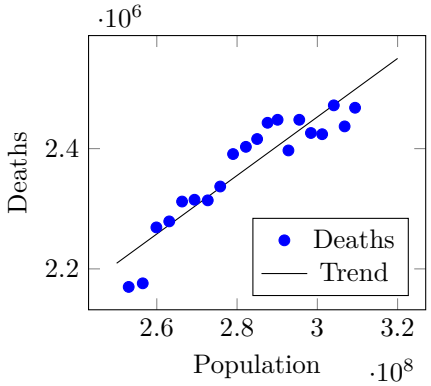
3. We now construct the parametrized model for the data. In Desmos, this is done by creating an equation using the tilde symbol `~` in place of an equals. If we want to use the parametrized model $\widetilde{D} = a\widetilde{P} + b$ with parameters a and b, we would type into the next formula `y1 ~ a x1 + b`.

4. Desmos will report values for the parameters a and b and draw the trend line through the scatter plot. The parameters are identified as $a = 0.486666$ and $b = 0.992711$ so that the model equation is

$$\widetilde{D} = 0.486666\widetilde{P} + 0.992711.$$

Specify static image with `@preview` attribute,
Or create and provide automatic screenshot as
`images/interactive-1-preview.png` via the `mbx` script

`www.desmos.com/calculator/my9xzvkfea`



We can now use the model to predict the number of deaths per year for a population of 300 million. This corresponds to $P = 300 \times 10^6 = 3 \times 10^8$ so that $\widetilde{P} = 3$. Using the parametrized model, we find

$$\begin{aligned}\widetilde{D} &= 0.486666(3) + 0.992711 \\ &= 2.452691\end{aligned}$$

Because $\widetilde{D}$ is the number of deaths in units of millions, the model predicts 2,452,691 deaths per year for a population of 300 million. Since the original data only had four significant digits, we should not expect any more digits accuracy in the model prediction. We would predict 2.453 million deaths. □

Example 1.3.11 Consider the electrophoresis gel data. Suppose we had another DNA sample of unknown length that traveled a distance of $D = 282$ pixels. Use a model to estimate the length of the DNA sample.

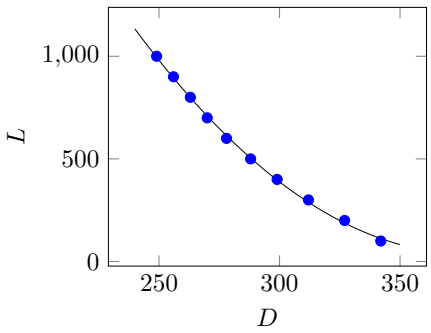
Solution. Because we know the distance displaced in the gel and want to predict the length of the polymer, we treat D as the independent variable and L as the dependent variable. We will look at the scatter plot (D, L) with the length of the DNA L graphed with respect to the distance traveled in the gel D (Figure 1.3.4). The data appear smooth with a slight upward curve. A nonlinear model will be required to model the bend, such as a quadratic parametrized model,

$$L = aD^2 + bD + c.$$

We enter the data in a table and apply regression with our model. In Desmos, we would create a table for (x_1, y_1) with values of D in x_1 and values of L in y_1. We then calculate model parameters using `y1 ~ a x1^2 + b x1 + c`. The resulting model parameters are $a = 0.0573428$, $b = -43.381$, and $c = 8241.57$. Consequently, the trend curve is modeled by

$$L = 0.0573428D^2 - 43.381D + 8241.57.$$

The graph of the data with the trend curve is shown below.



Using our value for D, we can find the value of L using the model,

$$L = 0.0573428(282)^2 - 43.381(282) + 8241.57 \approx 568.26.$$

Since our original data had 3 significant digits, we would estimate the length of the DNA in question as $L \approx 568$ nucleotides. In this way, a regression of known electrophoresis data allows us to estimate lengths of other molecules. □

You should note that the number of significant digits reported is not the same as the uncertainty in the prediction. The degree to which the original data vary around the trend curve leads to uncertainty in the coefficients of the regression model and subsequent uncertainty to the trend curve itself. In the last example, rounding the model parameters themselves to 3 significant digits would have changed the predicted length by 11 nucleotides. Analysis of this uncertainty is a topic for statistics and is outside the scope of this text. For simplicity, we use models to make predictions and then round to comparable precision as the data.

1.3.6 Summary

- Quantities that can be measured correspond to state variables. A system is the collection of all possible variables. The state of the system is the collection of values measured for all of the variables simultaneously. An important part of communication is describing all relevant variables and introducing their names.
- A relation between two variables can often be visualized graphically using a scatterplot. An equation is the mathematical idealization of a relation. The graph of an equation involving two variables, say x and y, shows all solutions as points (x, y).
- When the equation is written as a dependent variable equal to an expression of the dependent variable, points on the graph can be quickly tabulated using the formula.
- Using regression to find a trend line or regression curve can give an approximate relation corresponding to observed data. Treating the resulting equation as a model equation can give approximate predictions of states of the system.

1.3.7 Exercises

Each problem has an equation involving two variables. Determine whether each of the given states for those variables are in the solution set.

1. $3x - 2y = 8$

(a) $(x, y) = (0, -4)$

(b) $(x, y) = (1, -2)$

(c) $(x, y) = (4, 2)$

2. $2w + 5z - 3 = w^2 + z^2$

(a) $(w, z) = (-2, 2)$

(b) $(w, z) = (-1, 3)$

(c) $(w, z) = (3, 2)$

Each problem has an equation involving multiple variables. Solve for the indicated dependent variable.

3. **Perimeter of Rectangle.** Given $2L + 2W = P$, solve for W.

4. **Volume of Rectangular Prism.** Given $V = LWH$, solve for L.

5. **Volume of Cylinder.** Given $V = \pi r^2 h$, solve for h.

6. **Ideal Gas Law.** Given $PV = nRT$, solve for P.

Given an equation relating two variables, solve for the indicated dependent variable. Use your resulting expression to calculate the value for that variable given the values of the indicated independent variable. Make note of any values that are not defined. Plot the corresponding points in the solution set of the equation on a graph.

7. Given the equation $4x+5y = 20$, find y for each value $x \in \{1, 2, 3, 4, 5\}$.

8. Given the equation $np = 1000$, find n for each value $p \in \{10, 20, 25, 40, 50\}$.

Graph two dependent variables representing the expressions on each side of the equation. Use the points of intersection to identify solutions to the equation. Verify that the values you identify are solutions by testing whether make the equation true.

9. $3x - 5 = x + 2$

10. $\dfrac{20x}{x+4} = x + 3$

11. $4x^3 - 9x^2 = x - 6$

Additional DNA samples were run in the same electrophoresis gel as described in Example 1.3.11. Using the data and regression curve from that example, estimate the length of each sample. Indicate whether the approximation is appropriate.

12. Estimate the length of a DNA sample that traveled 200 pixels.

13. Estimate the length of a DNA sample that traveled 335 pixels.

14. Estimate the length of a DNA sample that traveled 350 pixels.

A Voltage–Resistance–Current Relationship A simple electric circuit has an applied voltage V (volts) and a variable load resistance R (kilohms). When the circuit is closed, current flows through the circuit, measured as the current I (amperes). When the voltage was held constant at $V = 9$ V, the resistance and current were measured with values recorded in the table below. The following group of problems are based on these data.

V (V)	R (kΩ)	I (A)
9.0	0.84	0.0107
9.0	1.2	0.0073
9.0	1.8	0.0050
9.0	2.7	0.0033
9.0	3.4	0.0026

15. Create a scatter plot of (R, I). Would a trend line make sense for this data? Explain.

16. Conductance G is the reciprocal of resistance, $G = 1/R$. Create a scatter plot of (G, I). Would a trend line make sense for this data? Explain.

17. One of the previous scatter plots should have had a meaningful trend line. State an appropriate regression equation as a model and use it to predict the current I when the resistance is $R = 2.1$ kΩ.

Population-Growth Relationships The number of births and of deaths in a population generally depends on the size of the population. The table below gives population data for ten of the twelve highest population cities in the state of Virginia for the year 2012. The data include the population P and the total number of births B and deaths D for the year recorded for each city. The following group of exercises are based on these data.

City	P	B	D
Virginia Beach	447021	6270	2828
Norfolk	245782	3773	1827
Chesapeake	228417	2805	1582
Richmond	210309	2939	1849
Newport News	180726	2905	1438
Alexandria	146294	2763	686
Roanoke	97469	1492	1172
Portsmouth	96470	1534	980
Suffolk	85181	1087	726
Lynchburg	77113	1062	779

18. Create a scatter plot of (P, B) and find the equation of the trend line. The cities of Hampton and Harrisonburg were left off the list with populations of $P = 136836$ and $P = 50981$, respectively. Use the trend line regression model to predict the number of births in these cities during 2012. Which calculation is an example of interpolation and which is extrapolation?

19. Create a scatter plot of (P, D) and find the equation of the trend line. Use the trend line regression model to predict the number of deaths $\cdot$ in Hampton and Harrisonburg during 2012. (See the previous problem for population values.) Which calculation is an example of interpolation and which is extrapolation?

Which of the calculations were examples of interpolation and which were examples of extrapolation?

1.4 Formulas as Models of Relations

1.4.1 Overview

Models are simplified abstract representations of something of interest. Airplane and automobile manufacturers create scale models to test aerodynamics in wind tunnels. Architects build models of future projects, whether a physical mock-up or a computerized 3-d representation, to see how their plan will fit together and give clients a vision of pending products. The models do not need to include every detail of the actual object of interest, just those details that are relevant to the purpose of the model.

Scientists also regularly use models. Physicists use high energy collisions of extremely fast particles to create conditions that they expect are comparable to the moments immediately after the big bang. A biologist may use mice from a well-controlled population as a model to study cancer, considering its biology to mimic that of humans at some level of approximation. A climatologist might use a computational model where a computer program tracks changes in the makeup of the air, pollutant levels and air and ocean temperatures according to known and assumed interactions.

A **mathematical model** is an abstract representation of measurable phenomena that is characterized through mathematical equations. Recall that we think of a **system** as the collection of all possible measurements associated with the objects and environment involved in the phenomenon. Each quantity is a **state variable**, even if we do not have a physical way to obtain the measurement. Many laws of science are described using mathematical equations that relate state variables. These equations are examples of mathematical models. Knowing the value of one state variable, we can use the model to predict the value of other variables.

In this section, we explore some of the most common parametrized formulas used in mathematical models. The most important concept that relates many of these families is the idea of proportionality. When we have a mathematical model, we can use values for the variables to solve for unknowns. We will introduce ways to solve equations using technology in this section and review some strategies for solving equations by hand in later sections.

1.4.2 Proportionality

The idea of proportionality occurs everytime there is a common ratio between two quantities.

Example 1.4.1 In chemistry, we know that the atomic mass of an element (daltons) represents the mass (grams) of exactly one mole of atoms of that element. The atomic mass of carbon-12 is exactly 12 Da. Thus, 1 mole of carbon-12 atoms has a mass of 12 grams, 2 moles of carbon-12 atoms has a mass of 24 grams, and 5 moles of carbon-12 atoms has a mass of 60 grams. The ratio of the number of moles to the mass is always the same constant matching the atomic mass. We say that the mass is **proportional to** the number of moles. □

So what does proportionality really mean? A **proportion** is a ratio between two quantities. The quantities are **proportional** if the ratio between the quantities always equals the same value. We sometimes say that the two quantities have a common proportion.

Definition 1.4.2 A quantity Q is **proportional to** a quantity P if the ratio Q/P is a constant, say $Q/P = k$. The value k is called the **proportionality**

constant and we can rewrite the equation as

$$Q = kP.$$

◊

Many laws of physics are statements of proportionality. Isaac Newton discovered that the force F acting on an object due to gravity is proportional to the mass m of the object. Newton's law of gravity could be written

$$F = mg,$$

where the constant g is called the gravity acceleration constant. The French physicist Charles-Augustin de Coulomb discovered a similar law, that the electrical force F acting on a charged object is proportional to the charge of the object q. Coulomb's law can be written

$$F = qE,$$

where E is the electrical field strength at the object's location. The German physicist Georg Ohm discovered that the voltage drop V across a conductor in a circuit is proportional to the current I flowing through that conductor. Ohm's law is written

$$V = IR$$

and R is called the resistance of the conducting component.

When we know that two quantities are proportional, we can find the proportionality constant by calculating the ratio given observed data. If there are errors or uncertainties in the data, we can approximate the proportionality constant using an average of calculated ratios.

Example 1.4.3 Suppose we know that the birth rate for a population (number of births per unit time) is proportional to the number of individuals in the population. If the population has 20 births per month when it consists of 5000 individuals, find the number of births per month when the population consists of 8000 individuals.

Solution. We start by assigning variables for our quantities. Let P be the size of the population and let B be the birth rate. Because the birth rate (births per month) is proportional to the population size (individuals), we know that the ratio B/P is equal to some constant, which we will name b:

$$b = \frac{B}{P} = \frac{20}{5000} = 0.004.$$

The constant b is called a **per capita birth rate**. Rewriting the equation $B/P = b = 0.004$, we have a model

$$B = bP \qquad \Leftrightarrow \qquad B = 0.004P.$$

We now use our model. When the population has 8000 individuals, we have $P = 8000$. Substituting this value in the model, we find

$$B = 0.004(8000) = 32.$$

That is, the population is predicted by our model to have 32 births per month. □

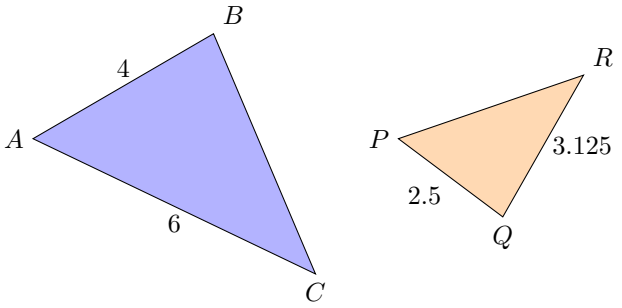
Mathematically, the geometric idea of similarity is one of the most common sources of proportional relations. Given two geometric polygons, we need a way to associate each vertex (points where edges meet) of one polygon with a

particular vertex on the other polygon. The polygons are usually labeled by vertices, like a triangle ABC or a quadralateral $PQRS$. If we relate triangle ABC to triangle JKL, then the corresponding vertices would be

$$A \leftrightarrow J \qquad B \leftrightarrow K \qquad C \leftrightarrow L.$$

Polygons are **similar** if the ratio of distances between any pair of corresponding vertices always equal the same constant.

Example 1.4.4 Suppose triangle ABC is similar to triangle PQR, shown in the figure below. The lengths of the edges AB and AC are 4 and 6, respectively. The lengths of the edges PQ and QR are 2.5 and 3.125, respectively. Find the lengths of the other two edges.



Solution. The association between vertices of the triangles are

$$A \leftrightarrow P \qquad B \leftrightarrow Q \qquad C \leftrightarrow R.$$

Because the triangles are similar, the ratios of corresponding edges must all have the same value,

$$\rho = \frac{AB}{PQ} = \frac{AC}{PR} = \frac{BC}{QR}.$$

Because we know the lengths of both AB and PQ, we can use those values to determine the common ratio,

$$\rho = \frac{AB}{PQ} = \frac{4}{2.5} = 1.6.$$

Using this ratio, we can solve for the remaining unknowns:

$$1.6 = \frac{AC}{PR} = \frac{6}{PR} \qquad \Rightarrow \qquad PR = \frac{6}{1.6} = 3.75$$
$$1.6 = \frac{BC}{QR} = \frac{BC}{3.125} \qquad \Rightarrow \qquad BC = 1.6(3.125) = 5$$

Thus, $PR = 3.75$ and $BC = 5$. □

The last example of simple proportionality arises in linear relations. Traditionally, we think of y as the dependent variable and x as the independent variable. Given any two points in the relation $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$, we calculate the ordered increments of change going from P_1 to P_2 as $\Delta x = x_2 - x_1$ and $\Delta y = y_2 - y_1$. We always calculate increments of change as the ending value minus where the starting value. A relation is **linear** if the ratio $\Delta y/\Delta x$ is always the same constant. This constant value is called the **slope**, traditionally using the symbol m,

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}. \tag{1.4.1}$$

The slope represents the proportionality constant between Δx and Δy. If we know any point $(x, y) = (a, b)$ that satisfies the linear relation and the slope, then if we take any other point (x, y), we must have

$$y - b = m\,(x - a). \tag{1.4.2}$$

This equation is called the point–slope equation of a line.

Example 1.4.5 A container of ice water containing 200 grams of ice and 600 grams of water requires heat added to raise its temperature. The relation between temperature increase and required energy in heat is linear. To raise the temperature 10 degrees Celsius, we need 110.19 kJ. To raise the temperature 20 degrees Celsius, we need 133.67 kJ. How much energy is required to raise the temperature 25 degrees Celsius?

Solution. Because the relation between the temperature increase and the energy added to the water is linear, we can calculate the slope as the ratio of change in these values. Let T be the amount the temperature rises and let Q be the energy in heat added. Thinking of Q as the dependent variable, the slope will be the ratio of the change in Q to the change in T,

$$m = \frac{\Delta Q}{\Delta T} = \frac{133.67 - 110.19}{20 - 10} = 2.348.$$

Knowing the slope, we can create an equation that models the relation between Q and T. The point–slope equation of a line using the point $(T, Q) = (10, 110.19)$ becomes

$$Q - 110.19 = 2.348\,(T - 10).$$

To find the energy required to raise the temperature 25 degrees Celsius, we substitute $T = 25$ and can then solve for Q.

$$\begin{aligned} Q - 110.19 &= 2.348\,(25 - 10) \\ Q - 110.19 &= 35.22 \\ Q &= 145.41 \end{aligned}$$

It will take 145.41 kJ to raise the temperature to 25 degrees Celsius. □

1.4.3 Common Models

Many common models are based on a relation of proportionality between a quantity and a dependent variable based on another quantity. Simple proportionality would be where a variable y is proportional to another variable x, $y = kx$. We are now interested in models where y is proportional to some simple function of x. We say that y is **inversely proportional to** x if y is proportional to the reciprocal of x,

$$y = k \cdot \frac{1}{x} = \frac{k}{x}.$$

This is equivalent to saying that the product $xy = k$ is constant.

A **power law** refers to a dependent variable being proportional to some power p of the independent variable,

$$y = Ax^p.$$

An **exponential relation** refers to a dependent variable being proportional to some positive base b being raised to the independent variable as the power,

$$y = A\,b^x.$$

Although these two laws look similar, they have very different behaviors because of the role of the independent variable. Find ways to train your mind to read these differently. For example, for the power law, you might read the formula as "x raised to a power p", but for the exponential relation, you would read the formula as "the exponential of base b raised to the power x". As you intentionally use language to distinguish between the two models, you will more easily remember appropriate methods for each.

We now consider how we would find the model parameters given data. Regression methods do exist. For example, in Desmos, it would be possible to put given data in a table and then compute parameters using a formula like `y1 ~ A x1^p` or `y1 ~ A b^x1`. You would get perfectly good approximate values that could be used for applications. However, our purpose is to introduce the role of equations in solving for exact values.

When we are given data that a model describes exactly, we should consider each point in the data as being a solution to the model equation. When we have multiple points, we have multiple equations, all of which need to be true simultaneously. We then treat the parameters as variables and solve the system of equations to find parameter values.

We should have a clear separation in our thinking between creating the system of equations and solving the resulting system. We first focus on creating the system of equations, and we will use a computational tool to solve for the values. We will look at graphical methods of approximating solutions as well as using computer algebra systems to find exact formulas.

Example 1.4.6 Suppose y has a power law relation with x, $y = Ax^p$. Further, suppose that we have two data points, $(x, y) = (2, 4)$ and $(x, y) = (3, 8)$. Find the equations that determine the model parameters. Graph the equations to find their values.

Solution. To find each equation, we substitute the values of x and y in the model equation. For each given point, this will leave an equation that still involves the unknown model parameters. Using the point $(x, y) = (2, 4)$, we substitute $x = 2$ and $y = 4$ in $y = Ax^p$:

$$4 = A \cdot 2^p.$$

Using the point $(x, y) = (3, 8)$, we substitute $x = 3$ and $y = 8$ in $y = Ax^p$:

$$8 = A \cdot 3^p.$$

To show that we have a system of equations that need to be solved together, we group the equations with a curly brace,

$$\begin{cases} A \cdot 2^p = 4, \\ A \cdot 3^p = 8. \end{cases}$$

Our first method for finding the values will be graphical. Because we have two parameters, A and p, we can think of these as two variables that define a plane of points (p, A). Each equation defines a curve in the plane of solutions to that equation. With two equations, we obtain two different curves. Intersection points are the points in common to both curves and are the solutions that we seek.

Some graphing tools, like Desmos and most graphing calculators, require the variables to be x and y. They also typically expect that we have solved for y as a dependent variable. If we replace $p \leftrightarrow x$ and $A \leftrightarrow y$ and then solve for y, our equations become

$$\begin{cases} y = 4/2^x, \\ y = 8/3^x. \end{cases}$$

In Desmos, you can click on the point of intersection and see approximate values $(1.71, 1.223)$. Using a handheld calculator, a menu option to find an intersection point gives a better approximation $(1.709511, 1.223055)$ meaning that $p \approx 1.7095$ and $A \approx 1.2231$. A graph showing the points and the approximate model $y = 1.2231 \cdot x^{1.7095}$ is shown below.

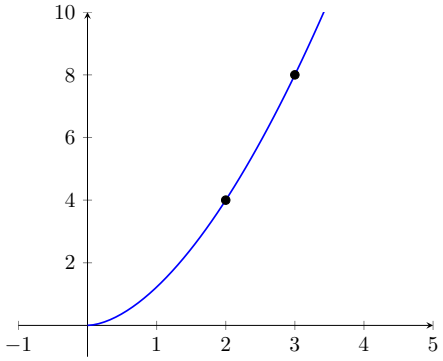


Figure 1.4.7 Graph of $y = 1.2231 \cdot x^{1.7095}$ with given points $(2, 4)$ and $(3, 8)$.

□

Example 1.4.8 Similar to the previous example, suppose y has an exponential relation with x, $y = Ab^x$ with the two data points, $(x, y) = (2, 5)$ and $(x, y) = (3, 8)$. Find the equations that determine the new model's parameters and solve the system of equations.

Solution. To find each equation, we substitute the values of x and y in the model equation. For each given point, this will leave an equation that still involves the unknown model parameters. Using the point $(x, y) = (2, 5)$, we substitute $x = 2$ and $y = 5$ in $y = Ab^x$ to obtain

$$5 = A \cdot b^2.$$

Using the point $(x, y) = (3, 8)$, we obtain the equation

$$8 = A \cdot b^3.$$

Our system of equations becomes

$$\begin{cases} A \cdot b^2 = 5, \\ A \cdot b^3 = 8. \end{cases}$$

The graphical method of solution used in the previous example results in approximate values. To find exact values, we need to perform an algebraic solution. A computational tool will expect us to provide it with our equations as well as the variables for which we are solving. In this text, we will work with the SageMath system, an open source computer algebra system. A blank interactive SageMath cell can be opened at https://sagecell.sagemath.org.

```
# Tell SageMath that A and b should be treated as variables
var('A','b')
# Define our equations
eq1 = (A*b^2 == 5)
eq2 = (A*b^3 == 8)
# Solve the system of equations for the system of variables
soln = solve([eq1, eq2], A, b)
# Display the result
show(soln)
```

When this script is executed, SageMath reports a solution

$$A = \frac{125}{64}, \quad b = \frac{8}{5}.$$

Our model becomes $y = \frac{125}{64} \cdot \left(\frac{8}{5}\right)^x$. A graph of the model with our given points is shown in the figure below.

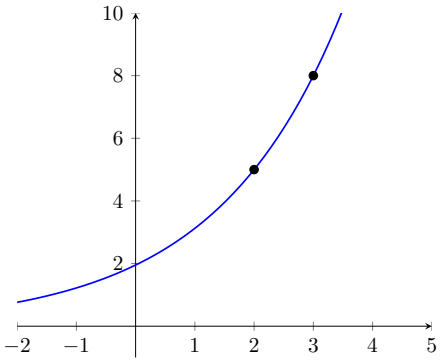


Figure 1.4.9 Graph of $y = \frac{125}{64} \cdot \left(\frac{8}{5}\right)^x$ with given points $(2, 5)$ and $(3, 8)$.

I want to note that SageMath fails to solve the similar system for the previous power law example. The change to the script above is minimal. You should try it and see what happens. This is a curious example where computer algebra systems sometimes know too much. In more advanced mathematics that account for complex numbers, there is an abiguity in that system of equations that is not obvious. You can see a glimpse of the complexity if you try to solve the system of equations using the WolframAlpha website. Try the request `solve A*2^p=4 and A*3^p=8 for A and p`. □

1.4.4 Constructing More Models

Once we have elementary models like powers and exponentials, we can construct more complicated models by using arithmetic. For example, **polynomials** are created by adding power functions where each of the powers are non-negative integers.

Definition 1.4.10 A **polynomial** is a sum of terms of the form $a_k x^k$ where k is a non-negative integer, a_k is a real number, and x is the independent variable. The values a_k are called **coefficients**. The largest power k is called the **degree** of the polynomial. ◊

The equation $y = 4x^3 - x + 5$ is a polynomial with degree 3, which we call a **cubic polynomial**. The coefficients are $a_0 = 5$, $a_1 = -1$, and $a_3 = 4$. A

model for y as cubic polynomial of x would look like

$$y = a_3x^3 + a_2x^2 + a_1x + a_0.$$

We usually include zero coefficients for any skipped powers smaller than the degree, so our example would also have $a_2 = 0$.

We can create a system of equations to find coefficients for a polynomial. Because a cubic polynomial has 4 coefficients, we will need four data points to find a unique solution.

Example 1.4.11 Find a quadratic polynomial (degree 2) that goes through the points $(-1, 1)$, $(1, 2)$, and $(2, 4)$.

Solution. A general degree 2 polynomial model would have the form

$$y = a_2x^2 + a_1x + a_0.$$

We substitute the values for x and y to get one equation for each point.

$$\begin{cases} 1 = a_2(-1)^2 + a_1(-1) + a_0 \\ 2 = a_2(1)^2 + a_1(1) + a_0 \\ 4 = a_2(2)^2 + a_1(2) + a_0 \end{cases} \quad \Leftrightarrow \quad \begin{cases} 1 = a_2 - a_1 + a_0 \\ 2 = a_2 + a_1 + a_0 \\ 4 = 4a_2 + 2a_1 + a_0 \end{cases}$$

We now use a computer algebra system to solve for the coefficients.

```
# Declare the parameters as variables
var('a2','a1','a0')
# Create the equations
eq1 = 1 == a2 - a1 + a0
eq2 = 2 == a2 + a1 + a0
eq3 = 4 == 4*a2 + 2*a1 + a0
# Solve the system
soln = solve([eq1,eq2,eq3], a0, a1, a2)
show(soln)
```

The result of the algebra solution is

$$a_0 = 1, \qquad a_1 = \frac{1}{2}, \qquad a_2 = \frac{1}{2}.$$

That is, our polynomial model is

$$y = \frac{1}{2}x^2 + \frac{1}{2}x + 1.$$

A graph of the data with the model is shown in the next figure.

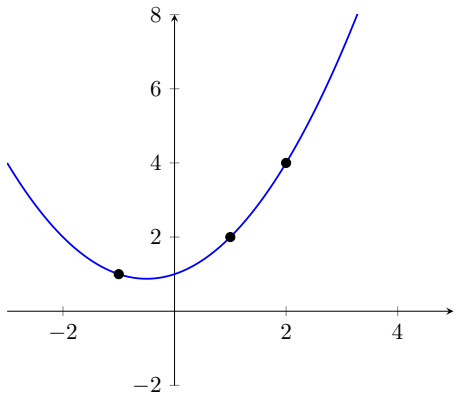


Figure 1.4.12 A graph of $y = \frac{1}{2}x^2 + \frac{1}{2}x + 1$ with the points $(-1, 1)$, $(1, 2)$, and $(2, 4)$.

□

We end this section by discussing the difference between regression and solving equations. Regression is used when we have data that we want to approximate with a model. This means that given data points will not necessarily be solutions—and most likely, they will not be. When we solve for a model passing through given data, we are finding a curve that has the given data points as solutions. Each point must actually lie on the curve. If we tried to solve for a model where we should be doing a regression, we will discover that there is no solution.

Example 1.4.13 Consider the three points $(1,-2)$, $(3,1)$, and $(6,6)$. Can we model these with a linear function $y = mx + b$?

Solution. Using the three data points, we can create a system of three equations for the model parameters.

$$\begin{cases} -2 = m(1) + b \\ 1 = m(3) + b \\ 6 = m(6) + b \end{cases} \quad \Leftrightarrow \quad \begin{cases} m + b = -2 \\ 3m + b = 1 \\ 6m + b = 6 \end{cases}$$

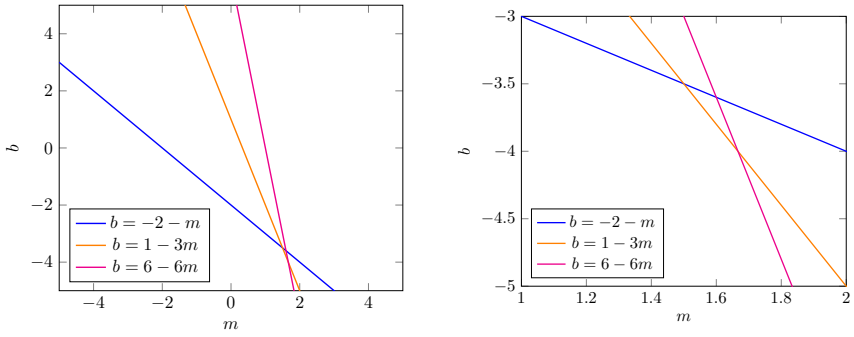
If we try to solve the system of equations, the solution set is empty.

```
var('m','b')
eq1 = m + b == -2
eq2 = 3*m + b == 1
eq3 = 6*m + b == 6
soln = solve([eq1,eq2,eq3],m,b)
show(soln)
```

To visualize why there is no solution for this system, let us consider the graphical approach. We can solve each equation for b as the dependent variable and graph the resulting equations. Recall that many graphing utilities require us to rename our variables, $m \leftrightarrow x$ and $b \leftrightarrow y$:

$$\begin{cases} b = -2 - m \\ b = 1 - 3m \\ b = 6 - 6m \end{cases} \quad \leftrightarrow \quad \begin{cases} y = -2 - x \\ y = 1 - 3x \\ y = 6 - 6x \end{cases}$$

When we graph these equations, we get three lines. Although they appear close to having a single point of intersection near $(m, b) = (1.6, -3.6)$, there is not point where all three lines intersect simultaneously. This is the graphical result of no solution.

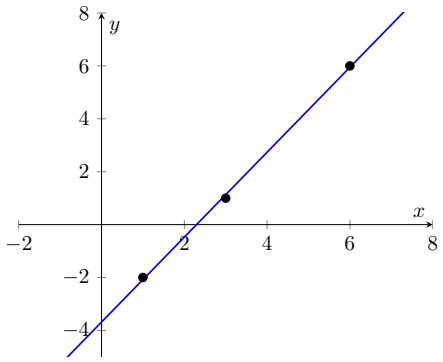


When we approach this problem as a regression model, we seek for a linear equation that is closest to all three points. Desmos reports regression coefficients $m \approx 1.60526$ and $b = -3.68421$ for a regression model

$$y = 1.60526x - 3.68421.$$

Notice that this model only approximates our points of interest:

$$\begin{aligned} x = 1 \quad &\Rightarrow \quad y = -2.07895 \\ x = 3 \quad &\Rightarrow \quad y = 1.13158 \\ x = 6 \quad &\Rightarrow \quad y = 5.94737 \end{aligned}$$



□

1.4.5 Summary

- Proportionality between two variables is when the ratio of their values is always the same constant. As a parametrized model, to say y is proportional to x means that $y = kx$ for some constant k.
- Many models are generated by expressing a dependent variable as being proportional to a function of the independent variable. Common exampes are inverse proportionality, power law relations, and exponential relations.
- Given a parametrized model, a data point for the variables establishes an equation in the model parameters. With enough data points, the resulting system of equations can be solved to find the parameters.
- Solutions to a system of equations can be approximated based on the intersection of graphs. Computer algebra systems can often help solve a system of equations to find exact solutions.

1.4.6 Exercises

1. The British physicist Robert Hooke observed that when a spring is stretched, the strength of the force is proportional to the length of the stretch, at least for small to moderate lengths. When a spring is stretched 5 cm, the force exerted by the spring is 1.8 N.

(a) If F is the force of the spring and L is the length the spring is stretched, write down the general equation describing the relation that F is proportional to L. Then use the given data to find the proportionality constant.

(b) Find the force exerted by the spring if it is stretched 8 cm.

(c) What lengths should the spring be stretched for a force of 1 N? 2 N?

2. The amount of heat (a form of energy) stored in a substance is proportional to the change in temperature. When a gram of water absorbs 10 J

of heat, the temperature rises by 2.389 degrees Celsius.

(a) If ΔT is the change in the temperature of the water and Q is the heat added to the water, write down the general equation describing the relation that Q is proportional to ΔT. Then use the given data to find the proportionality constant, which is called the specific heat of water.

(b) How much will the temperature change if 16 J of heat is added to the water.

(c) How much energy in added heat is required to raise a gram of water from 20 degrees to 100 degrees?

3. The surface area of a cube is proportional to the square of the length of one side.

(a) If A is the surface area and s is the length of a side, write down the general equation describing the relation that A is proportional to the square of s.

(b) Using geometrical reasoning, what is the proportionality constant?

(c) What is the surface area of a cube whose sides are each 5 cm in length?

(d) What the length is the side of a cube whose surface area is exactly $1\,\text{cm}^2$ in length?

4. The mass of a raindrop is proportional to the cube of its diameter. A raindrop with a diameter of 3 mm has a mass of 14.137 mg

(a) If m is the mass of a raindrop and d is the diameter, write down the general equation describing the relation that m is proportional to the cube of d.

(b) Find the constant of proportionality using the given data.

(c) What is the mass of a raindrop with a diameter of 5 mm?

(d) What is the diameter of a raindrop with a mass of 25 mg?

5. The time to complete a large manual labor job is inversely proportional to the number of people performing the labor. Suppose a job will take 20 days when 5 people are working.

(a) If T the time required to complete the job and L is the number of laborers, write down the general equation describing the relation that T is inversely proportional to L.

(b) Find the constant of proportionality using the given data. What is the physical interpretation of this constant?

(c) How long will the job take if there are 16 people working?

(d) What is the fewest number of people that can complete the job in 8 days?

6. The intensity of radiation from the sun is inversely proportional to the square of the distance from the sun. The earth, which is 1 AU (atronomical unit) from the sun, receives radiation from the sun at an intensity of 1367 W/tothe2.

(a) If I the radiation intensity and r is the distance fom the sun, write

down the general equation describing the relation that I is inversely proportional to the square of r.

(b) Find the constant of proportionality using the given data.

(c) Find the intensity of radiation at Mercury, which is 0.387 AU from the sun.

(d) Find the distance at which the sun's radiation is half the intensity as compared to earth.

7. A right triangle ABC has legs $AC = 4$ and $BC = 3$.

(a) Find the length of the hypotenuse AB by applying the Pythagorean theorem.

(b) Find the lengths of a triangle PQR that is similar to ABC whose hypotenuse has length $PQ = 1$.

(c) Find the lengths of a triangle STU that is similar to ABC whose leg SU has length $SU = 1$.

8. A right triangle ABC has a leg AC with length $AC = 2$ and a hypotenuse AB with length $AB = 3$.

(a) Find the length of the other leg BC by applying the Pythagorean theorem.

(b) Find the lengths of a triangle PQR that is similar to ABC whose hypotenuse has length $PQ = 1$.

(c) Find the lengths of a triangle STU that is similar to ABC whose leg SU has length $SU = 1$.

9. Consider the parametrized model $y = ax^2 + b$ and data points $(x, y) = (1, 3)$ and $(x, y) = (2, 9)$.

(a) Determine the system of equations in terms of a and b.

(b) Graph the system of equations for the parameters in the (a, b)-plane.

(c) Solve for a and b, using a computer to assist if needed.

(d) Find the value of y when $x = 4$.

10. Consider the parametrized model $y = ax^2 + bx$ and data points $(x, y) = (1, 3)$ and $(x, y) = (2, 9)$.

(a) Determine the system of equations in terms of a and b.

(b) Graph the system of equations for the parameters in the (a, b)-plane.

(c) Solve for a and b, using a computer to assist if needed.

(d) Find the value of y when $x = 4$.

11. Consider the parametrized power law model $y = a \cdot x^b$ and data points $(x, y) = (1, 3)$ and $(x, y) = (2, 9)$.

(a) Determine the system of equations in terms of a and b.

(b) Graph the system of equations for the parameters in the (a, b)-plane.

(c) Solve for approximate values for a and b using a graphing utility.

(d) Find the value of y when $x = 4$.

12. Consider the polynomial model $y = ax^2 + bx + c$ and data points $(x, y) = (1, 3)$, $(x, y) = (3, 6)$, and $(x, y) = (-2, 4)$.

(a) Determine the system of equations in terms of a, b, and c.

(b) Solve the system of equations to find exact values for the model parameters

(c) Find the value of y when $x = 6$.

1.5 Algebra and Equivalence

1.5.1 Overview

Algebra can sometimes feel complicated. This feeling often arises when algebra is viewed as a long list of rules of manipulation. Perhaps you think of algebra as rules for moving symbols and then feel overwhelmed by the number of possible different problems. Or you might feel like you are doing the same operation in different situations and are counted correct in some situations but incorrect in others.

To be effective in algebra, we should organize our thinking around a small number of core principles. The first principle is to distinguish between expressions and equations. Our goals when working with an expression are different from when we are working with an equation. The second principle is the idea of equivalence, whether that refers to equivalent expressions or equivalent equations. Third, we want to minimize the number of rules be relating them to the fundamental properties of algebra.

In this section, we will relate the algebra principles you should have previously learned around these core principles. We will focus on the idea that algebra identities allow us to replace one expression with another equivalent expression. The operations on expressions all originate with the basic properties of real number arithmetic. Equations are relations stating that two expressions have the same value. Our operations on equations will be designed to generate simpler equations that are equivalent to the original. These strategies focus on applying the same operation to both sides of the equation to maintain a balance, with the goal of creating an equation that is more easily solved than the original.

1.5.2 Expressions and Properties of Algebra

In algebra, we use variables as placeholders for numerical values. The variable is given a symbol, usually a letter like x, that takes the place of the value. Sometimes, this is because we want to describe a calculation without referencing a specific value. Other times, we want a specific but unknown value and use a variable as a name.

Example 1.5.1 Maybe you have seen a calculation described similarly to the following.

Think of a number. Add five. Double the result. Subtract four. Divide the answer in half. Subtract your original number.

Use the variable x to represent the number chosen. Write out the formula that describes this calculation.

Solution. We use parentheses to emphasize the order of calculations. Start with x and add five to get $x+5$. Doubling this means to multiply by two to get $2(x+5)$. Subtracting four gives $2(x+5)-4$. Dividing in half means divide this by two, resulting in $\dfrac{2(x+5)-4}{2}$. We end by subtracting the original value x. The formula that matches this calculation would be

$$\frac{2(x+5)-4}{2} - x.$$

□

An **expression** is any *formula* involving numbers and variables, such as the formula in the previous example. An expression itself represents a numerical value. When an expression involves variables, the value of the expression itself

is unknown until the values of all variables are known. Consequently, the expression itself is a **dependent variable**, and we often represent its value by another symbol.

Example 1.5.2 For the previous example, we could introduce a variable y to represent the final number. That gives a dependent variable defined by the expression

$$y = \frac{2(x+5)-4}{2} - x.$$

□

Dependent variables often depend on more than one variable.

Example 1.5.3 For a business that earns money by selling a number of items, each of which is sold for the same price, the **revenue** (money brought in through sales) is computed by multiplying the number of items sold by the price of each item. We can summarize this statement defining revenue using algebra if we represent the state variables by symbols. Let R represent the revenue, let n represent the number of items sold, and let p represent the price of each item. The product $n \cdot p$ is the expression representing the product of the number of items and the price per item. We use the equation

$$R = n \cdot p$$

to describe that the revenue is computed by this expression.

We see that the revenue R is defined here as a dependent variable based on the values of n and p. If we know the value of both n and p, then we will know the value of R. For example, if the company sells $n = 1000$ items and sets a price of $p = 1.25$ dollars, then the revenue is $R = 1000 \cdot 1.25 = 1250$ dollars. □

Different expressions can represent the same value. For example, $x + x$ and $2x$ are different expressions—they describe different calculations—but they always have the same value. We say that two expressions are **equivalent** if they result in the same value for all possible values of the involved variables. The properties of algebra describe the rules for how to create equivalent expressions.

Elementary Properties of Algebra.

For any expressions x, y, and z, the following expressions are equivalent.

- **Additive identity** (zero): $x + 0 = x$.
- **Multiplicative identity** (one): $1 \cdot x = x$.
- For every value x, there is an **additive inverse** value written $-x$ so that $x + -x = 0$. It is always the case that $-x = -1 \cdot x$.
- For every non-zero value x ($x \neq 0$), there is a **multiplicative inverse** value written $\div x$ so that $x \cdot \div x = 1$. When $x \neq 0$, we have $\div x = \frac{1}{x}$, which is why the inverse of x is also called the **reciprocal**.
- **Commutative properties**: $x + y = y + x$ and $x \cdot y = y \cdot x$.
- **Associative properties**: $(x+y)+z = x+(y+z)$ and $(x \cdot y) \cdot z = (x \cdot y) \cdot z$.
- **Distributive property** of multiplication over addition: $x \cdot (y + z) = x \cdot y + x \cdot z$.

Example 1.5.4 Our pick-a-number example with the dependent variable

$$y = \frac{2(x+5)-4}{2} - x$$

could be used as a mind-reader trick. If a volunteer from the audience chooses their own number for x and then does the math correctly, you (the mentalist) can always guess the final number y. Use algebra properties to determine what you should predict.

Solution. We start with the distributive property to rewrite $2(x+5) = 2x+10$. Subtracting 4 is equivalent to adding -4. We can then use the associative property to rewrite $(2x+10)+-4 = 2x+(10+-4) = 2x+6$. Dividing by two is equivalent to multiplying by $\frac{1}{2}$. This allows us to use the distributive property again,

$$\begin{aligned}\frac{2(x+5)-4}{2} = \frac{2x+6}{2} &= \frac{1}{2}(2x+6) = \frac{1}{2}(2x)+\frac{1}{2}(6)\\ &= (\frac{1}{2}\cdot 2)x + \frac{6}{2} = x+3\end{aligned}$$

The second line used the associative property and the multiplicative identity. Can you see where? The final step in the calculation is to subtract x,

$$y = \frac{2(x+5)-4}{2} - x = (x+3) - x = (x+-x)+3 = 3.$$

(What properties were used there?)

In conclusion, we have discovered $y = 3$. No matter what number is originally chosen, the final result of the described calculation will always be three. □

When finding equivalent expressions, we really are just using these basic rules. We can add zero by adding an expression and its additive inverse. We can multiply by one by multiplying by a nonzero expression and its multiplicative inverse, $1 = u/u$. The associative and commutative laws together allow us to reorder terms and operations *with respect to a single operation type.* A common mistake is to reorder terms or operations across different operations. For example, $2+3y$ might be incorrectly written as $5y$, which would be incorrectly using the idea of associativity, $2+(3\cdot y) = (2+3)\cdot y$ (false).

The distributive law is used to products of sums to sums of products and back. The reverse operation is usually called **factoring**. Adding fractions with common denominators is really about having a common factor. Because division is really multiplication by reciprocals, canceling common factors is an application of multiplicative inverses.

Example 1.5.5 Rewrite $\dfrac{3}{x} + x$ as a single fraction.

Solution. The first term $\frac{3}{x}$ has an inverse factor $\div x$. To combine terms as a fraction, this needs to be a common factor in both terms. We take x and multiply it by the inverse factors $x \div x$ and then factor:

$$3 \div x + x = 3 \div x + x\, x \div x = (3+x^2) \div x.$$

However, we usually do this using fraction notation:

$$\frac{3}{x} + x = \frac{3}{x} + \frac{x\cdot x}{x} = \frac{3+x^2}{x}.$$

□

Example 1.5.6 Simplify $\dfrac{3x^2y - 6xy^2}{9x^2y^2}$ by canceling common factors.

Solution. A common mistake would be to cancel the x^2 in the first term and y^2 in the second term. Terms do not cancel over addition; they cancel in multiplication. We need to rewrite the numerator as multiplication rather than addition (subtraction). The distributive law allows us to identify common factors:

$$\frac{3x^2y - 6xy^2}{9x^2y^2} = \frac{3xy(x - 2y)}{9x^2y^2}.$$

By recognizing $9x^2y^2 = (3xy)(3xy)$, we can rewrite our fraction and cancel the common factors:

$$\frac{3x^2y - 6xy^2}{9x^2y^2} = \frac{3xy(x - 2y)}{(3xy)(3xy)} = \frac{x - 2y}{3xy}.$$

This is the simplest way to rewrite as a fraction. We could also distribute the division to get a simplified sum:

$$\frac{x - 2y}{3xy} = \frac{x}{3xy} - \frac{2y}{3xy} = \frac{1}{3y} - \frac{2}{3x}.$$

□

1.5.3 Equations

An **equation** is a logical statement that two expressions are equal. As a logical statement, an equation can be **true** or **false**.

Example 1.5.7 $1 + 1 = 3$ is an equation. The two expressions are $1 + 1$ and 3. This equation is a false statement because the values of the expressions are different. □

When an equation involves variables, the truth of the statement depends on the values of the variables. If we specify particular values for each variable, then we calculate the exact numerical value of each expression and then test if the values are equal. For some values of the variables, the equation may be false; for other values, the equation will be true. A **solution** to the equation is a set of values for the variables in the equation that makes the statement true. The **solution set** of an equation is the set of all possible solutions. If the equation is true for all possible values of the variables, the equation is called an **identity**.

Example 1.5.8 $x + 1 = 3$ is an equation with a variable x. When $x = 1$, or when x represents the value 1, the equation is the same as our earlier example. In that case, the equation is false. However, when $x = 2$, the equation corresponds to $2 + 1 = 3$, which is true. The value 2 is a solution and is in the solution set. □

Example 1.5.9 The statement $2x + 5 = 13$ is an equation involving a single variable, x. We can test the equation using different values for x. For $x = 1$, the expression $2x + 5$ has a value $2(1) + 5 = 7$. Since $7 \neq 13$, the equation is false and $x = 1$ is not a solution. For $x = 4$, the expression $2x + 5$ has a value $2(4) + 5 = 13$. We see that $x = 1$ is a solution because the expressions $2x + 5$ and 13 have the same value. The value 1 is in the solution set. □

In the examples above, we took possible numbers and tested if they were solutions. Testing values to find solutions is impractical because there are infinitely many different values possible for each variable. Finding a solution by guessing would be a stroke of luck. If we did find a solution, we might use

our intuition to say that we found all of the solutions. But how do we *know*? What if our intuition is wrong? Finding one solution does not tell you whether there might be more solutions.

Instead, we use algebra to find solutions by **solving** the equation. You would have learned many strategies for solving equations in an earlier algebra class. Rather than attempt to address every strategy, we will focus on the overarching principles.

Most of these strategies rely on a principle of finding **equivalent** equations. Equations are equivalent when they true or false for exactly the same values of variables. The symbol $\Leftrightarrow$ is used to say that two logical statements are equivalent.

You may have learned that an equation is like a balance or scale. The two expressions are like two masses being balanced against one another. The equation is true if the masses are in balance. We create an equivalent equation if we apply the same operation to both sides of the equation, so long as the operation is invertible.

Balanced Operations Result in Equivalent Equations.

The following operations can be used to create equivalent equations, where each variable represents arbitrary expressions.

- Balanced Addition: $a = b$ is equivalent to $a + c = b + c$.
- Balanced Subtraction: $a = b$ is equivalent to $a - c = b - c$.
- Balanced Multiplication: $a = b$ is equivalent to $a \cdot c = b \cdot c$, so long as $c \neq 0$.
- Balanced Division: $a = b$ is equivalent to $\frac{a}{c} = \frac{b}{c}$, so long as $c \neq 0$.

Because multiplication and division include a condition $c \neq 0$, the new equation might have extra solutions corresponding to values where $c = 0$ that are not solutions to the original equation. These **extraneous** should not be confused with actual solutions.

In addition to the balanced arithmetic operations, we will later learn about invertible or one-to-one functions. An invertible function can be applied to both sides of an equation to create an equivalent equation, so long as the expressions have values in the function domain. Noninvertible functions potentially introduce extraneous solutions.

The primary strategy for solving an equation is to create an equivalent equation where the variable is isolated. If a variable appears only once in an equation, then our strategy would be to apply balanced operations until one side of the equation only has that variable. Generally, we can use the **inverse operation** for the last operation in the expression based on the order of operations. If we think about the operations involved in an expression as wrapping layers around the variable, then applying inverse operations would be like unwrapping the variable one layer at a time.

Example 1.5.10 Consider the earlier equation $2x + 5 = 13$. Use balanced operations to solve the equation.

Solution. The variable x only appears in the expression $2x+5$. Because order of operations applies multiplication before addition, the operation of addition $+5$ would be the last operation. The **inverse operation** is to add -5, which

we do in a balanced way.

$$2x + 5 = 13 \quad \Leftrightarrow \quad 2x + 5 + -5 = 13 + -5 \quad \Leftrightarrow \quad 2x = 8$$

The last operation in the expression $2x$ is now multiplication by 2. The next balanced operation is to multiply by the inverse $\div 2 = \frac{1}{2}$.

$$2x + 5 = 13 \quad \Leftrightarrow \quad 2x = 8 \quad \Leftrightarrow \quad \frac{1}{2} \cdot 2x = \frac{1}{2} \cdot 8 \quad \Leftrightarrow \quad x = 4$$

The equation $x = 4$ has isolated the variable, so the only solution is $x = 4$. The solution set $\{x : 2x + 5 = 13\}$—the set of values x that make $2x + 5 = 13$ true—has a single value $\{x : 2x + 5 = 13\} = \{4\}$. □

If an equation has the variable appearing in multiple locations, we generally have two strategies to consider. One strategy—isolating a variable—is to find an equivalent equation where the variable only appears once. To do this, we use balanced operations to put terms with the variable on the same side of the equation. We then use algebra properties, if possible, to solve for that variable.

Example 1.5.11 Solve the equation $\dfrac{3x}{x+2} = 2$.

Solution. The equation has the variable x appear twice. For the expression on the left to be defined, we know $x+2 \neq 0$. We can use balanced multiplication and multiply both sides of the equation by $x+2$ to find an equivalent equation

$$3x = 2(x + 2).$$

(This is also called cross-multiplication.) The right expression can be rewritten to obtain

$$3x = 2x + 4,$$

which can be solved as

$$x = 4.$$

We check our answer by testing the truth of the original equation. With $x = 4$, our equation is

$$\frac{3(4)}{4+2} = 2.$$

Because $3(4) = 12$ and $4 + 2 = 6$ and $12 \div 6 = 2$, the equation is true. The solution set is

$$\{x : \frac{3x}{x+2} = 2\} = \{4\}.$$

□

The other common strategy—factoring—is to find an equivalent equation with one expression exactly zero and the other expression is factored. The factoring strategy is based on the properties of zero in relation to multiplication. When *non-zero* numbers are multiplied, the product is also non-zero. The only way a product can equal zero is if one of the factors is zero.

Theorem 1.5.12 Product Equals Zero. *Given any expressions A and B, the equation $A \cdot B = 0$ is equivalent to the compound statement $A = 0$ or $B = 0$.*

Consequently, when an equation is written as a product equalling zero, we can identify all solutions for each factor individually equal to zero. The solution set will then be the **union** of the solutions of these separate equations.

Example 1.5.13 Solve the equation $x^3 = 4x$.

Solution. Because the variable x appears as a cube x^3 and alone as x, isolating the variable will not be a successful strategy. We use factoring instead, which requires moving all terms to one side. The balanced operation would be to add $-4x$ to both sides,

$$x^3 - 4x = 0.$$

Now that we have an equivalent equation written as an expression equal to zero, we need to factor our expression. The expression has a common factor x in all terms, so we write

$$x(x^2 - 4) = 0.$$

The factoring principle tells us that solutions satisfy either $x = 0$ or $x^2 - 4 = 0$. We can continue to factor:

$$x(x + 2)(x - 2) = 0.$$

Now, each factor might equal zero leading to a different solution: $x = 0$, $x = -2$, or $x = 2$. Because these are the only values that make a factor equal zero, they are the only solutions. The solution set is the union of the three values,

$$\{x : x^3 - 4x\} = \{-2, 0, 2\}.$$

□

In the next section, we will explore the strategy of factoring in more depth in the context of solving polynomial equations. We will also review using the quadratic formula to solve equations. The next example reminds you to be careful about what you think will be equivalent equations.

Example 1.5.14 Solve the equation

$$\frac{x}{x-3} = \frac{3x-4}{x-3}.$$

Solution. A common strategy for this equation that two fractions are equal is to cross-multiply. That is, multiply the x in the numerator on the left by the $x - 3$ in the denominator on the right, and then multiply the $3x - 4$ in the numerator on the right by the $x - 3$ in the denominator on the left. Then we can use the factoring method.

$$\begin{aligned} x(x-3) &= (3x-4)(x-3) \\ x^2 - 3x &= 3x^2 - 9x - 4x + 12 \\ x^2 - 3x &= 3x^2 - 13x + 12 \\ 0 &= 2x^2 - 10x + 12 \\ 2(x^2 - 5x + 6) &= 0 \\ 2(x-2)(x-3) &= 0 \end{aligned}$$

This final equation is factored. The equation $2 = 0$ has *no solution*. The equations $x - 2 = 0$ and $x - 3 = 0$ have solutions $x = 2$ and $x = 3$, respectively. However, because the denominators were the same, $x - 3$, our solution $x = 3$ was actually an extraneous solution.

Multiplying an equation involving fractions by an expression involving x always risks introducing extraneous solutions, particularly if it changes the domain of the expressions. Factoring is always preferable and only slightly more challenging. To use factoring, we find an equivalent equation by adding

expressions to get zero on one side,

$$\frac{x}{x-3} - \frac{3x-4}{x-3} = 0.$$

The common denominator is a common inverse factor, allowing us to combine the fractions,

$$\frac{x}{x-3} - \frac{3x-4}{x-3} = \frac{x-(3x-4)}{x-3} = \frac{x-3x+4}{x-3} = \frac{-2x+4}{x-3}.$$

Consequently, our equation is equivalent to

$$\frac{-2x+4}{x-3} = 0.$$

The factors are $-2x+4$ and the multiplicative inverse of $x-3$, which can never equal zero. The only solution is the solution to $-2x+4=0$ or $x=2$. (A quotient equals zero only if the numerator equals zero and the denominator is non-zero.) □

Finally, if an equation is equivalent to an equation that is always false, then the equation has no solutions. The solution set is the empty set, $\varnothing = \{\}$.

Example 1.5.15 Find the solution set for the equation $\dfrac{c}{c+3} = 1$.

Solution. When we cross-multiply the equation by the expression $c+3$ (assuming $c \neq -3$), we get an equation

$$c = c+3$$

which is equivalent to

$$0 = 3.$$

Both of these equivalent equations are never true. There are no solutions to the original equation. The solution set is the empty set $\varnothing$. □

1.5.4 Systems of Equations

We have just discussed solving an equation for a single variable. Equations might involve multiple variables. Such an equation establishes a relation between the variables. Solutions will require that the value of one variable depends on the values of any other variables.

Example 1.5.16 The equation

$$u^2 + v^2 = 16 + 6u$$

forms a relation between variables u and v.

- Find the possible values for v when $u = -1$.
- Find the possible values for u when $v = 2$.
- Find the possible values for u and v when $u = v$.

Solution. First, to solve the equation when $u=-1$, we substitute the value of $u=-1$ and then use algebra to isolate v.

$$u^2 + v^2 = 16 + 6u$$
$$(-1)^2 + v^2 = 16 + 6(-1)$$

$$1 + v^2 = 10$$
$$v^2 = 9$$
$$v = \pm 3$$

There are two values for v when $u = -1$. The solutions are the states $(u, v) = (-1, 3)$ and $(u, v) = (-1, -3)$.

Next, to solve the equation when $v = 2$, we substitute the value $v = 2$. However, because u appears in the equation with terms u^2 and $6u$, we can not combine terms to isolate u. Instead, we need to use the (((Unresolved xref, reference "thm-quadratic-formula"; check spelling or use "provisional" attribute)))quadratic formula.

$$u^2 + v^2 = 16 + 6u$$
$$u^2 + (2)^2 = 16 + 6u$$
$$u^2 + 4 = 16 + 6u$$
$$u^2 - 6u - 12 = 0$$
$$u = \frac{6 \pm \sqrt{(-6)^2 - 4(-12)}}{2}$$
$$u = \frac{6 \pm \sqrt{84}}{2} = \frac{6 \pm 2\sqrt{21}}{2}$$
$$u = 3 \pm \sqrt{21}$$

Again, two states are solutions, $(u, v) = (3 + \sqrt{21}, 2)$ and $(u, v) = (3 - \sqrt{21}, 2)$.

Finally, the equation $u = v$ is a constraint involving both variables. Because v is shown as a dependent variable in the constraint $v = u$, we substitute u in place of v in the original equation.

$$u^2 + v^2 = 16 + 6u$$
$$u^2 + (u)^2 = 16 + 6u$$
$$2u^2 - 6u - 16 = 0$$
$$u^2 - 3u - 8 = 0$$
$$u = \frac{3 \pm \sqrt{(-3)^2 - 4(1)(-8)}}{2}$$
$$u = \frac{3 \pm \sqrt{9 + 32}}{2} = \frac{3 \pm \sqrt{41}}{2}$$

For each value of u, we have $v = u$. One solution would be $(u, v) = (\frac{3+\sqrt{41}}{2}, \frac{3+\sqrt{41}}{2})$ while the other solution would be $(u, v) = (\frac{3-\sqrt{41}}{2}, \frac{3-\sqrt{41}}{2})$. □

When working with multiple variables, we often have multiple equations. For example, when we created equations for the parameters of a parametrized model from data, we created a different equation involving the parameters for each data point. A solution for one equation is a state giving values for each of the variables such that the equation is true. A solution for the *system* of equations is a state that makes all of the equations true.

A useful strategy for solving equations is to isolate a dependent variable in one equation and then substitute the resulting value or formula into the other equation. That equation then has one variable which can be solved using the usual methods.

Example 1.5.17 Find a model of the form $y = ax + bx^2$ that passes through the points $(x, y) = (1, 2)$ and $(x, y) = (2, 5)$.

Solution. Using the data provides us with an equation for a and b for each point:

$$\begin{aligned} (x, y) = (1, 2) \quad &\Rightarrow \quad 2 = a(1) + b(1)^2 \\ (x, y) = (2, 5) \quad &\Rightarrow \quad 5 = a(2) + b(2)^2 \end{aligned}$$

These equations form a system of equations that must be satisfied simultaneously,

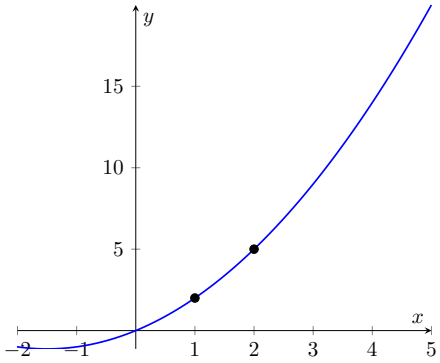
$$\begin{cases} a + b = 2, \\ 2a + 4b = 5. \end{cases}$$

This sets up the mathematical problem that we will solve.

We begin by solving one of the equations for one of the variables. If we take the equation $a + b = 2$ and solve for b, we obtain $b = 2 - a$. We can now substitute the expression $2 - a$ in place of b in the *other* equation, and then solve for a:

$$\begin{aligned} 2a + 4(2 - a) &= 5 \\ 2a + 8 - 4a &= 5 \\ -2a + 8 &= 5 \\ -2a &= -3 \\ a &= \frac{3}{2} \end{aligned}$$

Knowing that $a = \frac{3}{2}$ and that $b = 2 - a$, we find $b = \frac{1}{2}$. The model passing through the data is therefore $y = \frac{3}{2}x + \frac{1}{2}x^2$. A graph showing this solution is shown below.

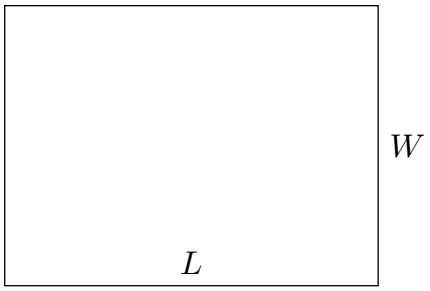


□

Systems of equations allow us to answer questions involving variables that are subject to multiple constraints. A constraint, which provides information about how the variables need to be related, is represented by an equation. Solving the system of equations finds the values that satisfy all constraints.

Example 1.5.18 Is it possible to enclose an area of 25 m^2 in a rectangle with perimeter of 18 m? If so, how?

Solution. In this problem, we need to identify the relevant variables for the system and the equations that constrain the state. We are working with a rectangle, which is characterized by a length and a width. Let us draw a figure and use variables L for the length and W for the width.



The perimeter P and the area A can be considered to be dependent variables, defined by the equations

$$P = 2L + 2W,$$
$$A = L \cdot W.$$

The problem gives us two additional pieces of information, $P = 18$ and $A = 25$. When we substitute those values of the state into the equations, we have two equations for two variables:

$$2L + 2W = 18, \qquad L \cdot W = 25.$$

In order to solve these equations, we use one equation to isolate one of the variables, say L, and then substitute the resulting expression into the other equation.

$$\begin{aligned} 2L + 2W = 18 \quad &\Rightarrow \quad L = 9 - W \\ L \cdot W = 25 \quad &\Rightarrow \quad (9 - W)W = 25 \end{aligned}$$

Then we solve the equation that only involves W.

$$\begin{aligned} (9 - W)W &= 25 \\ 9W - W^2 &= 25 \\ W^2 - 9W + 25 &= 0 \\ W &= \frac{9 \pm \sqrt{(-9)^2 - 4(25)}}{2} \\ W &= \frac{9 \pm \sqrt{81 - 100}}{2} = \frac{9 \pm \sqrt{-19}}{2} \end{aligned}$$

When solving this quadratic formula, we have the square-root of a negative number giving complex numbers.

In conclusion, we found that there are no real solutions. This means that it is not possible to create a rectangle with a perimeter of 18 m and an area of 25 m^2. □

1.5.5 Equivalence and Graphs

It is useful to think about how graphs relate to equivalent expressions, equivalent equations, and equivalent systems of equations. Understanding the interpretation relating to graphs should help us understand the concepts.

Because two expressions are equivalent if and only if they produce the identical values for any choice of the variables, the graphs of equivalent expressions should look identical. When there is only one variable involved, say x, then a simple graph suffices. Suppose u_1 and u_2 are the two expressions. In a graphing utility, if we graph the two expressions, $y = u_1$ and $y = u_2$, then the graphs

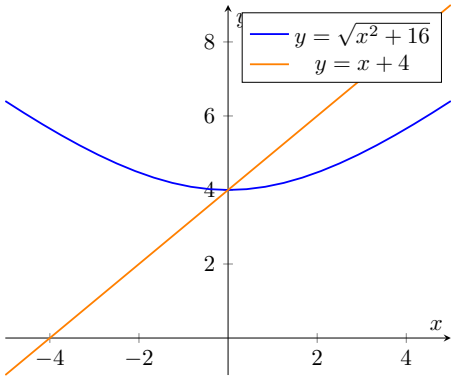
should exactly overlap. It can be difficult to check if an overlap is exact or only approximate.

We could instead create a single graph that subtracts one expression from the other, $y = u_1 - u_2$. When the expressions are equivalent, the graph will show $y = 0$ for all values x. However, because computers only approximately represent numbers, computer arithmetic can introduce small errors. Consequently, we should not be surprised to see a graph with small fluctuations.

Example 1.5.19 Use a graph to test whether the following expressions are equivalent.

1. $\sqrt{x^2+16}$ and $x+4$
2. $\dfrac{\sqrt{x+1}-1}{x}$ and $\dfrac{1}{\sqrt{x+1}+1}$

Solution. The first comparison between $\sqrt{x^2+16}$ and $x+4$ is checked by graph $y = \sqrt{x^2+16}$ and $y = x+4$ on the same figure.

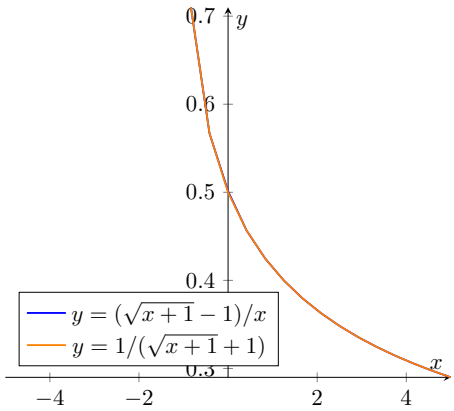


The graphs clearly are different. This is definitive evidence that the expressions are not equivalent. Even identifying just one point, like $x = 3$, and showing the formulas give different results proves that the expressions are not equivalent:

$$\begin{aligned} x = 3 \quad &\Rightarrow \quad \sqrt{x^2+16} = \sqrt{9+16} = \sqrt{25} = 5, \\ x = 3 \quad &\Rightarrow \quad x+4 = 3+4 = 7. \end{aligned}$$

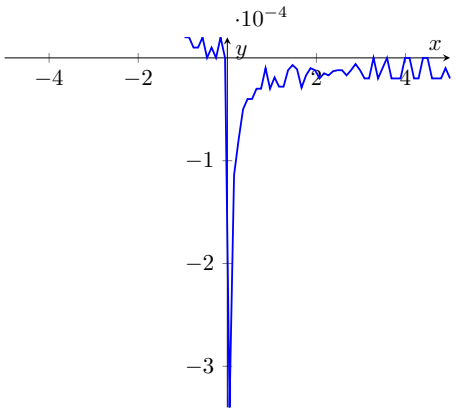
The value used provides a **counterexample** to the claim of equivalence.

The second claim is that $\dfrac{\sqrt{x+1}-1}{x}$ and $\dfrac{1}{\sqrt{x+1}+1}$ are equivalent. When we graph these two expressions, we only see one curve in the figure.



Is this because the graphs are the same? Or is it because one of the graphs is so different that it just doesn't appear in the window? To avoid that uncertainty, we can instead graph their difference,

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



What do we see? Something interesting seems to be happening at $x = 0$. Notice that the first expression is undefined for $x = 0$ (can't divide by zero). The other expression has a value $\frac{1}{2}$. So the two expressions are not really equivalent because of this one point.

What about the other points? The rest of the graph has very small values that appear to fluctuate. We must consider the possibility that this is due to computer error. Let us try some values for which the square root will be simple.

$$\begin{aligned} x = 3 \quad &\Rightarrow \quad \frac{\sqrt{x+1}-1}{x} = \frac{\sqrt{4}-1}{3} = \frac{1}{3} \\ x = 3 \quad &\Rightarrow \quad \frac{1}{\sqrt{x+1}+1} = \frac{1}{\sqrt{4}+1} = \frac{1}{3} \end{aligned}$$

That's a match.

$$\begin{aligned} x = 8 \quad &\Rightarrow \quad \frac{\sqrt{x+1}-1}{x} = \frac{\sqrt{9}-1}{8} = \frac{2}{8} \\ x = 8 \quad &\Rightarrow \quad \frac{1}{\sqrt{x+1}+1} = \frac{1}{\sqrt{9}+1} = \frac{1}{4} \end{aligned}$$

Again, it's a match. We start to think that the expressions probably are equivalent. □

Graphical evidence that expressions are equivalent can give us confidence but do not provide definitive evidence. Ultimately, that needs to come from algebraic arguments. In our example, $\frac{\sqrt{x+1}-1}{x}$ and $\frac{1}{\sqrt{x+1}+1}$ appear to be equivalent for $x \neq 0$. Let us use algebra to simplify the difference between the expressions.

Example 1.5.20 Show that $\frac{\sqrt{x+1}-1}{x}$ and $\frac{1}{\sqrt{x+1}+1}$ are equivalent for $x \neq 0$.

Solution. We will simplify the difference by finding a common denominator. Notice how we have to use the distributive property as we FOIL out the

product.

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

This proves that the expressions are equivalent for $x \neq 0$. □

The graphical interpretation of equivalent equations is not the same as equivalent expressions. That should make sense because expressions and equations are not the same type of objects. Recall that equations are equivalent if they have the same solution sets. Consequently, we need to understand the graphical meaning of solution sets.

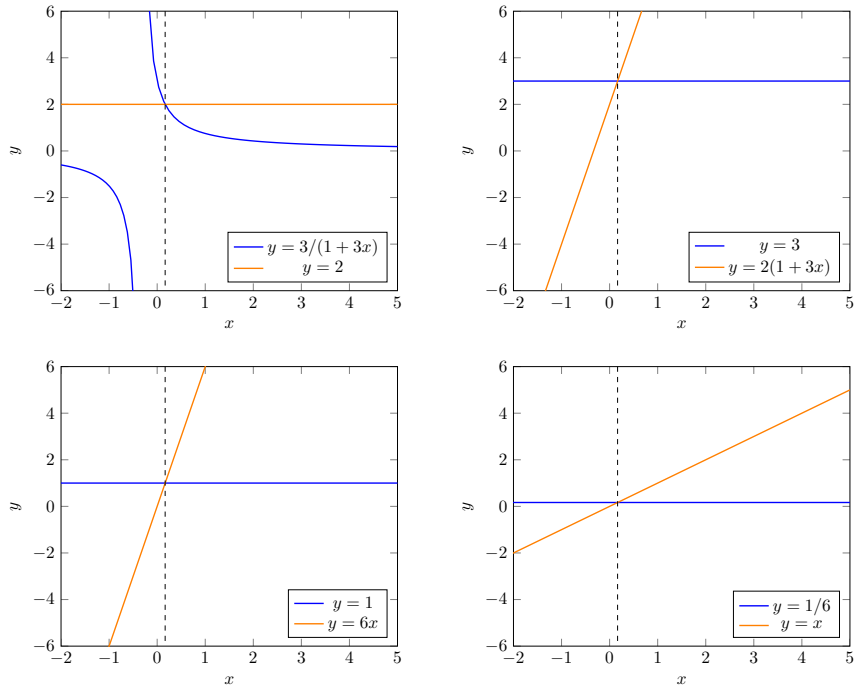
An equation is a statement that two expressions have the same value. We can graph the value of an expression in terms of its independent variable. Since an equation has two expressions, we consider two different graphs. A solution to an equation is a value for the variable where the expressions have the same value. This corresponds to a point where the graphs intersect. Consequently, two equations are equivalent if the values of the variable where the expressions agree are the same.

Example 1.5.21 Solve the equation $\dfrac{3}{1+3x}=2$ using equivalent equations. For each stage of the solution, graph the expressions in the equations to illustrate the equivalence.

Solution. To solve the equation, we will multiply both sides by $1+3x$ and then isolate the variable x. This gives us the following sequence of equivalent equations.

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

For each equation, we should see the expression intersect exactly at $x=\frac{1}{6}$. We graph the two expressions for each equation and the vertical line at $x=\frac{1}{6}$. Each figure shows the graphs intersect at the same x-value. The y-values for the intersection points are changing because the expressions are changing.



□

When we visualized one equation as a graph of two expressions, we introduced a new variable y so that the graph could be shown in the (x, y) plane. The value of y at the point of intersection is not necessarily the same when we change to an equivalent equation. However, when we work with systems of equations, all of the variables are essential. An equivalent system of equations needs solutions to keep the same values for all variables at the point of intersection.

Example 1.5.22 Solve the system of equations

$$\begin{cases} x + 4y = 6 \\ 3x - y = 5 \end{cases}$$

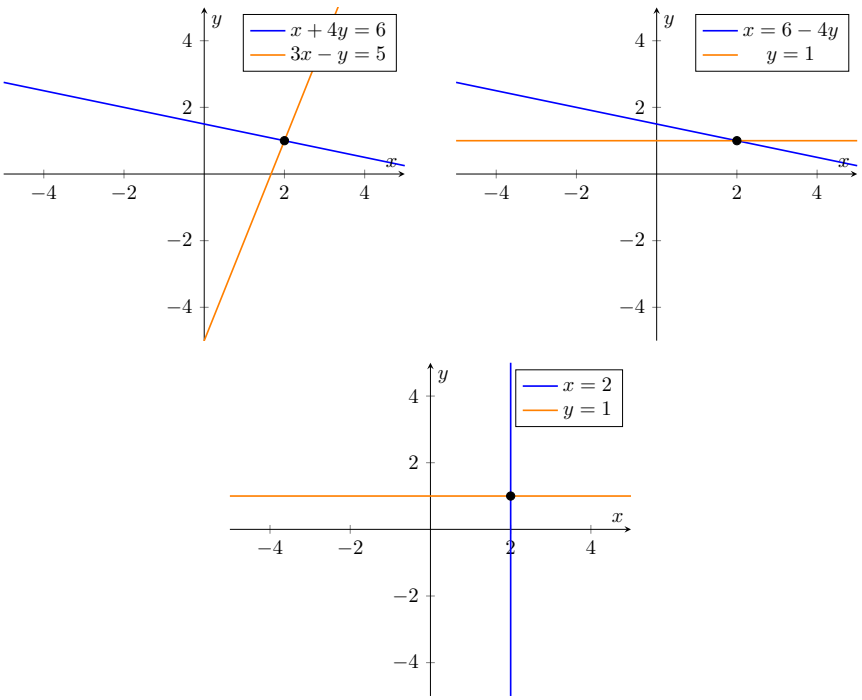
and graph the equivalent systems at each stage.

Solution. Solving the first equation for x, we get $x = 6 - 4y$. When we substitute this expression into the second equation, we have a new system.

$$\begin{cases} x = 6 - 4y \\ 3(6 - 4y) - y = 5 \end{cases} \quad \Leftrightarrow \quad \begin{cases} x = 6 - 4y \\ 18 - 13y = 5 \end{cases}$$

The second equation in this system simplifies to $-13y = -13$ or $y = 1$. When we substitute that value back into the first equation, we get $x = 6 - 4(1)$ or $x = 2$. The solution to the system is $(x, y) = (2, 1)$.

When graphing the equations in the equivalent systems, we note that each equation is the graph of a line. We can graph the lines quickly by plotting their intercepts found by setting $x = 0$ or $y = 0$ and solving for the other value. The equation $x + 4y = 6$ has a y-intercept $(x = 0)$ at $y = \frac{3}{2}$ and an x-intercept $(y = 0)$ at $x = 6$. The equation $3x - y = 5$ has a y-intercept $(x = 0)$ at $y = -5$ and an x-intercept $(y = 0)$ at $x = \frac{5}{3}$. In our second system, the equation $y = 1$ is a horizontal line because x can have any value. In the final system (the solution), the equation $x = 2$ corresponds to a vertical line where y can have any value.



□

1.5.6 Summary

- An expression is any value or a formula that represents a value. Expressions are equivalent if they have the same value for all possible assignments of the variables. Simplifying an expression is to find an equivalent expression in a form that meets an established convention.
-
- An equation is a logical statement that two expressions are equal. A solution to an equation is a state (values specified for all variables) that makes the equation true. The solution set is the set of all possible solutions. Equations are equivalent if they have exactly the same solution sets.
- The primary method for solving an equation is to find an equivalent equation that isolates the variable.
- A key fact about arithmetic is that the only way a product can equal zero is if one factor is zero. This fact is used to solve equations written as an expression equal to zero by factoring.
- A system of equations has solutions described by states with all variables having values that make every equation true. Equivalent systems of equations have the same solutions.

1.5.7 Exercises

1. Show using the elementary properties of addition and multiplication why $2(x+3)-1 = 2x+5$.

2. Show using the elementary properties of addition and multiplication why $(x+3)(x-1) = x^2+2x-3$.

3. A student made a mistake writing $\dfrac{3x+1}{x} = \dfrac{3+1}{1} = 4$. What did the student do? Why was it incorrect?

4. A student made a mistake writing $x \cdot \dfrac{2x+1}{x+3} = \dfrac{2x^2+x}{x^2+3x}$. What did the student do? Why was it incorrect?

Rewrite each of the following expressions as an equivalent sum instead of as a product.

5. $4(x+3)$

6. $3(x-2)(x+4)$

7. $(x-1)(x-2)(x-3)$

Rewrite each of the following expressions as an equivalent factored expression.

8. $3x-15$

9. $4x^2-6x$

Without solving the equation, which of the following values are in the solution set?

10. $x^2-2x = x+4$

(a) $x=-2$

(b) $x=-1$

(c) $x=0$

(d) $x=1$

(e) $x=2$

11. $z^3-5z = 2z^2-6$

(a) $z=-2$

(b) $z=-1$

(c) $z=0$

(d) $z=1$

(e) $z=2$

Find the solution set for each equation.

12. $2(x+5)-3=7$

13. $3t+5 = t-2$

14. $\dfrac{4u}{u+5} = 3$

15. $\dfrac{4u}{u+5} = 4$

16. $\dfrac{2y}{y-1} = 2$

17. $\dfrac{2y}{y-1} = 3$

18. $(2x-3)(x+2)=0$

19. $t(5t-1)(3-t)=0$

20. $\dfrac{p(p+1)}{p+2} = 0$

Find the solutions to the system of equations.

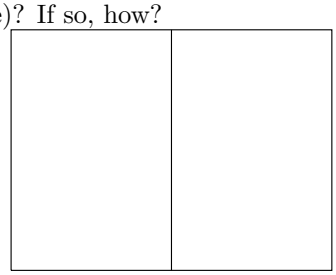
21. $3x + 2y = 15$ and $y = 3$.

22. $2x - 5y = 7$ and $x + 2y = 9$.

23. $x^2 - y = 4x$ and $2x + y = 3$.

24. Is it possible to enclose an area of 25 m^2 using a rectangle with perimeter of 25 m? If so, how?

25. Is it possible to enclose an area of 50 m^2 in two congruent rectangles that share an edge such that the total length of edges is 40 m (counting the shared edge only once)? If so, how?



1.6 Factoring

1.6.1 Overview

Factoring provides a powerful algebraic tools to analyze where an expression equals zero. In the previous section, we noted that a product of two factors can only equal zero if at least one of the factors itself equals zero (see Theorem 1.5.12). More generally, factoring helps us understand the relation between a product and zero, whether it is greater than or less than zero.

In this section, we will review some strategies for factoring expressions. We introduce some technological approaches to factoring. Then we review some strategies that you would have seen in an algebra course. We also use our knowledge of how factoring to develop meaningful models.

1.6.2 Factoring with Technology

Factoring is a process that is best accomplished using technology. Every technique that we learn by hand can be accomplished much more quickly and more reliably by a computer. Once the computer has been programmed correctly, it doesn't commit the arithmetic errors that we are prone to make. In a practical setting, except for simple problems, you will be better off obtaining factors from a computer algebra system.

A computer algebra system (CAS) is a computer program that is designed to apply the rules of algebra according to the user's request. The popular website WolframAlpha (`https://www.wolframalpha.com`) allows you to ask mathematical questions using natural language. It interprets your request and shows a variety of mathematical responses that might answer your question. For example, to factor a polynomial like $x^3 - 7x + 6$, we would submit a request `factor x^3-7x+6`. WolframAlpha would give a response that the result is $(x-1)(x-2)(x+3)$.

WolframAlpha can perform many other basic computations. It is built on the same CAS as a stand alone application called Mathematica, which is sold by Wolfram. A web-based system like WolframAlpha has a disadvantage that we can not create a chain of dependent calculation. Systems like Mathematica that use scripts or notebooks, on the other hand, do allow for interrelated calculations. Popular commercially available CAS programs include Wolfram's Mathematica and MapleSoft's Maple. Many college campuses have license agreements with one of these programs.

A free and open-source alternative CAS is SageMath (`http://www.sagemath.org/`). While you can download and use this program on your own computer, you can also access and use its capabilities through web-access. Similar to using WolframAlpha to use the power of Mathematica, you can use the power of SageMath in what are known as SageCells. A SageCell can be accessed at `https://sagecell.sagemath.org`. The online version of this text also has live SageCells embedded as interactive demonstrations.

In SageMath, we can create mathematical objects (like expressions or equations) and then perform actions on those objects. The creation and naming of a mathematical object occurs through an *assignment*. For example, to create the expression x^3-7x+6 and assign it to a name `expr1`, we type the command `expr1 = x^3-7*x+6` on its own line. Notice how we must explicity state that there is multiplication between `7` and x. To create a new expression that is the factored form and name it `expr2`, we perform a new assignment where the value is based on the factoring action applied to `expr1`. The relevant command would be `expr2 = expr1.factor()`.

The SageCell script below illustrates the commands working together. There are also two commands that show us a nicely formatted version of our expressions for comparison. Be sure to try this script. Push the `Evaluate` button in the live online cell, or copy this into a clean SageCell and evaluate it there. You should see the results:

$$x^3 - 7x + 6$$
$$(x+3)(x-1)(x-2)$$

```
expr1 = x^3-7*x+6
expr2 = expr1.factor()
show(expr1)
show(expr2)
```

Using named expressions is useful when we have additional actions to do later. For simple problems like this, we can actually skip naming the expressions. Try changing the script to the following command and re-evaluate: `show( factor(x^3-7*x+6) )`. What happens if you don't include the command `show`?

The reverse process of multiplying out a factored expression also is frequently needed. It is also tedious to do by hand and more reliable using a computer. The relevant command is to `expand` the expression. Suppose we want to know what $(x+1)(x+2)(x+3)(x+4)$ is as a polynomial in standard form. Using the SageCell script below reveals the answer to be

$$(x+1)(x+2)(x+3)(x+4) = x^4 + 10x^3 + 35x^2 + 50x + 24.$$

Again, notice how the CAS requires that we show explicitly where each multiplication occurs. What happens if there aren't parentheses? What do you think might be happening?

```
expr1 = (x+1)*(x+2)*(x+3)*(x+4)
expr2 = expr1.expand()
show(expr1)
show(expr2)
```

1.6.3 Strategies for Factoring by Hand

Although technology makes factoring fast and simple, we should be prepared to perform simple factoring by hand. We review some basic strategies for factoring that you would have learned in an algebra class.

It helps to remember that factoring is the reverse process of the distributive property of multiplication over addition. That is, when we expand $a\,(b+c) = ab + ac$, we see that ab and ac have the common factor of a from distribution. If we can identify a common factor, then we can reverse the process and write our expression as a multiplication over addition of terms.

Example 1.6.1 Factor $4x^2 + 6x^3$.

Solution. Recalling that $4 = 2 \cdot 2$ and $6 = 2 \cdot 3$, we recognize that the terms $4x^2$ and $6x^3$ have a common factor of $2x^2$:

$$4x^2 + 6x^3 = 2x^2 \cdot 2 + 2x^2 \cdot 3x.$$

Factoring this out, we have

$$4x^2 + 6x^3 = 2x^2(2 + 3x).$$

□

More advanced factoring approaches often are built on top of this idea. For example, a method known as **factoring by grouping** arises by expressing a sum of terms as a sum of two groups of terms that can be found to have a common factor. Some cubic polynomials (but not most) can be factored using this approach.

Example 1.6.2 Use factoring by grouping to factor $x^3 - 3x^2 - 4x + 12$.

Solution. The strategy is to group the x^3 and x^2 terms together and to group the x and constant terms together,

$$x^3 - 3x^2 - 4x + 12 = (x^3 - 3x^2) + (-4x + 12),$$

and then factor out common factors. The first group $x^3 - 3x^2$ has a common factor of x^2 to give

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

The second group $-4x + 12$ has a common factor of 4 to give

$$-4x + 12 = 4(-x + 3).$$

To get a common factor, we should recognize that we should have used a common factor of -4:

$$-4x + 12 = -4(x - 3).$$

We now have groups with a common factor:

$$\begin{aligned} x^3 - 3x^2 - 4x + 12 &= (x^3 - 3x^2) + (-4x + 12) \\ &= x^2(x - 3) + -4(x - 3) \\ &= (x^2 - 4)(x - 3) \end{aligned}$$

A full factorization would also factor $x^2 - 4 = (x + 2)(x - 2)$ to give

$$x^3 - 3x^2 - 4x + 12 = (x + 2)(x - 2)(x - 3).$$

□

In many cases, a mathematical solution to a problem is easier to find when we anticipate what it should look like. We use this concept to guide us in factoring quadratic polynomials. Quadratic polynomials result from expanding a product of the form $(ax+b)(cx+d)$. That expansion is often described using the acronym *FOIL* (First-Outside-Inside-Last):

$$(ax + b)(cx + d) = ac\,x^2 + ad\,x + bc\,x + bd = ac\,x^2 + (ad + bc)x + bd.$$

Notice that in the middle expression, we have four terms, similar to what we had with cubic polynomials. This means that we might be able to factor if we can find a clever way to do grouping.

If we want to factor a quadratic expression $Ax^2 + Bx + C$, then we are looking for values for a, b, c, d so that

$$Ax^2 + Bx + C = (ax + b)(cx + d).$$

This requires that $A = ac$, $C = bd$, and $B = ad + bc$. A clever observation is that $AC = (ac)(bd) = (ad)(bc)$, so that we are writing B as a sum of factors of AC. This will be how we create our grouping.

Example 1.6.3 Factor $2x^2 - x - 6$.

Solution. We begin by recognizing the coefficients $A = 2$, $B = -1$, and

$C = -6$. We want to write $B = -1$ as a sum of factors of $AC = -12$. Our strategy is to think through all of the simple factors of -12 and see if any pair of factors add to -1. What are the factors of -12?

$$-12 = (-1)(12) = (1)(-12) = (-2)(6) = (2)(-6) = (-3)(4) = (3)(-4)$$

When we check the sum of those pairs, we find:

$$\begin{aligned} -1 + 12 &= 11 \\ 1 + -12 &= -11 \\ -2 + 6 &= 4 \\ 2 + -6 &= -4 \\ -3 + 4 &= 1 \\ 3 + -4 &= -1 \end{aligned}$$

In practice, we would likely add the factor pairs in our head rather than write them down.

Once we find the pair of factors with the correct sum, $3 + -4 = -1$, we expand the term $-x$ as a sum $3x - 4x$ to rewrite the quadratic as a sum of four terms that can now be grouped.

$$\begin{aligned} 2x^2 - x - 6 &= 2x^2 + 3x - 4x - 6 \\ &= (2x^2 + 3x) + (-4x - 6) \\ &= x(2x + 3) + -2(2x + 3) \\ &= (x - 2)(2x + 3) \end{aligned}$$

We thus have the factors

$$2x^2 - x - 6 = (x - 2)(2x + 3).$$

□

When using the method of grouping for quadratics, be sure that you consider both positive and negative factors as pairs.

Another example of anticipating the form of a solution occurs when we know a polynomial's root. Knowing that a polynomial has a root means that we also know a factor.

Theorem 1.6.4 Root–Factor Theorem. *Suppose $p(x)$ is a polynomial of degree n for which $x = c$ is a root, $p(c) = 0$. Then $p(x)$ can be written in a factored form*

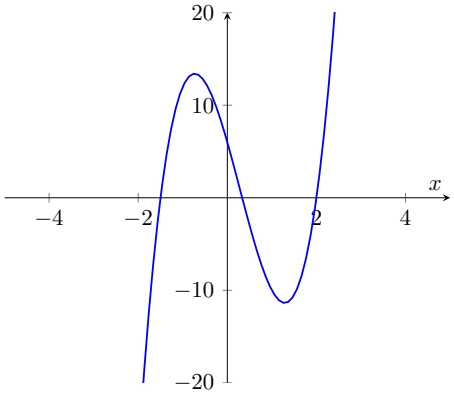
$$p(x) = (x - c) \cdot q(x)$$

where $q(x)$ is a polynomial of degree $n - 1$.

If we can find a root to a polynomial, then we know a simple factor. One way to find the root is by looking at its graph. Knowing the original polynomial and a factor, we can work out the other polynomial factor. Finding that other factor corresponds to polynomial division.

Example 1.6.5 Factor the polynomial $6x^3 - 5x^2 - 17x + 6$.

Solution. When we graph the polynomial, we can identify possible roots.



The graph suggests that there might be a root at $x = 2$, between -1 and -2, and between 0 and 1. We verify that $x = 2$ is a root by substituting that value into the formula:

$$\begin{aligned} 6x^3 - 5x^2 - 17x + 6 &= 6(2)^3 - 5(2)^2 - 17(2) + 6 \\ &= 6(8) - 5(4) - 17(2) + 6 \\ &= 48 - 20 - 34 + 6 \\ &= 0 \end{aligned}$$

The factor that corresponds to $x = 2$ as a factor is $x - 2$ (because $x = 2$ is equivalent to $x - 2 = 0$). We now know that there is a polynomial $q(x) = ax^2 + bx + c$ with degree 2 so that

$$6x^3 - 5x^2 - 17x + 6 = (x - 2) \cdot (ax^2 + bx + c).$$

We find $q(x) = ax^2 + bx + c$ by polynomial division. In effect, however, we are multiplying by this polynomial with unknown coefficients and determining the coefficient values so that the product equals the original polynomial.

$$\begin{aligned} (x - 2) \cdot (ax^2 + bx + c) &= x \cdot (ax^2 + bx + c) - 2(ax^2 + bx + c) \\ &= ax^3 + bx^2 + cx - 2ax^2 - 2bx - 2c \\ &= ax^3 + (b - 2a)x^2 + (c - 2b)x - 2c \end{aligned}$$

Because this product must equal $6x^3 - 5x^2 - 17x + 6$, we have a system of equations based on matching coefficients:

$$\begin{cases} a = 6 \\ b - 2a = -5 \\ c - 2b = -17 \\ -2c = 6 \end{cases} \Leftrightarrow \begin{cases} a = 6 \\ b = 2a - 5 \\ c = 2b - 17 \\ -2c = 6 \end{cases}$$

Substituting $a = 6$ into the second equation gives $b = 2(6) - 5 = 7$. Substituting $b = 7$ into the third equation gives $c = 2(7) - 17 = -3$. This matches the fourth equation solved for c. We can therefore write the factorization using the values of the coefficients for $q(x)$.

$$6x^3 - 5x^2 - 17x + 6 = (x - 2)(6x^2 + 7x - 3).$$

We finish the problem by factoring the new quadratic factor. The product $ac = 6(-3) = -18$ has factors $-18 = (-2)(9)$ that sum to $-2 + 9 = 7$. We can rewrite and group the quadratic to have common factors:

$$6x^2 + 7x - 3 = 6x^2 + -2x + 9x - 3$$

$$\begin{aligned} &= (6x^2 - 2x) + (9x - 3) \\ &= 2x(3x - 1) + 3(3x - 1) \\ &= (2x + 3)(3x - 1) \end{aligned}$$

Consequently, our final factorization can be written

$$6x^3 - 5x^2 - 17x + 6 = (x - 2)(2x + 3)(3x - 1)$$

The other roots can be found from the factors: $x = -\frac{3}{2}$ (from $2x + 3 = 0$) and $x = \frac{1}{3}$ (from $3x - 1 = 0$). □

1.6.4 Factors for Polynomial Modeling

Polynomials have easily understood behavior based on their factors. The roots of the factors are exactly the roots of the polynomial. Furthermore, these roots are the only possible locations where the polynomial might change sign. It is easy to show that each simple factor changes sign exactly at its root by solving an inequality directly. Then, knowing the signs of each factor allows us to determine the sign of the polynomial as a whole, based on the following theorem.

Theorem 1.6.6 *The relation of product of expressions $w = u_1 \cdot u_2 \cdots u_n$ with zero is based the relations of the factors:*

- *The product $w = 0$ if and only if at least one $u_k = 0$.*
- *The product $w > 0$ if and only if all $u_k \neq 0$ and there are an even number of $u_k < 0$.*
- *The product $w < 0$ if and only if all $u_k \neq 0$ and there are an odd number of $u_k < 0$.*

We can analyze the behavior of a polynomial, in terms of its relation to zero, using the factors. This process is called **sign analysis**.

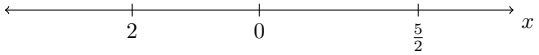
1. Find the factored version of the polynomial.
2. Identify the roots of all of the factors and order them on the number line.
3. The roots divide the number line into a collection of intervals. On each interval between roots, count the number of factors that will be negative.
4. On each interval, if the number of negative factors is even, then the polynomial will be greater than zero. If the number of negative factors is odd, then the polynomial will be less than zero.

Sometimes, a polynomial has repeated roots that appear as a factor raised to a power. When counting factors, the power is used as multiplicity of repetition.

Example 1.6.7 Perform sign analysis of the polynomial

$$p(x) = x^2(x + 2)(2x - 5).$$

Solution. The polynomial has four factors—x (double), $x + 2$, and $2x - 5$. The roots of these factors are $x = 0$, $x = -2$, and $x = \frac{5}{2}$. We order these roots graphically on a number line.

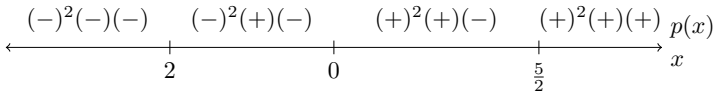


Once the roots are ordered on the number line, we can see the intervals of interest. Intervals represent continuous segments of the number line and are described as a range from left to right. There are four intervals: $(-\infty, -2)$,

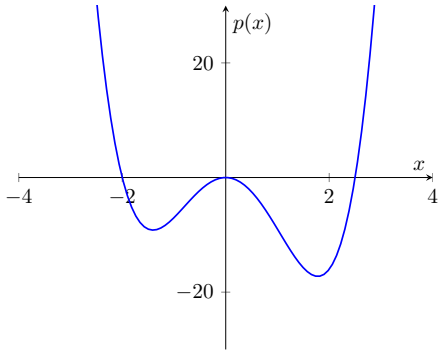
$(-2, 0)$, $(0, \frac{5}{2})$, and $(\frac{5}{2}, \infty)$. On each interval, we can determine the sign of each factor, illustrated in the table below. Then, when we multiply these factors together, we find the overall sign of the polynomial. The polynomial is negative when there is an odd number of negative factors and positive when there is an even number of negative factors.

Interval	x^2	$x+2$	$2x-5$	$p(x)$
$(-\infty, -2)$	$(-)^2$	$-$	$-$	$+$
$(-2, 0)$	$(-)^2$	$+$	$-$	$-$
$(0, \frac{5}{2})$	$(+)^2$	$+$	$-$	$-$
$(\frac{5}{2}, \infty)$	$(+)^2$	$+$	$+$	$+$

Rather than create a table of signs, we can label the signs of the factors directly on the number line above each interval.



The table or the number allows us to interpret the relation of the polynomial with zero. The polynomial is positive (greater than zero) on the intervals $(-\infty, -2)$ and $(\frac{5}{2}, \infty)$ and negative (less than zero) on the intervals $(-2, 0)$ and $(0, \frac{5}{2})$. When we look at a graph of the polynomial, we can see that the graph is above the axis on the outer intervals and below the axis on the inner intervals.



□

Performing sign analysis on a polynomial might seem like a silly exercise. After all, with our graphing calculators and computers, it is easy enough to graph the polynomial directly and look where the graph is above or below the axis. It is in the reverse process that we start to see the power.

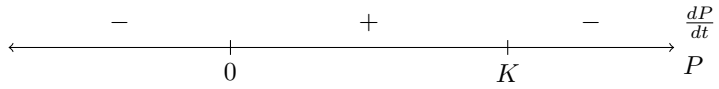
Suppose that we want to explore a mathematical model for a phenomenon where we know how a quantity relates with zero on the number line. However, we would like to have a simple mathematical formula that captures that relation. By constructing a polynomial with factors that will match the sign analysis, we can use that polymomial as our model.

Example 1.6.8 A simple model for density-dependent population growth has a population's growth rate as positive when the population is between zero and a carrying capacity and negative when the population is above the carrying capacity. Create a simple polynomial model that will capture this behavior.

Solution. We will let P be our symbol to represent the value of the population. In biology, the carrying capacity is most often symbolized by the symbol K. The symbol for a quantity's growth rate is called the **derivative** with respect to time and has the symbol $\frac{dP}{dt}$. (The goal of calculus is to understand what this derivative really means.) We want to create a model that describes

how the growth rate $\frac{dP}{dt}$ depends on the population P such that this formula is positive when P is in the interval $(0, K)$ and negative when P is in (K, ∞).

We can start by creating a number line summary of the behavior we want. Because $P < 0$ is not physically relevant, the sign used on $(-\infty, 0)$ doesn't matter. However, since polynomials change sign at roots unless a factor has an even power, we will choose to make $\frac{dP}{dt}$ negative when $P < 0$.



The roots help us know our basic factors. A root at $P = 0$ corresponds to a factor of P. A root at $P = K$ corresponds to a factor of $P - K$. The product $P(P - K)$, however, would have the opposite signs on the intervals. This is corrected by multiplying by a third constant factor, $-a$, where a itself is some positive constant. This gives us our basic model,

$$\frac{dP}{dt} = -aP(P - K).$$

If we multiply the negative sign by the factor $P - K$, we obtain an equivalent model

$$\frac{dP}{dt} = aP(K - P).$$

The constants a and K become model parameters.

In biology, slightly different model parameters are more commonly used. We rewrite our factor $K - P = K - \frac{PK}{K}$ and then factor out the common factor K We now have a model

$$\frac{dP}{dt} = aK\, P(1 - \frac{P}{K}).$$

The product aK can be replaced by another parameter, r, to obtain

$$\frac{dP}{dt} = r\, P(1 - \frac{P}{K}).$$

The parameter r is called the intrinsic per capita growth rate. This model is known as the **logistic growth rate model** for density-dependent growth. □

In our model, we used a factor $1 - \frac{P}{K}$ instead of the factor $K - P$. These two factors have the same roots. One advantage to a factor like $1 - \frac{P}{K}$ is that the units of P and K cancel so that the factor is dimensionless. Polynomial models are often written with factors of the form $\frac{x}{a} - 1$ or $1 - \frac{x}{a}$ where $x = a$ is the desired root. In particular, a linear model with known x- and y-intercepts is most easily modeled using factors and does not require finding the slope.

Example 1.6.9 Suppose S is the number of seeds a plant produces and D is the density of competing plants around it. When there is no competition, $D = 0$, the plant can produce its highest output, $S = M$ seeds. When the competition reach a critical level, $D = D_M$, the plant no longer produces seed, $S = 0$. Develop a linear model for how S relates to D for $0 \le D \le D_M$.

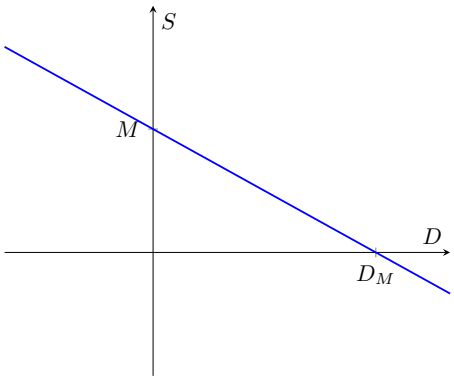
Solution. A linear model has a single root, which in this modeling scenario is $S = 0$ at $D = D_M$. This means we have a factor $D - D_M$ or $\dfrac{D}{D_M} - 1$. We need another constant factor, say A to account for the vertical scale, such as

the other intercept. Our model has the form

$$S = A\left(\frac{D}{D_M} - 1\right).$$

When we substitute $D = 0$, we find $S = -A = M$ so that the parameter $A = -M$. The model can then be written in terms of the data provided,

$$S = -M\left(\frac{D}{D_M} - 1\right) = M\left(1 - \frac{D}{D_M}\right).$$



□

1.6.5 Summary

1. In practical applications, technology is the most efficient method to factor expressions. Computer Algebra Systems (CAS) like SageMath provide tools to perform operations on mathematical objects.

2. The first priority in factoring involves common factors. Grouping terms sometimes allows us to find common factors among the groups.

3. Quadratic polynomials with integer coefficients $Ax^2 + Bx + C$ can be factored by grouping if you can find factors of AC that add to the value of B. This is accomplished by rewriting the term Bx as two terms using the factors and then grouping terms.

4. Knowing a root $x = a$ of a polynomial tells us that $x - a$ is a factor. The other factor can be found by polynomial division.

5. The factored form of an expression allows us to perform sign analysis.

   (a) Identify the roots of all of the factors and order them on the number line.

   (b) The roots divide the number line into a collection of intervals. On each interval between roots, count the number of factors that will be negative.

   (c) On each interval, if the number of negative factors is even, then the expression will be greater than zero. If the number of negative factors is odd, then the expression will be less than zero.

6. Alternatively, knowing the roots and desired results for sign analysis, we can use a factored polynomial to generate a model for the relation between two variables.

1.6.6 Exercises

Use technology to factor the following formulas.

1. $x^3 + x^2 - 5x + 3$

2. $x^4 - 4x^3 - 11x^2 + 30x$

3. $x^4 + 10x^3 + 35x^2 + 50 + 24$

4. $24x^3 + 14x^2 - 11x - 6$

Use the method of grouping to factor each cubic.

5. $x^3 - 2x^2 + 3x - 6$

6. $x^3 - 5x^2 - 3x + 15$

7. $4x^3 - 12x^2 - x + 3$

Factor each quadratic polynomial.

8. $x^2 - 2x - 3$

9. $x^2 - 9x + 20$

10. $x^2 + 4x - 21$

11. $2x^2 + 3x - 2$

12. $2x^2 - 7x - 15$

13. $6x^2 + 11x - 10$

For each problem, verify that the given value is a root of the given polynomial. Use the Root-Factor theorem and polynomial division to factor the polynomial.

14. $x^3 - 6x^2 - x + 30$; $x = -2$

15. $x^3 - 2x^2 - 11x + 12$; $x = 4$

16. $2x^3 - x^2 - 13x - 6$; $x = 3$

For each factored polynomial, complete sign analysis to describe the intervals where the polynomial is greater than zero and where it less than zero. Then compare your results with a graph.

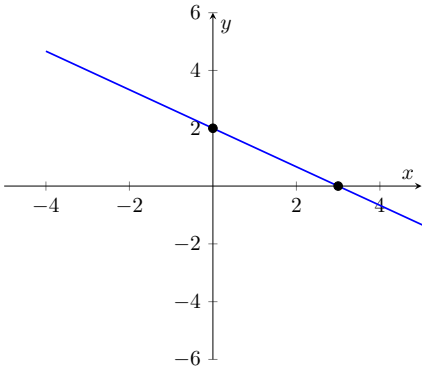
17. $3x(x - 4)$

18. $2x(2 - x)(3 - x)$

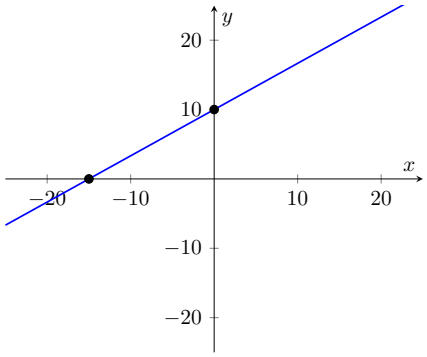
19. $x(x + 2)(x - 1)^2$

20. $(x - 4)(2x - 3)(3x + 1)^3$

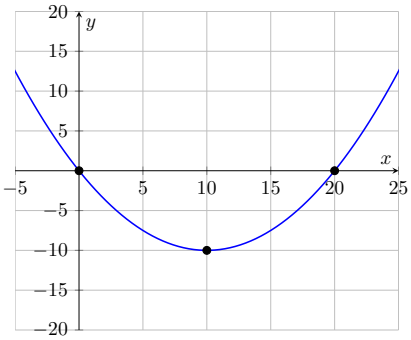
21. Find a linear model for the following graph without finding the slope.



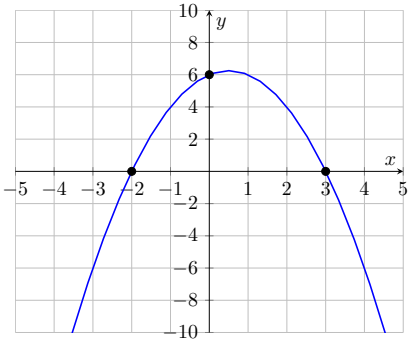
22. Find a linear model for the following graph without finding the slope.



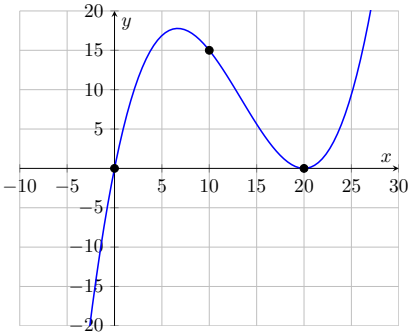
23. Find a polynomial model for the following graph.



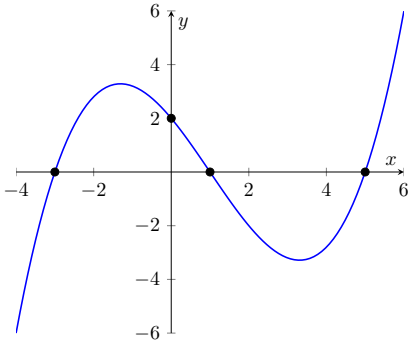
24. Find a polynomial model for the following graph.



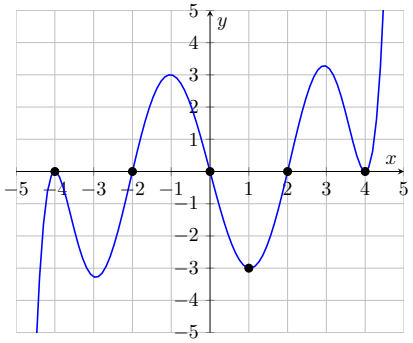
25. Find a polynomial model for the following graph.



26. Find a polynomial model for the following graph.



27. Find a polynomial model for the following graph.



28. In economics, the demand for a product, which measures the number of units a company can sell, is related to the price charged to purchase the product. Let p be the price being charged and let n be the demand or number of units that are sold. A simple model is that the demand has a linear relation with the price. Suppose that if the company gave away product for free, $p = 0$, the demand is $n = 5000$. On the other hand, if the price charged is $p = 50$, there is no demand, $n = 0$.

Create a linear model for the demand as it depends on the price. Then, create a corresponding model for the revenue as it relates to the price, where the revenue is the price charged times the number of units sold.

29. Some populations will die off (negative growth rate) unless the size of the population is above some minimum value. Populations that exhibit this behavior are said to have the Allee effect. Let K be the population carrying capacity and M be the minimum threshold for population growth, and assume that $0 < M < K$.

Create a polynomial model for the population's growth rate such that the growth rate is positive for $M < P < K$ and negative for $0 < P < M$ and $P > K$. The growth rate should be zero at $P = 0$, $P = M$, and $P = K$.

30. Optical tweezers (also called laser tweezers) use a tightly focused laser beam to create a force on small transparent objects. The force depends on the location of the object from the center of the beam. When the object is further from the beam than a particular distance, δ, the force pushes the object away. When the distance is less than δ, the force pulls the object toward the center.

Create a polynomial model for the force F as it depends on the position x of the object. Assume that $x = 0$ is the center of the tweezers. The force will have roots at $x = \pm\delta$ and $x = 0$. A positive force pushes the

object to the right and a negative force pushes it to the left. Include a constant factor k and use dimensionless factors for the roots at $x = \pm\delta$.

1.7 Exponents, Roots, and Logarithms

Overview. We frequently think of simplifying expressions as looking for opportunities to cancel terms. We similarly might think of moving terms to the opposite side of an equation, but somehow the opposite term moves to the other side. Although thinking about inverses as opposites can be useful, it is perhaps difficult to generalize properly. A more productive framework with which to think about working with expressions and equations is in terms of operations. We will later generalize the idea of an operation to the mathematical concept of a function. In that sense, inverse operations will correspond to inverse functions.

One of the most common areas relating to inverses where novices encounter trouble is in terms of powers, where the inverse operations are roots and logarithms. For example, a common conceptual error is thinking that logarithms and bases cancel in the same way that factors in fractions cancel. A root is the inverse operation of raising a quantity to a given power. A logarithm is the inverse operation of an exponential operation.

In this section, we review the properties of exponents and focus on distinguishing between the power and exponential operations. We discuss the concept of inverse operations and introduce roots and logarithms as inverses to these operations. We learn to apply inverses to simplify expressions and solve equations involving powers and exponentials.

1.7.1 Properties of Exponents

Do you remember why we have powers as a mathematical notation? When the power is a positive integer, it is to represent repeated multiplication. For example, 3^4 means that we multiply four threes together,

$$3^4 = 3 \cdot 3 \cdot 3 \cdot 3.$$

How did we go from this simple notational convenience to be able to interpret negative powers or fractional powers or even irrational powers? We make sense of these more complex ideas by thinking about what properties the notation should satisfy.

We know that when we multiply and divide by the same number, the net effect is equivalent to multiplying by 1. We say that the terms cancel. We should think of these *actions* as **inverse operations**. That is, the action of multiplying a number by 3 and dividing a number by 3 are inverse operations. If you do them in succession (one after the other), the net effect is equivalent to having applied no operation at all,

$$x \cdot 3 \div 3 = x.$$

Extending this idea allows us to simplify repeated factors represented by powers. How would we simplify $\dfrac{3^5}{3^2}$? If we realize that 3^5 in the numerator means that we multiply by five threes and that the 3^2 in the denominator means that we divide by two threes, then we recognize that there are two pairs of inverse operations:

$$\begin{aligned}\frac{3^5}{3^2} &= \frac{3 \cdot 3 \cdot 3 \cdot 3 \cdot 3}{3 \cdot 3} \\ &= 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \div 3 \div 3 \\ &= 3 \cdot 3 \cdot 3 \cdot (3 \div 3) \cdot (3 \div 3)\end{aligned}$$

$$= 3 \cdot 3 \cdot 3$$
$$= 3^3$$

The net effect of dividing by 3^2 is that we removed two of the threes in the product 3^5. This is why we say that division causes powers to subtract, $\frac{3^5}{3^2} = 3^{5-2} = 3^3$.

All of the basic properties of powers are motivated by the idea that an integer power corresponds to repeated multiplication. For each property, can you think of how it would be a consequence of this idea?

Properties of Powers.

- Zero Power: $b^0 = 1$ for $b \neq 0$
- Inverse Power: $b^{-x} = \frac{1}{b^x}$
- Product with Common Base: $b^x\, b^y = b^{x+y}$
- Quotient with Common Base: $\frac{b^x}{b^y} = b^{x-y}$
- Power of a power: $(b^x)^y = b^{xy}$
- Product with Common Exponent: $b^x\, c^x = (bc)^x$
- Quotient with Common Exponent: $\frac{b^x}{c^x} = \left(\frac{b}{c}\right)^x$

We illustrate several additional properties in the context of integer powers. For integer exponents, a power means repeated multiplication (similar to how multiplication by an integer means repeated addition). So $b^3 = b \cdot b \cdot b$. The product properties are just about counting.

Example 1.7.1

$$b^3 \cdot b^2 = (bbb) \cdot (bb) = b^5 = b^{3+2}$$
$$(b^2)^3 = (bb)(bb)(bb) = b^{2 \cdot 3}$$
$$(ab)^3 = (ab)(ab)(ab) = (aaa)(bbb) = a^3 b^3$$

□

The zero power property is necessary for the power of a sum rule to remain consistent. We know that $b^{x+0} = b^x$. But the properties of powers also mean $b^{x+0} = b^x \cdot b^0$. For these to both be true requires $b^x = b^x \cdot b^0$ or that $b^0 = 1$.

The properties of powers relating to products and quotients behave similar to the distributive property of multiplication over addition. This is because multiplication originates as repeated addition, just as powers originate as repeated multiplication. However, addition and powers have no convenient properties. Many mistakes occur when students forget this and imagine that powers distribute over addition like multiplication. (It doesn't!)

Example 1.7.2 To illustrate that $(a+b)^2 \neq a^2 + b^2$, consider the numbers $a = 2$ and $b = 3$. The first expression gives

$$(a+b)^2 = (2+3)^2 = 5^2 = 25$$

while the second expression gives

$$a^2 + b^2 = 2^2 + 3^2 = 4 + 9 = 13.$$

The proper way to expand the first expression is to think of the power as repeated multiplication and apply the distributive property. This is often called the FOIL method:

$$(a+b)^2 = (a+b)(a+b) = a^2 + 2ab + b^2.$$

□

1.7.2 Exponent Operations and Their Inverses

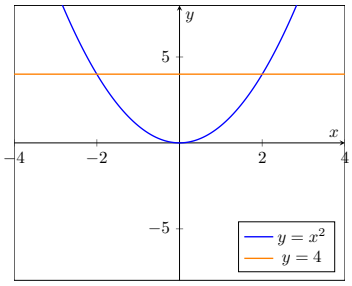
Expressions have a dual interpretation. On the one hand, an expression is a mathematical object that represents some numerical value. On the other hand, an expression also represents a sequence of operations that act on other values. For example, the expression $3x^2 + 5$ is a formula that for each value of x represents a particular value. It also describes a sequence of operations: "Take a value, x. Square it. Multiply by three. Add 5."

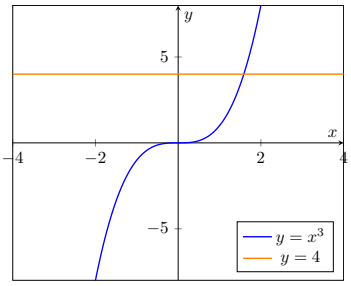
Thinking of powers as operations will require some caution. In the expression, 2^3, the numbers 2 and 3 play different roles. The number 2 is called the **base** and the number 3 is called the exponent. As an action, the expression 2^3, raising 2 to the power 3, can be interpreted in two ways, depending on which number we think of as being acted on. We could say that take the number 2 and apply the power 3. This is likely the more familiar interpretation. Alternatively, we could say that we take the number 3 and apply the base 2. The second interpretation corresponds to the exponential operation.

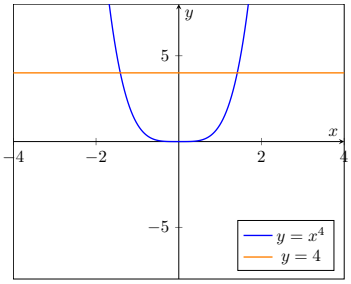
Introducing variables might help make this distinction clearer. This corresponds to thinking of the expression as the value of a function. We will let x represent the number being acted on. The operation of a power 2^3 corresponds to $f(x) = x^3$ with $x = 2$. The same expression interpreted as an exponential operation would be the formula $g(x) = 2^x$ with $x = 3$. An **elementary power function** (applying a power) raises a value to a *fixed power*, $f(x) = x^p$, for a constant p. An **elementary exponential function** (applying a base) uses a value as the exponent of a *fixed base*, $g(x) = b^x$, for a constant b.

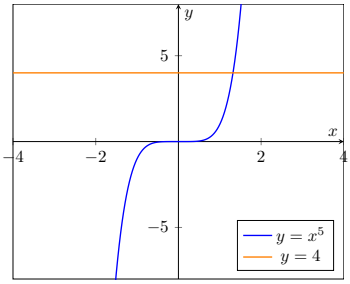
Power and exponential functions have corresponding inverse operations. A root provides the inverse operation to an integer power. A logarithm provides the inverse operation to an exponential. We begin by focusing on powers and roots.

If we wish to solve the equation $x^n = 4$ where n is a positive integer, we might graph $y = x^n$ and $y = 4$ and look for where they intersect.









For even powers n, the equation $x^n = 4$ has two solutions. The graph is symmetric across the y-axis because the product of an even number of negative values is positive, $(-1)^2 = 1$. Notice that if we were trying to solve $x^n = -4$, there would be no solutions when n is even but there would be a single solution when n is odd. The solution is called the nth root.

Definition 1.7.3 For an integer $n > 1$, the nth root $y = \sqrt[n]{x}$ is the value such that $y^n = x$. If n is even, we require $x \geq 0$ and $y \geq 0$. If n is odd, there is no restriction. ◊

A root can be written as a fractional power. The properties of exponents imply that $(x^p)^n = x^{pn}$. If $p = \frac{1}{n}$, then $(x^{\frac{1}{n}})^n = x$. That is, $x^{\frac{1}{n}} = \sqrt[n]{x}$. This equivalence means that for any rational number $p = \frac{k}{n}$, the power x^p can be computed using integer powers (repeated multiplication) and extracting roots:

$$x^{\frac{k}{n}} = (\sqrt[n]{x})^k.$$

Note 1.7.4 When the base is positive, then the choice of representation in the exponent does not matter. Negative values in the base create complications. For example, we know that a fraction can have multiple representations, like $\frac{1}{3} = \frac{2}{6} = \frac{3}{9}$. Because $(-2)^3 = -8$, we know that $\sqrt[3]{-8} = -2$. This is equivalent to saying $(-8)^{1/3} = -2$. However, $(-8)^{2/6}$ is undefined because the 6th root of -8 is not a real number. On the other hand, it will be true that $(-8)^{3/9} = -2$ as well as for any other equivalent fraction with an odd denominator.

We can simplify expressions that have roots and powers applied consecutively. For example, $\sqrt[3]{a^3}$ takes a value a, and then applies the cubing operation followed by the cube-root operation. Because these are inverses, we recover the original value,

$$\sqrt[3]{a^3} = a.$$

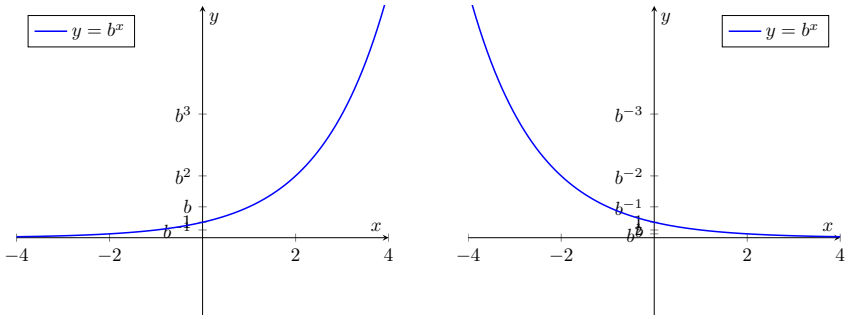
We must be careful, again, when the inverse operations involve even powers. For example, the expression $\sqrt{a^2}$ does not actually simplify to a in all cases. If $a = -2$, then squaring this gives $a^2 = 4$ and then the square root gives $\sqrt{a^2} = \sqrt{4} = 2$. In general, when n is an even power, then $\sqrt[n]{a^n} = |a|$. Applying the power first always makes a^n a positive value, and then the nth root is also defined to return a positive value.

On the other hand, the expression $(\sqrt{x})^2 = x$ rather than $|x|$. In this situation, the first operation is the square root which requires that $x \geq 0$. Squaring the square root of x always recovers the value of x.

How would we define a power with an irrational value? We might approximate the irrational power. That is, we find a rational number that is close to the irrational number and use it instead. This should raise questions. Does our choice of approximation matter? How close do we need to be? Calculus helps us here by introducing the idea of limits. Limits will be central to the ideas of approximation that occur throughout calculus.

Once we know that we can raise any positive base to an arbitrary number as the exponent, we can think of an exponential as a valid operation for a given positive base. Typical graphs of $y = b^x$ for $b > 1$ and for $0 < b < 1$ are shown

below. The special case when $b = 1$ corresponds to a horizontal line and is not shown.



The graphs suggest that for any $y > 0$, the equation $b^x = y$ will have a unique solution x for each value y. The solution is called the logarithm of y for the base b.

Definition 1.7.5 For any base b with $b > 0$ and $b \neq 1$, the logarithm with base b of a value $x > 0$ is written $\log_b(x)$. The value $y = \log_b(x)$ is defined for $x > 0$ as that value y such that $b^y = x$. ◊

Notice that both roots and logarithms are defined through the equation that they solve. We can interpret them as operations that will cancel their inverse operations. That is, the consecutive operations of an exponential and a logarithm with the same base cancel one another. In a similar way, a power and its corresponding root cancel one another, although we will have to be careful with even powers because of the even symmetry.

Example 1.7.6 Solve $\sqrt[3]{x} = 2$.

Solution. The equation $\sqrt[3]{x} = 2$ has an isolated cube root on the left. The inverse operation of the cube root is cubing. Starting with a value x, finding its cube root, and then cubing the result just gets back to x. We use this inverse operation in a balanced way to solve the equation.

$$\begin{aligned} \sqrt[3]{x} &= 2 \\ (\sqrt[3]{x})^3 &= 2^3 \\ x &= 2^3 \end{aligned}$$

The solution is $x = 8$. □

Example 1.7.7 Solve $x^4 = 4$.

Solution. The equation $x^4 = 4$ has an isolated integer power on the left. The inverse operation is a fourth root. Because the power is even, there are two solutions.

$$\begin{aligned} x^4 &= 4 \\ \sqrt[4]{x^4} &= \sqrt[4]{4} \\ x &= \pm\sqrt[4]{4} \end{aligned}$$

Because $4 = 2^2$, we could rewrite this as

$$x = \pm(2^2)^{\frac{1}{4}} = \pm 2^{\frac{2}{4}} = \pm 2^{\frac{1}{2}} = \pm\sqrt{2}.$$

□

Example 1.7.8 Solve $\log_3(x) = 2$.

Solution. The equation $\log_3(x) = 2$ has an isolated logarithm. The inverse operation is an exponential with the same base $b = 3$. We use the function representation of this operation, $\exp_3(x) = 3^x$. An equivalent equation is formed by applying this exponential to both sides of the equation.

$$\log_3(x) = 2$$
$$\exp_3(\log_3(x)) = \exp_3(2) = 3^2$$
$$x = 9$$

This is saying that the equation $\log_3(9) = 2$ is equivalent to $3^2 = 9$. Notice that we could have just written down the inverse equation immediately, since that is how the logarithm is defined. □

Example 1.7.9 Solve $\log_4(x) = 3$.

Solution. The equation $\log_4(x) = 3$ is defined by the inverse equation $4^3 = x$. So $x = 64$. □

Example 1.7.10 Solve $4^x = 8$.

Solution. The equation $4^x = 8$ has an isolated exponential. The inverse operation is a logarithm with the same base $b = 4$.

$$4^x = 8$$
$$\log_4(\exp_4(x)) = \log_4(8)$$
$$x = \log_4(8)$$

To go further on this problem, we need more properties.

For this particular problem, we can proceed if we recognize that both 4 and 8 are powers of 2. Because $4 = 2^2$ and $8 = 2^3$, we can rewrite our equation as

$$4^x = 8 \quad \Leftrightarrow \quad 2^{2x} = 2^3.$$

This means that $2x = 3$ or $x = \frac{3}{2}$. We can verify this, since $4^{3/2} = (\sqrt{4})^3 = 2^3 = 8$. That is, $\log_4(8) = \frac{3}{2}$. □

The previous example illustrated a way that we can simplify a logarithm when the base and input are both powers of the same value. In that example, we had $\log_4(8)$ and the base 4 and input 8 were powers of 2. We used the equivalence of the equations

$$x = \log_4(8) \quad \Leftrightarrow \quad 4^x = 8$$

and then rewrote that equation in terms of the common base b to find x.

Example 1.7.11 Simplify $\log_9(\frac{1}{27})$.

Solution. We start by assigning this value to the variable x so that we have an equation,

$$x = \log_9(\frac{1}{27}).$$

The equivalent equation using an exponential instead of a logarithm is

$$9^x = \frac{1}{27}.$$

We recognize that $9 = 3^2$ and $27 = 3^3$ so that the equation can be rewritten

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

This means that $2x = -3$ or $x = -frac32$. That is, $\log_9(\frac{1}{27}) = -\frac{3}{2}$ □

We will learn more general techniques for simplifying logarithms later. Most scientific calculators only have logarithms for base $b = 10$ and for base $b = e$. The logarithm for $b = 10$ is called the common logarithm and appears on a calculator with out a base `log`. The logarithm for $b = e$ is called the natural logarithm and appears on a calculator as `ln`. We will later prove that every logarithm can be found using one of these by the change of base formula

$$\log_b(x) = \frac{\log(x)}{\log(b)} = \frac{\ln(x)}{\ln(b)}.$$

Example 1.7.12 Solve $3^x = 5$.

Solution. The unknown x has an exponential with base $b = 3$ operation acting on it. To isolate the variable, we need to apply the inverse operation to both sides.

$$\begin{aligned} 3^x &= 5 \\ \log_3(3^x) &= \log_3(5) \\ x &= \log_3(5) \end{aligned}$$

We have solved the equation, but we don't have a good sense of what that number might be. We know that $3^1 = 3$ and $3^2 = 9$, so $x = \log_3(5)$ must be somewhere between 1 and 2. The change of base formula allows us to approximate the value on a calculator,

$$x = \log_3(5) = \frac{\ln(5)}{\ln(3)} \approx 1.46497.$$

□

1.7.3 Applications

We encountered equations that require roots and logarithms to solve when we considered exponential and power function models for data. Recall that a **general power function** has the form

$$f(x) = A\,x^p,$$

where A and p are the model parameters. A **general exponential function** has the form

$$f(x) = A\,b^x,$$

where A and b are the model parameters, with $b > 0$ and $b \neq 1$.

Example 1.7.13 A colony of bacteria is observed to cover an area of $2.5\,\text{mm}^2$. Six hours later, the colony has expanded to cover a space of $100\,\text{mm}^2$. Assuming that the bacteria is growing according to an exponential growth model, develop a model for the area of the colony as a function of the time since the first observation. If the population continues to follow this model, at what time will the bacteria colony fill a dish with $6000\,\text{mm}^2$?

Solution. An exponential model relates the area of the colony C (mm^2) as a function of time t (hours) according to the formula

$$C = A\,b^t,$$

where A and b are model parameters to be determined by the data. The first observation, $t = 0$ and $C = 2.5$, corresponds to a parameter equation

$$2.5 = A\,b^0.$$

The second observation, $t = 6$ and $C = 100$, corresponds to a parameter equation

$$100 = A\,b^6.$$

Because $b^0 = 1$, the first equation gives $A = 2.5$. Substituting this into the second equation gives an equivalent equation

$$100 = 2.5\,b^6.$$

We solve the equation for b by isolating the variable. The expression currently has a product with 2.5, so we apply the inverse operation of dividing by 2.5 on both sides,

$$\frac{100}{2.5} = b^6.$$

The expression now has a power operation, and the inverse is a root,

$$b = \left(\frac{100}{2.5}\right)^{1/6} = \sqrt[6]{40} \approx 1.84931.$$

Consequently, our model for the colony area is given by

$$C = 2.5 \cdot 1.84931^t.$$

To answer the final question, we see that t is unknown but $C = 6000$. Substituting this into the model equation, we obtain an equation only involving t,

$$6000 = 2.5 \cdot 1.84931^t.$$

To isolate the variable, we first need to divide by 2.5,

$$\frac{6000}{2.5} = 1.84931^t.$$

Now, the variable has an exponential operation acting on it. The inverse operation is a logarithm with base $b = 1.84931$, so that

$$t = \log_{1.84931}(2400) = \frac{\ln(2400)}{\ln(1.84931)} \approx 12.6595.$$

The model predicts that the bacteria colony will fill the dish after abou 12.66 hours, which is approximately 12 hours and 40 minutes. □

Example 1.7.14 In the 1930s, a Swiss biologist named Max Kleiber observed that the metabolic rate of mammals approximately follows a power law relation with the mass of the animal. Let Q represent the average metabolic rate (in kJ per day) and let M represent the average mass (in kg). A mouse has an average mass $M = 0.021$ kg and an average metabolic rate of $Q = 20.9$ kJ/day. A horse has an average mass $M = 400$ kg and an average metabolic rate of $Q = 32000$ kJ/day. Find a power law model that matches this data. Use the model to predict the metabolic rate for a cat which has an average mass $M = 3$ kg.

Solution. We start with the model equation, $Q = A\,M^p$, with model parameters A and p. Using the data allows us to create a system of equations for our parameters.

$$\begin{aligned}(M, Q) = (0.021, 20.9) \quad &\Rightarrow \quad 20.9 = A\,0.021^p \\ (M, Q) = (400, 32000) \quad &\Rightarrow \quad 32000 = A\,400^p\end{aligned}$$

To solve a system of equations, we solve one equation for one variable. For this problem, we use the first equation to solve for A:

$$20.9 = A\,0.021^p \qquad \Leftrightarrow \qquad A = \frac{20.9}{0.021^p}.$$

We now substitute this expression in place of A into the other equation.

$$32000 = \frac{20.9}{0.021^p}\,400^p$$
$$32000 = 20.9\frac{400^p}{0.021^p} = 20.9\left(\frac{400}{0.021}\right)^p$$

To solve for the remaining unknown, we need to apply inverse operations. The current expression involves multiplication by 20.9, so the inverse operation is division by 20.9.

$$\frac{32000}{20.9} = \left(\frac{400}{0.021}\right)^p$$

Now the expression is an exponential of p with base $b = \frac{400}{0.021}$. The inverse operation is the logarithm,

$$\log_b\left(\frac{32000}{20.9}\right) = p.$$

We use the change of base formula to find the decimal approximation,

$$p = \frac{\ln(32000/20.9)}{\ln(400/0.021)} \approx 0.744187.$$

Knowing p, we go back to find the value for A,

$$A = \frac{20.9}{0.021^p} \approx 370.45.$$

Our approximate power law model is therefore $Q = 370.45 \cdot M^{0.744187}$. Using the average mass of a cat $M = 3$, we predict

$$Q = 370.45 \cdot 3^{0.744187} \approx 839.067.$$

The observed metabolic rate for cats is actually $Q = 546$ kJ/day, which is lower than predicted. This should not disappoint us too much, as Kleiber's law was really based on a regression model for many animals. Some values will be above the model's prediction and some will be below. □

1.7.4 Summary

- Properties of exponents are motivated by the idea that integer powers correspond to repeated multiplication.
- Exponential functions (exponential growth or exponential decay) have the form $f(x) = A \cdot b^x$, with the independent variable in the exponent.
- Power functions have the form $f(x) = A \cdot x^p$, with the independent variable as the base of the power.
- Roots, such as the square root or cube root, are inverse operations for the power operation:

$$x^n = y \qquad \Leftrightarrow \qquad x = \sqrt[n]{y}.$$

When n is an even integer, we require $x \geq 0$ and $y \geq 0$.

- Roots can be written as reciprocal powers:
$$\sqrt[n]{x} = x^{1/n}.$$
- Logarithms are inverse operations for the exponential operation, defined for every base $b > 0$ and $b \neq 1$:
$$b^x = y \qquad \Leftrightarrow \qquad x = \log_b(y).$$
- The change of base rule allows us to find decimal approximations for any base using the common or natural logarithm:
$$\log_b(x) = \frac{\log(x)}{\log(b)} = \frac{\ln(x)}{\ln(b)}.$$

1.7.5 Exercises

Identify a property of exponents to rewrite an equivalent expression. Note that each property can be applied in either direction.

1. 3^{x+2}

2. $(2x)^3$

3. $2^x \cdot 3^x$

4. $\dfrac{x^3}{4^3}$

5. $\dfrac{3^u}{3^4}$

6. 2^{3x}

7. A student made a mistake writing $3 \cdot 2^x = 6^x$. What did the student do? Why was it incorrect?

8. A student made a mistake writing $2^x \cdot 3^y = 6^{x+y}$. What did the student do? Why was it incorrect?

Find an exact value for each root or logarithm. Do not use a calculator.

9. $\sqrt{9a^2}$

10. $\sqrt[3]{-8a^6}$

11. $\log_2(8)$

12. $\log_2(\frac{1}{8})$

13. $\log_3(\frac{1}{81})$

14. $\log_3(1)$

15. $\log_4(2)$

16. $\log_4(32)$

17. $\log_{1/2}(4)$

18. $\log_{1/4}(2)$

19. $\log_8(\frac{1}{2})$

20. $\log_8(16)$

21. $\log_{25}(\frac{1}{125})$

Solve the equations.

22. $x^7 = 4$

23. $3x^3 = 8$

24. $\sqrt[4]{x} = 3$

25. $3\sqrt[3]{2x} = 4$

26. $5^x = 10$

27. $3^{2x} = 4$

28. $\log_4(x) = 2$

29. $\log_3(2x) = 9$

30. $4\log_5(x) = 15$

31. $3x^2 - x^6 = 0$

32. $4 \cdot x^5 = 3$

33. $4 \cdot 5^x = 3$, writing solutions in terms of the natural logarithm `ln`.

34. $2 \cdot 3^{-x} = 3 \cdot 2^x$, writing solutions in terms of the natural logarithm `ln`. Hint: Find an equivalent equation with a single exponential after using properties of exponents.

Applications.

35. Find a power law for y as a function of x that includes data $(x, y) = (1, 5)$ and $(x, y) = (4, 10)$.

36. Find a power law for y as a function of x that includes data $(x, y) = (2, 20)$ and $(x, y) = (4, 5)$.

37. Find a power law for y as a function of x that includes data $(x, y) = (2, 4)$ and $(x, y) = (20, 8)$.

38. Find an exponential law for y as a function of x that includes data $(x, y) = (0, 5)$ and $(x, y) = (4, 15)$.

39. Find an exponential law for y as a function of x that includes data $(x, y) = (0, 5)$ and $(x, y) = (10, 2)$.

40. Find an exponential law for y as a function of x that includes data $(x, y) = (1, 5)$ and $(x, y) = (6, 10)$.

41. The average body mass and life span of mammals have been observed to follow an approximate power law. A mouse has an average body mass of 0.021 kg and a life span of 1.5 years. A horse has an average body mass of 400 kg and a life span of 40 years. Find a power function model for the life span as a function of body mass. Predict the life span of a typical hare, which has a body mass of 3.4 kg.

42. The fraction of carbon in organic matter that is radioactive (carbon-14) decays exponentially from the time of death. At the time of death, the fraction of radiocarbon would be 1.25 parts per trillion. A sample that is 1000 years old has a fraction of radiocarbon measured at 1.1075 parts per trillion. Model the fraction in parts per trillion as an exponential function of time since death. Estimate the age of a sample that has radiocarbon measured at 0.8 parts per trillion.

43. P-32 is a radioactive isotope of phosphorus used in labeling biological molecules. P-32 has a half-life of 14.29 days. Suppose an experiment begins with 10 µg of P-32. Find a parametrized model for the mass (in µg) as a function of time (in days) measured from the start of the experiment, $t \mapsto M$, in order to determine how much P-32 remains after 10 days and after 100 days.

Hint. Create two constraints using $t = 0$ and $t = 14.29$.

44. The isotope of plutonium Pu-239 has a half-life of 24,110 years, which is the time after which half of the mass has decayed. For an initial mass of 1 kg, how much plutonium remains after 100 years?

45. An exponentially growing population that doubles in size every 5 years currently has 1000 individuals.

(a) What will the population be in 4 years?

(b) How long does it take for the population to triple?

1.8 Logarithms and Their Properties

We previously learned about the algebraic properties of exponents. Because logarithms are the inverses of exponential operations, they inherit related algebraic properties. The properties of logarithms allow us to rewrite expressions involving products and exponents in new ways.

In this section, we will introduce the properties of logarithms and how they relate to the properties of exponents. These properties are closely related to the logarithmic scale. We will learn new methods for solving equations involving exponentials using the properties of the logarithm. In addition, we will learn how to rewrite exponential functions in terms of other bases and justify the change of base formula for logarithms.

1.8.1 The Logarithmic Scale

Historically, the logarithm was invented so that calculations involving multiplication and division could be solved using the much simpler operations of addition and subtraction. Addition and multiplication share many of the same properties—commutativity, associativity, and the existence of an identity and inverses. These shared properties suggest that there might be a way to think about multiplication in terms of addition.

The properties of exponents establish the relationship required to make this connection. When two numbers can be written as powers of the same base, say $u = b^x$ and $v = b^y$, then the product uv can be written as a power of that base as $uv = b^{x+y}$. Division $u \div v$ can similarly be written as a power, $u \div v = b^{x-y}$. In this way, multiplication and division of numbers directly corresponds to addition and subtraction of their associated powers for a given base.

The logarithm relates a number and its associated power for a given base. For every number $u > 0$ and base $b > 0$ with $b \neq 1$,

$$u = b^x \qquad \Leftrightarrow \qquad x = \log_b(u).$$

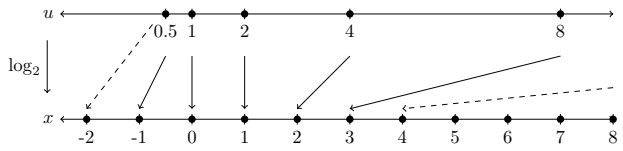
We will try to develop an intuition by creating the map between a number and its logarithm. In the process, we will develop something called a logarithmic number line or logarithmic scale.

A map is a way of thinking about the relationship between two variables. We use one variable as the independent variable, or input, for the map. This is the value we start with. The relation tells us that this value for the independent variable is associated with a particular value of the dependent variable. The value of the dependent variable is the output of the map.

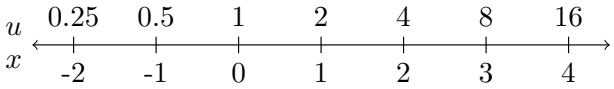
To make our example precise, we will create the map for a base $b = 2$. However, the process could be done for any base $b > 0$ with $b \neq 1$. With a base $b = 2$, we find $\log_2(u)$ by solving the equation $2^x = u$. The logarithm is the map $u \mapsto x$. The equation is easy to solve when u is a known power of 2.

$$\begin{aligned}
1 = 2^0 &\quad \Leftrightarrow (u, x) = (1, 0) \quad \Leftrightarrow \quad \log_2(1) = 0 \\
2 = 2^1 &\quad \Leftrightarrow (u, x) = (2, 1) \quad \Leftrightarrow \quad \log_2(2) = 1 \\
4 = 2^2 &\quad \Leftrightarrow (u, x) = (4, 2) \quad \Leftrightarrow \quad \log_2(4) = 2 \\
8 = 2^3 &\quad \Leftrightarrow (u, x) = (8, 3) \quad \Leftrightarrow \quad \log_2(8) = 3 \\
\tfrac{1}{2} = 2^{-1} &\quad \Leftrightarrow (u, x) = (\tfrac{1}{2}, -1) \quad \Leftrightarrow \quad \log_2(\tfrac{1}{2}) = -1
\end{aligned}$$

These simple logarithms are illustrated in the following map.



Now, instead of drawing two number lines, let us use a single number line labeling the points with the input above the line and the output below the line.



This relation could be extended to *every* point on the number line. The value above the number line is just the value 2^x where x is the value below the number line. Some special values should be explicitly identified. The point with $x = \frac{1}{2}$ corresponds to $u = 2^{1/2} = \sqrt{2} \approx 1.4142$. Similarly, the point with $x = \frac{3}{2}$ corresponds to $u = 2^{3/2} = \sqrt{8} \approx 2.8284$.

The resulting locations of numbers above the number line is called a **logarithmic scale**. A more detailed logarithmic scale is provided below. Because consecutive integers are closer and closer together in a logarithmic scale, the figure only shows the tick mark location for some of the values.

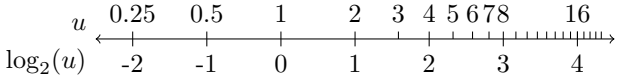


Figure 1.8.1 Logarithmic scale showing the logarithm base two, $b = 2$.

In our example, we developed the logarithmic scale using a base $b = 2$. We could have done it with any base. The logarithmic scale with base $b = 10$, corresponding to the common logarithm, is shown below.

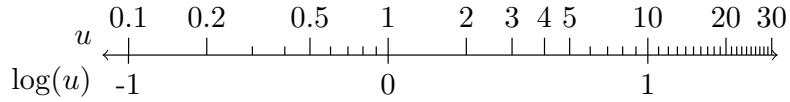


Figure 1.8.2 Logarithmic scale showing the common logarithm, $b = 10$.

Comparing the logarithmic scales for bases $b = 2$ and $b = 10$, you should notice that the logarithmic scale itself appears to be the same. The only thing that is different is the scale of the tick marks below the number line corresponding to the value of the logarithm.

1.8.2 Properties of the Logarithm

When considering the logarithmic scales, we always had $\log_b(1) = 0$. This is because $b^0 = 1$ for every valid base. This creates a mapping from the multiplicative identity $u = 1$ to the additive identity $x = 0$. That is, the value $u = 1$ becomes the origin of the logarithmic scale.

In a similar way, because $b^1 = b$, the logarithm will always have $\log_b(b) = 1$. Be careful that you do not think of this as a cancellation. Instead, think of the logarithmic scale. The value $u = 1$ sets the origin. The logarithm measures the position of each number as if using a ruler starting at $u = 1$. The value $u = b$ sets the length scale for the ruler.

The core properties of logarithms are summarized below.

Theorem 1.8.3 Properties of Logarithms. *For every base $b > 0$ with $b \neq 1$ and values $u > 0$ and $v > 0$, logarithms satisfy the following properties.*

- *Identities Map:* $\log_b(1) = 0$.
- *The Base Sets the Scale:* $\log_b(b) = 1$.
- *Inverse Property:* $b^{\log_b(u)} = u$ *and* $\log_b(b^x) = x$.
- *Product Rule:* $\log_b(u \cdot v) = \log_b(u) + \log_b(v)$.
- *Quotient Rule:* $\log_b\left(\frac{u}{v}\right) = \log_b(u) - \log_b(v)$.
- *Power Rule:* $\log_b(u^p) = p \cdot \log_b(u)$.

Proof. The first two properties were proved in the paragraphs preceding the theorem. For the remaining three properties, because $u > 0$, we know that $b^x = u$ has a solution $x = \log_b(u)$. Similarly, for $v > 0$, we know that there is a value $y = \log_b(v)$ so that $b^y = v$. The inverse property is simply stating what it means to be a logarithm.

The power rule considers the logarithm of the value u^p. By rewriting $u = b^x$, we are looking for the logarithm of $u^p = (b^x)^p$. By the Power of a Power property, we have $u^p = b^{(xp)}$, which means that $\log(u^p) = xp = p \cdot \log_b(u)$.

The product rule considers the logarithm of the value $u \cdot v$. By rewriting $u = b^x$ and $v = b^y$, we are looking for the logarithm of $uv = b^x \cdot b^y$. By the Product with a Common Base property, we have $uv = b^{x+y}$, which means that $\log(uv) = x + y = \log_b(u) + \log_b(v)$. The proof of the quotient rule for logarithms is proved in a similar way. ■

The properties of logarithms allow us to compute the logarithm of a product (and quotient) in terms of the logarithms of the individuals factors.

Example 1.8.4 A reference table shows $\log(2) \approx 0.30103$, $\log(3) \approx 0.47712$ and $\log(5) \approx 0.69897$. Use properties of logarithms to determine $\log(1.2)$.

Solution. Start by writing 1.2 as a product of the factors 2, 3, and 5.

$$1.2 = \frac{12}{10} = \frac{4(3)}{2(5)} = \frac{2(3)}{5}$$

The properties of logarithms allow us to rewrite $\log(1.2)$ as

$$\begin{aligned}
\log(1.2) &= \log\left(\frac{2(3)}{5}\right) \\
&= \log(2(3)) - \log(5) && \text{(Quotient)} \\
&= \log(2) + \log(3) - \log(5) && \text{(Product)} \\
&\approx 0.30103 + 0.47712 - 0.69897 = 0.07918.
\end{aligned}$$

□

Historically, the logarithm was invented so that multiplication and division calculations could be solved using the much simpler operations of addition and subtraction. Engineers and scientists would often reference logarithm tables to find the logarithms of the factors and add the values by hand. Then they would look in the table for the number that had the resulting logarithm. The slide rule was a mechanical implementation, where the lengths corresponding to logarithm values were added physically to get a new length. For an interactive demonstration, check out http://educ.jmu.edu/~waltondb/webapp/log_scale_explore.html.

Although modern calculators and computers have eliminated this particular

need for logarithms in calculations, we still use logarithms in order to expand symbolic formulas expressed as products and powers in terms of sums of the logarithms of the factors. To expand a logarithm, we look at the structure of the expression to which the logarithm is applied. There are likely many parts, but we look at the final operation that is used to form that expression.

1. If the input of the logarithm is a *product* of expressions, then we expand the logarithm by rewriting it as a sum of logarithms each with one of the factors as input.

2. If the input of the logarithm is a *quotient* of expressions, then we expand the logarithm by rewriting it as a difference of logarithms, adding the logarithm of the numerator and subtracting the logarithm of the denominator.

3. If the input of the logarithm is a *power* applied to an expression, then we expand the logarithm by rewriting it as the value of the power times the logarithm of the base.

4. If the input of the logarithm is a *anything else*, then the rules of logarithms do not apply. In particular, we can not expand the logarithm of a sum.

By changing any quotients into multiplication by negative powers, we can reduce the process of expanding logarithms to only two rules: the product and power rules.

Example 1.8.5 Expand $\log\left(\dfrac{4x^3\sqrt{2x+5}}{(x^2+3)^5}\right)$ as far as possible.

Solution. Each factor of the expression inside the logarithm will get its own term using the product and quotient rules for logarithms. If we think of every division in terms of negative powers, then we only need to deal with products. In particular, the inner expression can be rewritten

$$\frac{4x^3\sqrt{2x+5}}{(x^2+3)^5} = 4x^3(2x+5)^{\frac{1}{2}}(x^2+3)^{-5}.$$

The factors identified are 4, x^3, $(2x+5)^{\frac{1}{2}}$, and $(x^2+3)^{-5}$. Using the logarithm's product rule followed by the logarithm's power rule, we find

$$\begin{aligned}\log\left(\frac{4x^3\sqrt{2x+5}}{(x^2+3)^5}\right) &= \log(4)+\log(x^3)+\log\left((2x+5)^{\frac{1}{2}}\right)+\log\left((x^2+3)^{-5}\right)\\ &= \log(4)+3\log(x)+\tfrac{1}{2}\log(2x+5)-5\log(x^2+3).\end{aligned}$$

Notice that we could have used the quotient rule of logarithms instead of negative powers to get the term $-5\log(x^2+3)$. □

The rules for expanding logarithms can also be used in reverse to collect terms into a single logarithm.

Example 1.8.6 Collect the terms of

$$2+4\log x+2\log y-\frac{1}{2}\log(x+y)$$

to write the expression in terms of a single logarithm.

Solution. We first recognize that every logarithm times an expression can

be rewritten as a logarithm with the expression in the exponent.

$$2 + 4\log x + 2\log y - \frac{1}{2}\log(x+y) = 2 + \log(x^4) + \log(y^2) + \log\left((x+y)^{-1/2}\right)$$

Now that the expression is a sum of logarithms. What about the isolated number 2? Using the common logarithm with base $b = 10$, we know $\log(10^2) = 2$.

$$\begin{aligned} 2 + 4\log x + 2\log y - \frac{1}{2}\log(x+y) &= 2 + \log(x^4) + \log(y^2) + \log\left((x+y)^{-1/2}\right) \\ &= \log(10^2\, x^4\, y^2\, (x+y)^{-1/2}) \\ &= \log\left(\frac{100x^4y^2}{\sqrt{x+y}}\right) \end{aligned}$$

□

1.8.3 Using Logarithms to Solve Equations

The properties of logarithms help solve us equations, particularly where the variable is in an exponent. Logarithm and exponential operations are invertible, so we can use them as balanced operations on an equation to obtain new equivalent equations. The properties of the logarithm can then be used to our advantage.

Example 1.8.7 Solve the equation $3 \cdot 2^{3x} = 5$ using the logarithm base 10.

Solution. It is usually best to isolate the exponential term first.

$$\begin{aligned} 3 \cdot 2^{3x} &= 5 \\ 2^{3x} &= \frac{5}{3} \end{aligned}$$

We next apply the logarithm base 10 to both sides of this equation, which then allows us to apply the logarithm power rule on the left. Then we can isolate x.

$$\begin{aligned} \log_{10}(2^{3x}) &= \log_{10}(\tfrac{5}{3}) \\ 3x \cdot \log_{10}(2) &= \log_{10}(\tfrac{5}{3}) \\ x &= \frac{\log_{10}(\frac{5}{3})}{3\log_{10}(2)} \end{aligned}$$

Alternatively, we could have applied the logarithm at the very first. This would require using the logarithm product rule on the left.

$$\begin{aligned} \log_{10}(3 \cdot 2^{3x}) &= \log_{10}(5) \\ \log_{10}(3) + \log_{10}(2^{3x}) &= \log_{10}(5) \\ \log_{10}(2^{3x}) &= \log_{10}(5) - \log_{10}(3) \\ 3x\log_{10}(2) &= \log_{10}(5) - \log_{10}(3) \\ x &= \frac{\log_{10}(5) - \log_{10}(3)}{3\log_{10}(2)} \end{aligned}$$

□

The properties of logarithms allow us to compute logarithms with uncommon bases using logarithms that we know using the change of base formula.

Theorem 1.8.8 *For any two exponential bases b and c,*

$$\log_b(u) = \frac{\log_c(u)}{\log_c(b)}.$$

Proof. Consider $x = \log_b(u)$, which solves $b^x = u$. If we solve this equation using $\log_c$ to both sides, we find the change of base formula.

$$\begin{aligned} b^x &= u \\ \log_c(b^x) &= \log_c(u) \\ x\log_c(b) &= \log_c(u) \\ x = \log_b(u) &= \frac{\log_c(u)}{\log_c(b)}. \end{aligned}$$

■

This theorem has a nice geometric interpretation in relation to the logarithmic scale. The value $\log_b(u)$ measures the distance in the logarithmic scale from the origin at 1 to the location of u in terms of the scale set by the location of b. The change of base formula is based on knowing the positions of u and b on the logarithmic scale in terms of the base c. We take the position of u and divide it by the length to get to b. In effect, this is a change of units calculation, similar to finding a distance measured in inches when we know the distance measured in centimeters.

Closely related to the change of base formula is the fact that we can rewrite any power with a positive base using a composition with an elementary exponential. We will soon discover that the number e is the natural exponential base. Thus, every power can be rewritten in terms of the natural exponential function. The logarithm with this base is the **natural logarithm** ln.

Example 1.8.9 Rewrite $f(x) = 4 \cdot 3^{2x}$ using the natural exponential function.

Solution. One approach is to use the inverse property, $e^{\ln(u)} = u$, on the factor with the power. Then use logarithm properties to simplify (expand) the power.

$$\begin{aligned} f(x) &= 4 \cdot 3^{2x} \\ &= 4 \cdot e^{\ln(3^{2x})} && \text{(Inverse Property)} \\ &= 4 \cdot e^{2x\ln(3)} && \text{(Power Rule)} \\ &= 4 \cdot e^{2\ln(3)\,x} && \text{(Commute)} \end{aligned}$$

Another approach is to just rewrite the base using the inverse property, $3 = e^{\ln(3)}$, and then finish by using properties of powers.

$$\begin{aligned} f(x) &= 4 \cdot 3^{2x} \\ &= 4 \cdot (e^{\ln(3)})^{2x} && \text{(Inverse Property)} \\ &= 4 \cdot e^{2x\ln(3)} && \text{(Power of Product)} \\ &= 4 \cdot e^{2\ln(3)\,x} && \text{(Commute)} \end{aligned}$$

□

We use base e for exponentials so much that we summarize the statement as a theorem.

Theorem 1.8.10 *Every exponential function $f(x) = Ab^x$ can be written using the natural exponential $f(x) = Ae^{kx}$ where $k = \ln(b)$.*

Example 1.8.11 A population P is an exponential function of time t, $P = Ae^{kt}$. Suppose that $P = 500$ when $t = 0$ and the population triples every 5 years. Find the formula for P.

Solution. This is an exponential model $P = Ae^{kt}$ with unknown parameters A and k. We use the data $(t, P) = (0, 500)$ and $(t, P) = (5, 1500)$ (the population triples in 5 years) to create equations based on our model that constrain the parameters.

$$\begin{cases} 500 = Ae^{0k} \\ 1500 = Ae^{5k} \end{cases}$$

The first equation gives $A = 500$. Substituting that into the second equation, we can solve for k.

$$\begin{aligned} Ae^{5k} &= 1500 \\ 500e^{5k} &= 1500 \\ e^{5k} &= 3 \\ 5k &= \ln(3) \\ k &= \tfrac{1}{5}\ln(3) \end{aligned}$$

Using these parameters, we have our model

$$P = 500e^{\frac{1}{5}\ln(3)t}.$$

□

1.8.4 Summary

- The logarithmic scale is based on the inverse map of an exponential function. The multiplicative identity $u = 1$ represents the origin of the scale. The base b establishes the length scale used to calculate logarithms.
- The properties of logarithms are consequences of properties of exponents.
  - $\log_b(1) = 0$ (Identities Map) and $\log_b(b) = 1$ (Base Sets Scale).
  - $b^{\log_b(u)} = u$ and $\log_b(b^x) = x$ (Inverse Property).
  - $\log_b(uv) = \log_b(u) + \log_b(v)$ (Product Rule).
  - $\log_b(u \div v) = \log_b(u) - \log_b(v)$ (Quotient Rule).
  - $\log_b(u^p) = p \cdot \log_b(u)$ (Power Rule).
- To expand a logarithm is to identify if the input of a logarithm is a product, quotient, or power, and then to rewrite that logarithm as a sum of logarithms, a difference of logarithms, or a product with a logarithm. This is repeated until no logarithm has an input that is a product, quotient, or power.

1.8.5 Exercises

1. For an unknown base b, we have $\log_b(2) = 0.3$. Use the rules of logarithms to find each of the following.

(a) $\log_b(8)$

(b) $\log_b(\frac{1}{\sqrt{2}})$

(c) $\log_b(4b^2)$

Can you identify the value b?

2. For an unknown base b, we have $\log_b(2) \approx 0.3562$ and $\log_b(3) \approx 0.5646$. Use the rules of logarithms to find each of the following.

(a) $\log_b(6)$

(b) $\log_b(72)$

(c) $\log_b(\frac{4}{9})$

3. Expand $\log\left(3x^5(2x+1)^4\right)$ as far as possible.

4. Expand $\log\left(\dfrac{(x^2+4)^3}{x^4(3x+1)}\right)$ as far as possible.

5. Expand $\log\left(\sqrt{5(x^2+1)}\right)$ as far as possible.

6. Rewrite the expanded formula $2+3\log(x)-\log(2x+1)$ as the logarithm (base 10) of a single expression.

7. Rewrite the expanded formula $\frac{1}{2}\ln(x)-\ln(x+1)-\ln(x-1)$ as the logarithm (base e) of a single expression.

8. Use the natural logarithm to solve the equation $4 \cdot 5^x = 3$.

9. Use the natural logarithm to solve the equation $2 \cdot 3^{-x} = 3 \cdot 2^x$.

10. Write the function $f(x) = 4^x$ using an exponential with base 10.

11. Write the function $f(x) = 4^x$ using an exponential with base e.

12. Write the function $f(x) = 5 \cdot 0.25^x$ using an exponential with base e.

13. Write the function $f(x) = x^4$ using an exponential with base e for $x > 0$.

14. Write the function $f(x) = x^{2x}$ using an exponential with base e for $x > 0$.

15. Find an exponential model $y = Ae^{kx}$ satisfying the states $(x, y) = (0, 3)$ and $(x, y) = (5, 9)$.

16. Find an exponential model $y = Ae^{kx}$ satisfying the states $(x, y) = (1, 3)$ and $(x, y) = (4, 6)$.

17. In a living organism, 1 gram of carbon would result in about 840 carbon-14 atoms disintegrating per hour. After death, the rate of radiocarbon disintegrations decays exponentially. Carbon-14 has a half-life of 5730 years, meaning the rate has dropped to half its original value after this time. Determine the radioactive disintegration rate for 1 gram of carbon using the natural exponential base e. What is the radioactive disintegration rate of the sample after 1000 years?

18. A money market account starting at \$2000 grew by 10% in one year. Determine the value of the money market account assuming the rate of growth remains constant by using an exponential growth model. What will be the value in 4 years? How long does it take for the value to double?

1.9 Applications of Logarithms

The properties of logarithms are useful for a variety of applications. In this section, we discuss using a logarithm to transform data. We will see that data following an exponential model look linear in a semi-log transformation; data following a power law model look linear in a log-log transformation. We also consider an application to probability in relation to log-likelihood.

1.9.1 Logarithmic Transformations

Sometimes we look at data that are at many different scales. On a standard number line, we think of the numbers 1, 10, 100, and 1000 as spread very far apart. However, if we consider looking at a usual number line that will include all of these values, the relative space between 1 and 10 is very small and both seem very close to zero.

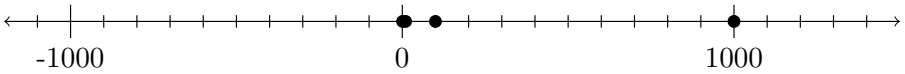


Figure 1.9.1 A number line showing the numbers 1, 10, 100, and 1000.

In a similar way, we might normally think of the numbers 0.1, 0.01, 0.001, and 0.0001 as all very close to 0. However, each value is a different order of magnitude exactly like the values 1, 10, 100, and 1000. The common logarithm is the logarithm for base ten ($b = 10$). Consequently, the common logarithm of these numbers would give us equally spaced integers.

$$\begin{aligned}
0.0001 = 10^{-4} &\quad\Leftrightarrow\quad \log 0.0001 = -4 \\
0.001 = 10^{-3} &\quad\Leftrightarrow\quad \log 0.001 = -3 \\
0.01 = 10^{-2} &\quad\Leftrightarrow\quad \log 0.01 = -2 \\
0.1 = 10^{-1} &\quad\Leftrightarrow\quad \log 0.1 = -1 \\
1 = 10^{0} &\quad\Leftrightarrow\quad \log 1 = 0 \\
10 = 10^{1} &\quad\Leftrightarrow\quad \log 10 = 1 \\
100 = 10^{2} &\quad\Leftrightarrow\quad \log 100 = 2 \\
1000 = 10^{3} &\quad\Leftrightarrow\quad \log 1000 = 3
\end{aligned}$$

Because the logarithm spaces values apart according to the order of magnitude, we can often use logarithms to visualize data that occur at multiple orders of magnitude. Quality plotting tools allow us to plot the data but use the logarithm for their position in the graph. This is called using a logarithmic scale. We can choose to use a logarithmic scale for one or both axes.

Example 1.9.2 Brain size is strongly correlated with overall body mass in mammals. However, mammals cover a wide range of different sizes. The graph of brain size versus body mass for 96 species is shown in Figure 1.9.3, based on data from *The Statistical Sleuth* by Ramsey and Schafer (2013). Because the elephant is so large relative to many other species, its data point requires a wide window in the figure. The majority of species, however, are much smaller and form a crowded cluster of points near the origin.

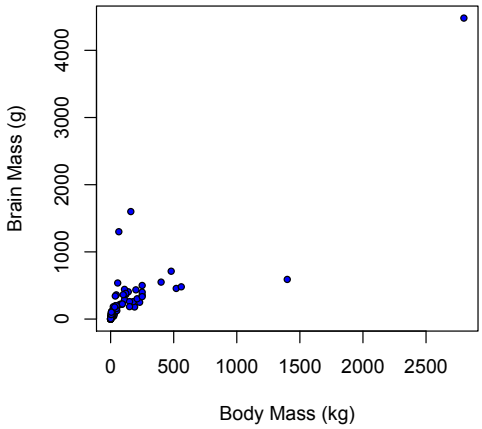


Figure 1.9.3 Plot of body mass (kg) and brain size (g) for 96 species of mammals.

This suggests replotting the data using a logarithmic scale. The same data a shown with logarithmic scales for both variables in the Figure 1.9.4. Using a logarithmic scale on both axes is called a **log-log plot**. The transformed data spreads the points out more uniformly across the figure. In addition, the log-log plot suggests that the transformed data is approximately linear.

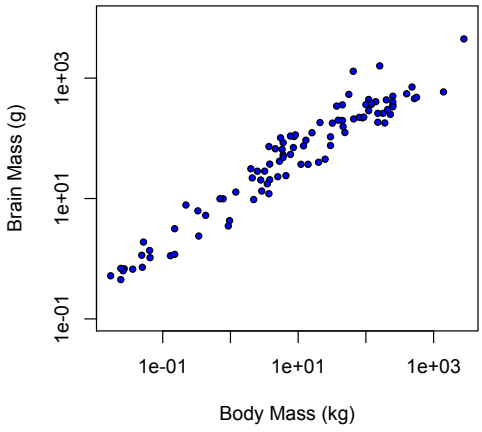


Figure 1.9.4 Log-log plot of body mass (kg) and brain size (g) for 96 species of mammals.

□

The previous example illustrated a dataset where transformed data look linear. Let us work out what that relation must be like.

Suppose we have raw data with variables (x, y) and we transform the data with logarithms. This creates two new variables, $u = \log(x)$ and $v = \log(y)$. The log-log plot is a figure showing data (u, v) but with the axes showing the original values on a logarithmic scale. If the transformed data are linear, there must be a model

$$v = a\,u + b.$$

We now substitute our original variables and solve for y. We collect terms

in the logarithm.

$$\begin{aligned}\log(y) &= a\,\log(x) + b \\ &= \log(x^a) + \log(10^b) \\ &= \log(10^b \cdot x^a)\end{aligned}$$

Thus, we find $y = 10^b \cdot x^a$, which is a power law model. We summarize our result as a theorem for future reference.

Theorem 1.9.5 *Data (x, y) such that the transformed data $(\log(x), \log(y))$ (a log-log plot) has a linear relation will satisfy a power law relation.*

Another common transformation is a semi-log plot. This occurs when only the dependent variable is transformed. In other words, only the y-axis is transformed to a logarithmic scale. What relationship does this reveal?

Suppose we have raw data with variables (x, y) and we only transform y with a logarithm,

$$v = \log(y).$$

The semi-log plot is a figure showing data (x, v) but with the axes showing the original values on a logarithmic scale. If the transformed data are linear, there must be a model

$$v = a\,x + b.$$

We now substitute our original variables and solve for y.

$$\log(y) = a\,x + b.$$

To solve for y, we use the inverse operation to the logarithm, which is an exponential.

$$y = 10^{a\,x+b}.$$

Using the properties of exponents, we can rewrite this

$$y = 10^{a\,x} \cdot 10^b = 10^b \cdot (10^a)^x.$$

Thus, we find $y = A\,B^x$, with $A = 10^b$ and $B = 10^a$, which is an exponential model.

Theorem 1.9.6 *Data (x, y) such that the transformed data $(x, \log(y))$ (a semi-log plot) has a linear relation will satisfy an exponential relation.*

We can use the log-transformations to find the power law and exponential relations for actual data. If we know that (x, y) satisfies a power law for given data, then we know $(\log(x), \log(y))$ satisfies a linear model. We can calculate the slope and intercept of the transformed linear model and then solve for y. If we know that (x, y) satisfies an exponential model for given data, then we can find the equation of a line for $(x, \log(y))$ and then solve for y.

Example 1.9.7 Find the power law for (x, y) that includes data $(x, y) = (2, 5)$ and $(4, 8)$.

Solution. Power law data is linear under a log-log transformation, $u = \log(x)$ and $v = \log(y)$. The transformed points are $(u, v) = (\log x, \log y) = (\log(2), \log(5))$ and $(u, v) = (\log x, \log y) = (\log(4), \log(8))$ The slope is calculated and simplified using properties of logarithms,

$$\begin{aligned}m = \frac{\Delta v}{\Delta u} &= \frac{\log(8) - \log(5)}{\log(4) - \log(2)} \\ &= \frac{\log(8/5)}{\log(4/2)}\end{aligned}$$

$$= \frac{\log(8/5)}{\log(2)}$$

Using the point-slope form of a line, we have

$$\begin{aligned} v - \log(5) &= m(u - \log(2)) \\ &= \frac{\log(8/5)}{\log(2)}(u - \log(2)) \end{aligned}$$

Now we substitute back the original variables with $u = \log(x)$ and $v = \log(y)$. Alternatively, we could have done the work above using $\log x$ and $\log y$ in place of u and v. Our final equation then would be written

$$\log(y) - \log(5) = \frac{\log(8/5)}{\log(2)}(\log(x) - \log(2)).$$

We now proceed to apply the rules for logarithms to simplify our expression.

Our goal is to have a logarithm on the left equaling a logarithm on the right. The first step is to use the quotient rule of logarithms:

$$\log(y/5) = \frac{\log(8/5)}{\log(2)} \log(x/2).$$

On the right, we have a product of values. A linear log-log plot corresponds to a power law, and we ultimately want our equation to have x raised to a power. Because x currently appears within the logarithm, we will use the power rule for logarithms. The logarithm with x, $\log(x/2)$, is multiplied by an expression. That expression, which originally served as our slope, will become the power. For simplicity in writing, we introduce a new symbol, $p = \dfrac{\log(8/5)}{\log(2)}$.

$$\begin{aligned} \log(y/5) &= p \cdot \log\left(\frac{x}{2}\right) \\ \log(y/5) &= \log\left(\frac{x}{2}\right)^p \end{aligned}$$

Now that we have the logarithm of an expression on the left and on the right, the expressions within the logarithms must be equal.

$$\begin{aligned} \frac{y}{5} &= \left(\frac{x}{2}\right)^p \\ y &= 5 \cdot \left(\frac{x}{2}\right)^p \end{aligned}$$

This equation is our model for the power law relation. □

We now illustrate the similar process for a semi-log transformation. Exponentially related data will appear linear on a semi-log plot.

Example 1.9.8 Find the exponential law for (x, y) that includes data $(x, y) = (2, 4)$ and $(5, 10)$.

Solution. Exponential law data is linear under a semi-log transformation, $u = x$ and $v = \log(y)$ The transformed points are $(u, v) = (x, \log y) = (2, \log(4))$ and $(u, v) = (x, \log y) = (5, \log(10))$. The slope is calculated and simplified using properties of logarithms,

$$m = \frac{\Delta v}{\Delta u} = \frac{\log(10) - \log(4)}{5 - 2}$$

$$= \frac{\log(10/4)}{3}$$
$$= \frac{\log(5/2)}{3}$$

Using the point-slope form of a line, we can create our equation relating our variables.

$$v - \log(4) = m(u - 2)$$
$$\log y - \log 4 = \frac{\log(5/2)}{3}(x - 2)$$

We now seek to find an equivalent equation where a logarithm of an expression appears alone on each side of the equation. We first apply the quotient rule for logarithms on the left.

$$\log y - \log 4 = \frac{\log(5/2)}{3}(x - 2)$$
$$\log(y/4) = \frac{\log(5/2)}{3}(x - 2)$$

Because data that are linear in a semi-log plot have an exponential relation, we want to see how to get x into the exponent. On the right-hand side, we have a logarithm, $\log(5/2)$, that is multiplied by $(x - 2)$ and divided by 3. We can group those together as a single multiplication and then apply the power rule for logarithms.

$$\log(y/4) = \frac{x - 2}{3} \cdot \log(5/2)$$
$$\log(y/4) = \log\left((5/2)^{(x-2)/3}\right)$$

We can now eliminate the logarithm from both sides of the equation.

$$\frac{y}{4} = (5/2)^{(x-2)/3}$$
$$y = 4 \cdot (5/2)^{(x-2)/3}$$

□

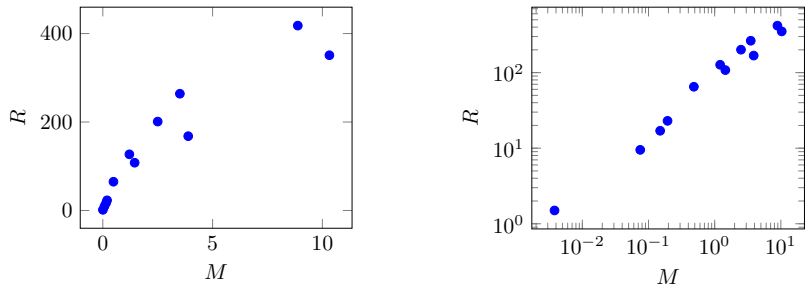
We end with an example of how this might relate to actual data.

Example 1.9.9 In 1967, Lasiewski and Dawson published an article in *The Condor* relating the body mass M (in kg) and resting metabolic rate R (in kcal/day) for birds. They tabulated the recorded body mass and metabolic rate for individual birds based on published studies. The following table includes twelve of these birds. Graph the data and determine if the data appear linear in a log-log plot. Then use a linear regression of transformed data to estimate the model

Table 1.9.10 Selected body mass and resting metabolic rate of birds, as tabulated in Lasiewski and Dawson (1967).

Bird	M (kg)	R (kcal/day)
Rufous hummingbird	0.0038	1.5
Common nighthawk	0.075	9.5
Common wood pigeon	0.150	17.0
Northern bobwhite	0.194	23.0
Wood duck	0.485	65
Pacific gull	1.21	127
Great horned owl	1.450	108
Wood stork	2.5	201
Brown pelican	3.51	264
Sandhill crane	3.89	168
Trumpeter swan	8.88	418
Andean condor	10.32	351

Solution. After entering the data into a spreadsheet or other graphing utility, we generate a scatterplot of the points (M, R), as shown below on the left. You should see that the relationship of the data is increasing and concave down. If we modify both axes to use a logarithmic scale, as shown below on the right, we see that the transformed relationship looks reasonably linear. This suggests using a linear relation on the transformed coordinates.



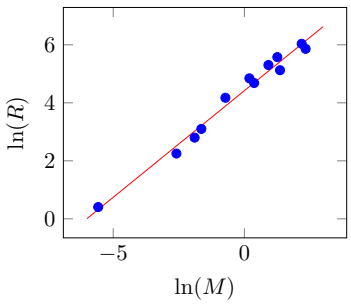
To form the linear model, we need to generate actual transformed values and not just a graph using logarithm scales. In the spreadsheet, we will create two new columns for $\ln(M)$ and for $\ln(R)$. For example, suppose the mass M of Selasphorus rufus appears in the spreadsheet in cell `B2` and we want to generate the transformed variable $\ln(M)$ in cell `D2`. In cell `D2`, we would type the formula `=ln(B2)`. Copying this formula and pasting it into other cells will preserve the relative location. If you paste it into cell `D3`, you will discover it automatically changed the formula to `=ln(B3)`.

Once we have new columns $\ln(M)$ and $\ln(R)$, we can create a new scatterplot of data $(\ln(M), \ln(R))$ using *linear* scales. This new graph will have the same appearance as the original data using *logarithmic* scales, except that the axes show the logarithm of the data rather than the original data using logarithmic scales. With this new graph, we can find the linear trend line. The graph below shows the graph of the transformed data, along with a trend line using the formula

$$y = 0.7356x + 4.4192.$$

We change the variables to those plotted to give a transformed model

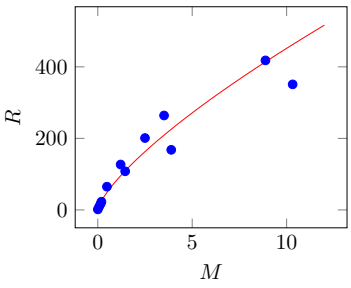
$$\ln(R) = 0.7356 \ln(M) + 4.4192.$$



We find the model of the relation for the original variables $M \mapsto R$ by solving for R and simplifying. To eliminate the logarithm, we apply the inverse operation of the natural exponential. Because our data are approximate, we can use decimal approximations for our formulas.

$$\begin{aligned} \ln(R) &= 0.7356 \ln(M) + 4.4192 \\ e^{\ln(R)} &= e^{0.7356 \ln(M) + 4.4192} \\ R &= e^{0.7356 \ln(M)} \cdot e^{4.4192} \\ &= e^{\ln(M^{0.7356})} \cdot 83.030 \\ &= 83.030 \cdot M^{0.7356} \end{aligned}$$

We see that the model is a power function, $R = 83.030M^{0.7356}$. The figure below shows the original data using linear axes along with this approximating model.



□

In practice, spreadsheets have built-in tools to accomplish the transformed calculations. When a spreadsheet application allows you to add a trend-line to a given plot, it often allows you to select a variety of different models. When it allows you to use an exponential model, the spreadsheet is internally using a semi-log transform, finding the linear trend-line, and then reporting back the resulting exponential model. Similarly, when a spreadsheet allows you to choose a power law model, the spreadsheet is internally finding the linear equation for a log-log transform and then reporting back the resulting power law model.

1.9.2 Enrichment: Log-Likelihood

Note 1.9.11 This section is included as an example of how logarithms play a more fundamental role in a more advanced sense than just transforming data. The content is optional. Subsequent sections do not rely on students having learned this material.

Suppose that we are performing an experiment that has a random outcome with two possibilities. We do not know in advance the probabilities associated with the two outcomes. For example, flipping a coin results in heads or tails.

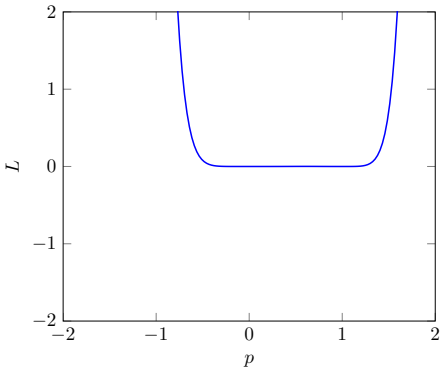
A fair coin has equal probabilities. A biased coin has unequal probabilities. If we suspected a coin was biased, we might want to determine the coin's true odds for heads versus tails. We would like to use repetition of an experiment in order to determine these probabilities.

In statistics, there is a method commonly used to estimate unknown parameters called the maximum likelihood principle. Each observation is assumed to have outcomes governed by a probability distribution characterized by certain model parameters. The likelihood L is the product of the probabilities densities associated with each observation. The maximum likelihood method adopts the parameter values that makes the likelihood as large as possible.

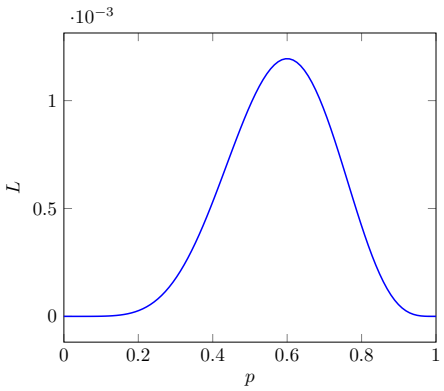
For our experiment with two different outcomes, the probability distribution is characterized by one parameter, p, which gives the probability of the first outcome (often called a success). The probability of the second outcome (often called a failure) will be $1-p$ since probabilities must add to 1. Suppose that we repeated the experiment ten times and counted six successes and four failures. The likelihood is the product of the probabilities for these outcomes, using the expressions involving the parameter. The likelihood will be the product of six factors with p and four factors with $1-p$. Writing these with powers, the likelihood is a function of p,

$$L = p^6(1-p)^4.$$

How will we maximize this value? Until we learn some calculus, we will need to find the maximum using a graph. The graph of this formula is shown below. (To make a graph, most graphing utilities require that you use the independent variable x in place of p.)



How do we interpret this graph? Because the parameter p is supposed to be a probability, we require $0 < p < 1$. But the graph doesn't seem to show a maximum there. This is because values of p outside the meaningful domain dominate the figure. If we redo the graph so that the domain only include $[0, 1]$, we get a better picture.



This graph has a maximum value at $p = 0.6$. We can also see why the earlier graph didn't show the maximum. The scale on the vertical axis for L for the restricted interval has an order of magnitude of 10^{-3}. If we had even more data than ten observations—and to estimate probabilities we need many more—this magnitude would be even smaller. Because of this effect that the likelihood shrinks in magnitude with more data, the likelihood value often drops below the smallest number a computer can represent. It would then be impossible to find the maximum likelihood parameter value.

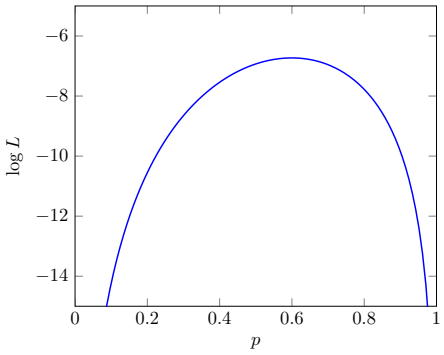
To avoid this issue, data scientists typically record the log-likelihood rather than the likelihood. Maximizing the log-likelihood will always give the same values as maximizing the likelihood itself. The log-likelihood is calculated as the natural logarithm of the likelihood,

$$\log L = \ln(L).$$

Because the logarithm of a product is equal to the sum of the logarithms of the factors, the log-likelihood is calculated by *adding* the logarithms of the probability densities corresponding to the observations. For our example,

$$\begin{aligned} \log L &= \ln\left(p^6\,(1-p)^4\right) \\ &= \ln\left(p^6\right) + \ln\left((1-p)^4\right) \\ &= 6\ln(p) + 4\ln(1-p) \end{aligned}$$

A graph of the log-likelihood $\log L$ versus p is shown in the figure below. The maximum value again occurs at $p = 0.6$.



Example 1.9.12 An exponential time is a random time until some event occurs that is characterized by gaining no information by knowing how long has already passed without the event yet occurring. The time until a radioactive particle decays is an example of an exponential time. The mathematical model for the probability density of an exponential time t has a single parameter,

usually represented by the Greek letter lambda λ,

$$f(t) = \lambda e^{-\lambda t}.$$

In a series of five experiments, the observed exponential times were recorded as $t_1 = 12.3$, $t_2 = 4.6$, $t_3 = 23.1$, $t_4 = 0.4$, and $t_5 = 10.5$. Calculate the log-likelihood for this collection of data, plot the log-likelihood, and determine the maximum likelihood value for the parameter λ.

Solution. The logarithm of the density is

$$\begin{aligned} \ln f(t) &= \ln\left(\lambda e^{-\lambda t}\right) \\ &= \ln(\lambda) + \ln(e^{-\lambda t}) \end{aligned}$$

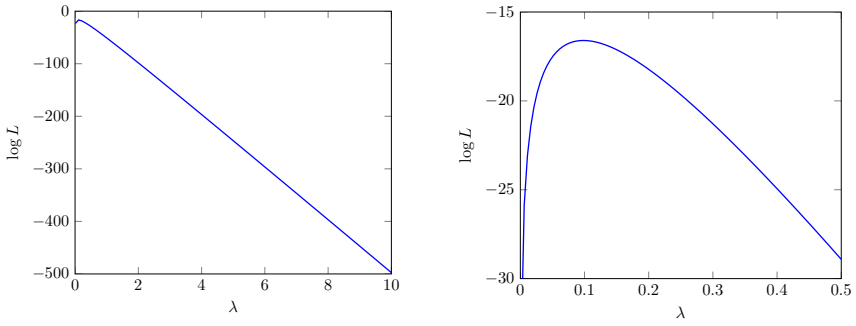
Because the natural logarithm and the exponential with the natural base e are inverses, we can simplify further to obtain

$$\ln f(t) = \ln(\lambda) - \lambda t.$$

The log-likelihood is the sum of the logarithms of the densities using the observed times. Each observation will result in adding $\ln(\lambda)$, so we obtain

$$\begin{aligned} \log L &= 5\ln(\lambda) - \lambda(12.3 + 4.6 + 23.1 + 0.4 + 10.5) \\ &= 5\ln(\lambda) - 50.9\lambda \end{aligned}$$

The parameter λ only needs to be a positive number. If we plot values $0 < \lambda < 10$ to explore where the maximum might be, we get the figure on the left. It shows the graph steadily decreasing, which means the maximum is close to zero. If we plot value $0 < \lambda < 0.5$, we get the figure on the right. The maximum value occurs at $\lambda = 0.098231$, which is our maximum likelihood estimate of the parameter.



□

1.9.3 Summary

- Transforming data with a logarithm allows us to view the distribution of data spread over a wide range of magnitudes.
- Data that appear linear in a log-log plot (both axes in logarithmic scale) follow a power law relation.
- Data that appear linear in a semi-log plot (only y-axis in logarithmic scale) follow an exponential relation.
- Estimating parameters for probability distributions is frequently based on maximum likelihood estimation. To avoid numerical underflow (expo-

nentially small magnitudes) of the likelihood, this is more common done using the log-likelihood.

1.9.4 Exercises

1. Suppose data for (t, M) appear linear in a semi-log plot. If the data include the points $(t, M) = (2, 5)$ and $(t, M) = (5, 2)$, find a linear model for the tranformed data and use it to find the appropriate model for the original data.

2. Suppose data for (P, S) appear linear in a log-log plot. If the data include the points $(P, S) = (2, 5)$ and $(P, S) = (5, 2)$, find a linear model for the tranformed data and use it to find the appropriate model for the original data.

3. A random experiment has two possible outcomes, high or low. The probability the result is high is represented by p, with $0 < p < 1$, and the probability the result is low is represented by $1 - p$. Twenty independent replicates of the experiment resulted in six highs and fourteen lows. Calculate the formula for the likelihood and use it to compute the log-likelihood. With a graph, estimate the maximum likelihood value for p.

4. An experiment results in randomly distributed exponential times. The probability density used in the likelihood has a single parameter λ,

$$f(t) = \lambda e^{-\lambda t}.$$

Replicating the experiment six times results in measured times $t_1 = 0.826$, $t_2 = 0.293$, $t_3 = 0.218$, $t_4 = 0.024$, $t_5 = 0.561$, and $t_6 = 0.233$. Calculate the formula for the likelihood and use it to compute the log-likelihood. With a graph, estimate the maximum likelihood value for λ.

5. A manufacturer tracks quality control by testing random samples for proper performance. The number n of identified flaws is a random value that occurs with a probability

$$f(n) = a_n e^{-\lambda} \lambda^n,$$

where a_n does not depend on the model parameter λ. To find the maximum likelihood value for λ, the value of a_n does not matter. For five days of quality control tracking, the number of observed flaws were recorded. $n_1 = 4$, $n_2 = 2$, $n_3 = 5$, $n_4 = 4$, and $n_5 = 8$. Calculate the formula for the likelihood using $a_n = 1$ and use it to compute the log-likelihood. With a graph, estimate the maximum likelihood value for λ.

Chapter 2

Functions to Model Relationships

2.1 An Introduction to Functions

Overview. Calculus studies the relationships between variables. We have been learning about relationships described by equations where a dependent variable, say y, is equal to an expression involving only the independent variable, say x. In mathematics, such relationships are most are generalized to create the concept of functions. A function is a predictive relationship between an independent and a dependent variable which we interpret as a **map**, $x \mapsto y$, meaning that knowing x we can predict the value of y.

In this section, we will study an overview of the core concepts relating to functions. Functions will generalize our idea of operations that can act on expressions. We will learn how to think of a function as a map between two variables. Associated with a function as a map are the sets known as the domain, codomain, and range.

2.1.1 Models as Functions

We have previously used equations as models of relations between variables. When we think of one variable as a dependent variable based on the other variable as the independent variable, we are mentally thinking of the equation as defining a map. Each model equation where one variable is defined as an expression in terms of the other variable represents such a map. We define the **domain** as the set of all possible values of the independent variable and the **range** as the set of all resulting values of the dependent variable.

A graphical view of a map is using two number lines, one for the domain and one for the range. For each value in the domain, we imagine that the map defines an arrow originating at the point in the domain and ending at the point in the range. If a value on the number line does not belong to the domain, there just isn't any arrow originating at that point.

Example 2.1.1 Consider a function defined by a linear relation with states $(x, y) = (1, 2)$ and $(x, y) = (3, 8)$. We can find the equation of the line by finding the slope and using the slope–intercept equation. The slope is interpreted as the ratio of the change in the output to the change in the input. The increment of the input is $\Delta x = x_2 - x_1 = 3 - 1 = 2$. The increment of the output is $\Delta y = y_2 - y_1 = 8 - 2 = 6$. The slope or rate of change is therefore $m = \frac{6}{2} = 3$. The equation of the line using the point $(x, y) = (1, 2)$ becomes

$$y - 2 = 3\,(x - 1).$$

If we solve for y, we get y as an explicit function of x,

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

The interactive figure below illustrates the idea of a map. The top number line contains the domain, while the bottom number line contains the range. A slider on the top line allows you to choose a value in the domain. The arrow dynamically moves to connect the point in the domain to the point in the range.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.1.2 The linear function that maps $x = 1 \mapsto y = 2$ and $x = 3 \mapsto y = 8$ has rate of change $m = 2$.

This map view of the linear function gives an interesting visual interpretation of the slope. We consider the initial point $(x, y) = (1, 2)$ as giving reference

values on each number line. The map sends $x = 1 \mapsto y = 2$. As we move the slider away from $x = 1$, the gap between the value of x and $x = 1$ defines Δx. The slope of the line then forces the gap from $y = 2$ and the value y coming from x will be $\Delta y = 3\,\Delta x$. That is, the slope is the scaling factor going from Δx to Δy. □

The algebraic interpretation of a map defined by an equation uses variable substitution. To find the value of the dependent variable, we substitute the assigned value of the independent variable into the equation. After we simplify the expression, we determine the value of the dependent variable.

Example 2.1.3 In 1990, the population of Harrisonburg, Virginia, was 30,707. In 2010, the population was 48,914. If the rate of change of the population increased at a constant rate, find a model for the population of Harrisonburg as a function of the year. What does the model predict for the year 2020?

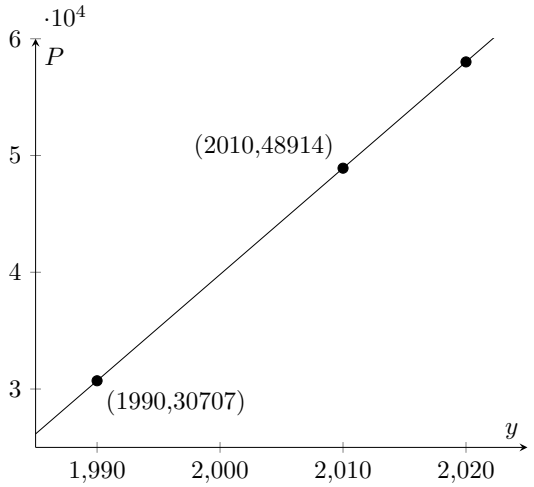
Solution. The variables involved in the model are the year y and the population P. By saying that the rate of change is constant means that the model uses a linear function. The given data show $y = 1990 \mapsto P = 30707$ and $y = 2010 \mapsto P = 48914$. We find the slope as the ratio of ΔP (change in output) to Δy (change in input).

$$\begin{aligned} m &= \frac{\Delta P}{\Delta y} \\ &= \frac{48914 - 30707}{2010 - 1990} \\ &= \frac{18207}{20} = 910.35 \end{aligned}$$

Using the point-slope form of the line, we start with the known mapping $1990 \mapsto 30707$ and the constant rate of change to give

$$P = 30707 + 910.35\,(y - 1990).$$

The graph of this line is shown below.



To find the predicted population for 2020, we use the model with substitution. Let $y = 2020$ and substitute this into the model:

$$\begin{aligned} P &= 30707 + 910.35\,(y - 1990) \\ &= 30707 + 910.35\,(2020 - 1990) \\ &= 30707 + 910.35 \cdot 30 \end{aligned}$$

$$= 58017.5$$

Consequently, the model predicts the map $y = 2020 \mapsto P = 58017.5$. Of course, the real population will not be a non-integer value. This example reminds us to distinguish between a prediction of a model and an actual value. □

Because we use functions to find values of the dependent variable by substitution, we have a special notation called **function notation**. The map or function is assigned a name. The name of the function followed by an expression inside of parentheses represents the value of the dependent variable when the input expression is substituted for the independent variable.

For example, if we write $f : x \mapsto y = 3x^2$, then the name of the function is f. The equation defining the relation is $y = 3x^2$. This is more commonly written $y = f(x) = 3x^2$. The expression $f(1)$, substituting x with $x = 1$, we find $y = 3(1)^2 = 3$. More simply, we write $f(1) = 3(1)^2 = 3$. Substitution can involve entire expressions as well as single values. The expression $f(2+h)$ represents the expression $3x^2$ when $2 + h$ is substituted for x,

$$f(2 + h) = 3\,(2 + h)^2.$$

Example 2.1.4 Suppose we have a function $f : t \mapsto B$ where t measures the time in years since we opened a bank account and B measures the account balance. If the account grows according to the model $B = 500(1.02)^t$, we can write $f(t) = 500\,(1.02)^t$. Find the balance after two, five, and ten years using function notation.

Solution. The values that we want, using function notation, are $f(2)$, $f(5)$, and $f(10)$. Each of these expressions represent values for B when the value of t is replaced by $t = 2$, $t = 5$, and $t = 10$. The three values are

$$\begin{aligned} f(2) &= 500(1.02)^2 = 520.20, \\ f(5) &= 500(1.02)^5 \approx 552.04, \\ f(10) &= 500(1.02)^{10} \approx 609.50. \end{aligned}$$

Thus, the model predicts balance values of $B = 520.20$ when $t = 2$, $B = 552.04$ when $t = 5$, and $B = 609.50$ when $t = 10$. □

Note 2.1.5 Although writing parentheses in mathematics next to a number or a variable means multiplication, a function is not a variable. The parentheses after the function do *not* mean multiplication but evaluation. It is unfortunate that the same symbols have different meanings, so you will need to pay close attention to the context. When reading aloud, a function expression like $f(x)$ should be read "f of x".

If we have an equation involving two variables and can solve to isolate one of the variables to be equal to some expression involving the other variable, then the equation defines a function. The isolated variable is the dependent variable (output) and the other variable is the independent variable. This is a function because once you know the value of the independent variable, the equation allows you to substitute that value and determine a single value of the dependent variable. An equation where the dependent variable is isolated defines the dependent variable as an **explicit function** of the independent variable.

Example 2.1.6 Given the equation $\frac{xy}{x+y} = 3$, rewrite y as an explicit function of x.

Solution. To show that y is a function of x, we need to solve the equation for y. We can cross multiply the equation and then collect terms involving y.

Once we factor the common factor of y, we can isolate the variable.

$$\begin{aligned}\frac{xy}{x+y} &= 3\\ xy &= 3(x+y)\\ xy &= 3x+3y\\ xy-3y &= 3x\\ (x-3)\,y &= 3x\\ y &= \frac{3x}{x-3}\end{aligned}$$

The equation shows that we have a function $x \mapsto y = \dfrac{3x}{x-3}$, an explicit function of x. □

If an equation can be solved for two different values of the dependent variable for a single value of the independent variable, then the equation does not define the dependent variable as a function of the independent variable. We will later learn that we can often restrict the equation to define an implicit function.

Example 2.1.7 Show that $x^2 + y^2 = 25$ does not define y as a function of x.

Solution. The equation $x^2 + y^2 = 25$ defines a circle with radius $r = 5$. For each value a with $-5 < a < 5$, the graph of the circle will intersect the vertical line $x = a$ at two different points. To demonstrate this, consider $x = 3$. When we substitute $x = 3$ into the equation $x^2 + y^2 = 25$, we obtain the equation $9 + y^2 = 25$ which is equivalent to $y^2 = 16$. There are two values, $y = \pm 4$, that solve this equation. Because there are two values for the dependent variable for this point, we see that the equation does not define an explicit function. □

2.1.2 Domain and Range

The **domain** of a function is the set of possible input values for the function. The domain might be all possible real numbers. It might also be a set restricted by algebraic constraints or by physical considerations. A second set associated with a function is the **range**. The range is the set of all possible output values for the function. We often consider a set called the **codomain**, which is a set to which all output values must belong. The range is always a subset of the codomain.

Definition 2.1.8 Function. A function f is a rule or relation from a given set D (the domain) to another set D' (the codomain) such that *every* value $a \in D$ is related (mapped) to a *unique* value $b \in D'$. We write $f : D \to D'$. If D' is not stated, it is assumed that $D' = \mathbb{R}$. ◊

A function must assign a value for the output for *every* value in the domain. Functions using the same rule but for different domains are different functions. Choosing a domain is part of the modeling process. The domain specifies what type of values are acceptable for the independent variable (input). For example, some modeling scenarios might require that the independent variable must be an integer, so we choose a domain $D = \mathbb{Z}$. Other modeling scenarios might require that the independent variable is constrained to be between two values $a < b$, so we choose a domain $D = (a, b)$.

The codomain could always be chosen to be the set of real numbers $\mathbb{R}$. In mathematics, the codomain is usually required to distinguish different types of functions, such as whether the value of the function is a real number, a

complex number, or even maybe a more complicated object like a matrix. For modeling, we would specify a codomain that is more precise in order to characterize some aspect of a function. For example, if the output variable is meant to represent a probability, then only values from 0 to 1 make sense, and we choose $D' = [0, 1]$. Frequently, we want to have a function that is non-negative. We can communicate this by saying $D' = [0, \infty)$.

Example 2.1.9 Consider a grocery store that charges \$0.25 per ear of corn. We can define a function $n \mapsto c$ that maps the number of ears purchased n to the pre-tax subtotal to charge c in dollars. Describe the function.

Solution. The function is characterized by the rule and by the domain. The rule is one of proportionality, $c = 0.25n$. The independent (input) variable n is discrete and only makes sense for non-negative integers. We include $n = 0$ because a customer might not purchase corn. Consequently, $D = \{0, 1, 2, \ldots\} = \mathbb{N}_0$. If a store had a customer limit on how many ears could be purchased, then our domain D would have to be modified.

If we named our function f, the following notation communicates our summary:

$$f : \mathbb{N}_0 \to \mathbb{R}; n \mapsto c = 0.25n.$$

The notation states the name of the function f, the domain $\mathbb{N}_0$, the input and output variables n and c, and the rule $c = 0.25n$. □

Example 2.1.10 Consider a grocery store that charges \$0.25 per pound of bananas. We can define a function $w \mapsto c$ that maps the weight of bananas purchased w to the pre-tax subtotal to charge c in dollars. Describe the function.

Solution. The function is characterized by the rule and by the domain. The rule is again one of proportionality, $c = 0.25w$. However, weight is a continuous variable; the weight of bananas purchased could conceivably be any positive real number. To allow for no purchase, we again include $w = 0$. The domain is therefore an interval of values $0 \leq w$, written $D = [0, \infty)$. If we named our function g, the following notation communicates our summary:

$$g : [0, \infty) \to \mathbb{R}; w \mapsto c = 0.25w.$$

□

In the two previous examples, the formulas for two different functions described the same calculation—multiply the input by 0.25. In the examples, we used mapping notation to describe the functions. Mapping notation has the advantage of being precise but a disadvantage in being a little cumbersome. We can use function notation to define each function if we include a restriction on the domain. For f we would write

$$f(n) = c = 0.25n, \quad n = 0, 1, 2, \ldots.$$

For g we would write

$$g(w) = c = 0.25w, \quad w \geq 0.$$

Notice that if we were to substitute a generic variable x for the input values, we would have $f(x) = 0.25x$ and $g(x) = 0.25x$. Nevertheless, the functions are not the same because they have different domains.

The graphs of these functions are also related but different. The discrete function f will have isolated points for its graph. The continuous function g will have a connected graph. Both functions will have all points on their graphs sitting on the line $y = 0.25x$.

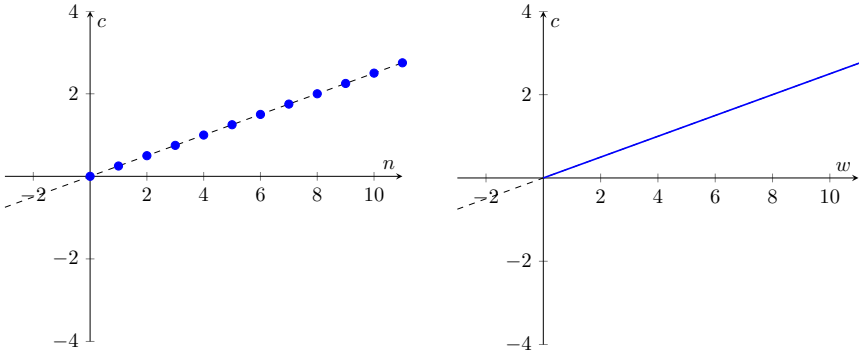


Figure 2.1.11 Graphs of the functions $f : \mathbb{N}_0 \to \mathbb{R}; n \mapsto c = 0.25n$ and $g : [0, \infty) \to \mathbb{R}; w \mapsto c = 0.25w$, each overlaid with the graph of the line $y = 0.25x$.

Unless specified otherwise, the domain of a function defined by a formula will be the largest set of real numbers for which the formula is defined. This type of domain is called the **natural domain** of the function or of the defining expression. However, the codomain might include values that are not necessarily in the range. For example, the function $x \mapsto y = 3$ can be defined for a domain $\mathbb{R}$ (all real numbers) and the range is the set with a single value $\{3\}$. The codomain could be defined to be any set that includes 3, such as the set of non-negative numbers $[0, \infty)$ or the set of all real numbers $\mathbb{R}$. The default codomain will be $\mathbb{R}$.

Definition 2.1.12 For a function f defined by a formula, such as $y = f(x)$, the **natural domain** is the set of all real numbers for which the formula is defined. ◊

Definition 2.1.13 For a function $f : D \to D'$, the **range** is the set of all values y for which there exists a state (x, y). That is, there exists $x \in D$ so that $f(x) = y$. ◊

We find the natural domain by identifying which operations might not be defined for all values and then solve either equations or inequalities that will identify where the function is defined. Our elementary operations and functions use the following constraints to find the domain.

- Division is undefined if the denominator equals zero.
- Even roots (e.g., square roots) and irrational powers are undefined if the inner expression is negative.
- Logarithms are undefined if the inner expression is non-positive (zero or negative).

The inequalities that arise in finding the domain can be solved directly or by using factor analysis.

Example 2.1.14 Determine the domain of $f(x) = \dfrac{2x + 3}{x^2 - 4}$.

Solution. Because $f(x)$ is defined as a quotient, the domain will be the set of all values where $x^2 - 4 \neq 0$. We solve this inequality by factoring and considering the complementary equation $x^2 - 4 = 0$, since a product can only equal zero if one of the factors equals zero.

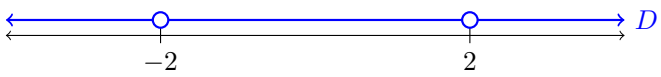
$$x^2 - 4 = 0$$
$$(x + 2)(x - 2) = 0$$
$$x + 2 = 0 \quad \text{or} \quad x - 2 = 0$$

$$x = -2 \quad \text{or} \quad x = 2$$

This means $f(x)$ is defined for all inputs except $x = -2$ or $x = 2$.

To describe the domain using intervals, we think of the real number line and remove $x = \pm 2$. A graphical representation of the set using a number line is shown below. Intervals are read from the line left-to-right. It starts at $-\infty$ and continues until -2, then goes from -2 to 2, and finally goes from 2 until $+\infty$. We write

$$D = (-\infty, -2) \cup (-2, 2) \cup (2, +\infty).$$



□

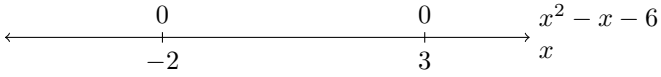
Sometimes finding the domain of a function involves performing sign analysis (such as for a square root or a logarithm). We identify end points of intervals where the expression of interest *might* change sign by solving equations. These end points only occur where the expression equals zero or where the expression itself is undefined (a discontinuity). We test the sign of the expression in each of the resulting intervals by using test points or counting the number of negative factors. Testing the sign at single points is often more efficient than counting negative factors.

Example 2.1.15 Find the domain of the function $g(x) = \log_4(x^2 - x - 6)$.

Solution. The logarithm in $g(x)$ will only have a real value when the input expression is positive, $x^2 - x - 6 > 0$. Our task becomes determining the signs of the expression $x^2 - x - 6$. To illustrate the process of testing points in intervals, we first find possible sign-changing points. The expression is always defined (no discontinuities) so we just solve for zeros $x^2 - x - 6 = 0$ by factoring.

$$\begin{gathered} x^2 - x - 6 = 0 \\ (x - 3)(x + 2) = 0 \\ x - 3 = 0 \quad \text{or} \quad x + 2 = 0 \\ x = 3 \quad \text{or} \quad x = -2 \end{gathered}$$

If we mark these points on a number line, we can easily identify the intervals to test for signs. It is helpful to use the same number line to record the resulting signs, so we can label x-values below the line and the resulting sign or value of the expression above the line.



The number line shows we need to test the intervals $(-\infty, -2)$, $(-2, 3)$, and $(3, \infty)$. Choosing one value from each interval, we can evaluate the expression at that point and identify the sign.

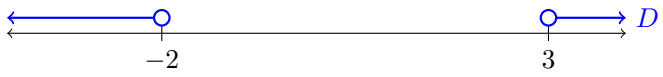
$$\begin{aligned} x = -3 \quad &\Rightarrow \quad x^2 - x - 6 = (-3)^2 - (-3) - 6 = 6 \\ x = 0 \quad &\Rightarrow \quad x^2 - x - 6 = 0^2 - 0 - 6 = -6 \\ x = 4 \quad &\Rightarrow \quad x^2 - x - 6 = 4^2 - 4 - 6 = 6 \end{aligned}$$

We can now update the number line we started by recording either + or − above each interval that we tested. These signs are identical to what we would get by using the number of negative factors for values in each interval. In fact, we could test the intervals by thinking about the factors instead of completely evaluating the expression's value.

We were finding the domain of $g(x) = \log_4(x^2 - x - 6)$, which requires $x^2 - x - 6 > 0$. Based on our summary, we need to find all values which result in the expression having a positive sign. So our solution is the set D formed from the union of intervals $(-\infty, -2)$ and $(3, \infty)$,

$$D = (-\infty, -2) \cup (3, \infty).$$

A visualization of the domain on the number line might also help solidify the connections between the sign analysis number line and the domain set.



□

Example 2.1.16 Find the domain of the function $h(x) = \sqrt{\dfrac{4x}{x^2 - 9}}$.

Solution. A square root (any even root) requires that the input expression is non-negative. Our domain is to solve the inequality

$$D = \{x : \frac{4x}{x^2 - 9} \geq 0\}.$$

To use sign analyis, we need to know the zeros and discontinuities and then test each resulting interval. Discontinuities occur when we try to divide by zero.

$$\begin{gathered} x^2 - 9 = 0 \\ (x+3)(x-3) = 0 \\ x + 3 = 0 \quad \text{or} \quad x - 3 = 0 \\ x = -3 \quad \text{or} \quad x = 3 \end{gathered}$$

Zeros for a quotient require that the numerator equals zero.

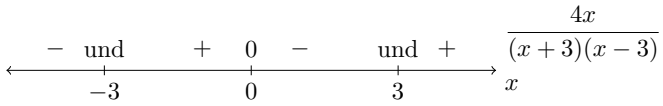
$$\begin{gathered} 4x = 0 \\ x = 0 \end{gathered}$$

Our sign analysis number line will have three points.

Checking one point in each resulting interval gives us the sign. To find the sign, we count the number of negative factors, including the factors in the denominator.

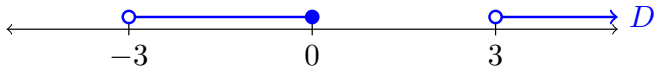
$$\begin{aligned} x = -4 \quad &\Rightarrow \quad \frac{4x}{(x+3)(x-3)} = \frac{4(-4)}{(-4+3)(-4-3)} = \frac{(-)}{(-)(-)} \\ x = -1 \quad &\Rightarrow \quad \frac{4x}{(x+3)(x-3)} = \frac{4(-1)}{(-1+3)(-1-3)} = \frac{(-)}{(+)(-)} \\ x = 1 \quad &\Rightarrow \quad \frac{4x}{(x+3)(x-3)} = \frac{4(1)}{(1+3)(1-3)} = \frac{(+)}{(+)(-)} \\ x = 4 \quad &\Rightarrow \quad \frac{4x}{(x+3)(x-3)} = \frac{4(4)}{(4+3)(4-3)} = \frac{(+)}{(+)(+)} \end{aligned}$$

The signs can be summarized on the number line.



We interpret our analysis to find the domain of $h(x)$. The domain must include intervals where the inner expression is positive, $(-3, 0)$ and $(3, \infty)$, along with points where the expression equals zero, $x = 0$. The set is visualized below. We do not include the points where the expression was undefined, $x = \pm 3$. The domain is the set

$$D = (-3, 0] \cup (3, \infty).$$



□

2.1.3 Other Representations for Functions

A function can be defined by rules other than formulas. Any method that creates a unique output result for each given input value in the domain is a valid function. When the domain is a small finite set, we can simply define the output values with a table. We could also define a function through a graph of the pairs (x, y). Sometimes, a function can be defined according to an algorithm that is not described by a formula.

Example 2.1.17 One round of a game called "Pig" involves the throw of a single six-sided die. If the die shows one dot, the round scores 100 points. If the die shows five dots, the round scores 50 points. Any other face on the die results in 0 points.

The score for a round is a function of the thrown face. The input for the function, or independent variable, is the number of dots showing on the thrown die's face, say D. The output for the function, or dependent variable, is the score for the round, say S. The domain involves six possible values,

$$D = \{1, 2, 3, 4, 5, 6\}.$$

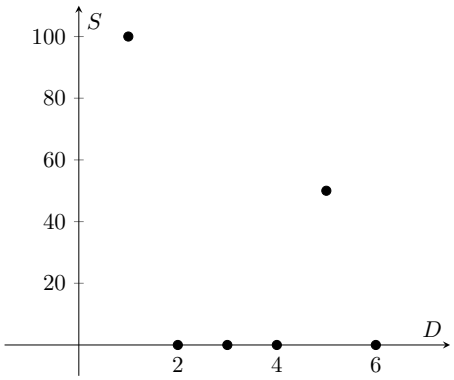
The range has three values, $R = \{0, 50, 100\}$.

The function can be characterized by a table. The table has two columns, one for the input (dots on the die) and one for the output (round score). The key feature of this table representing a function is that *every* value in the domain appears as exactly one entry in the input column. The function maps $D \mapsto S$ We can never allow a single input value to map to two different output values.

Number of Dots (D)	Score for Round (S)
1	100
2	0
3	0
4	0
5	50
6	0

A graph of this function has exactly six points, corresponding to the six values in the domain. Each point is a state (D, S) appearing in the table.

Notice how the graph and the table show the same information in different ways.



The inverse relation using the same table (S, D) is not a function. Multiple die throws correspond to the same score. Thus, knowing the score S is not enough information to know the value of the number of dots showing on a throw D. □

2.1.4 Summary

- A function is a relation between an independent variable (input) and a dependent variable (output) such that for each value of the input, there is exactly one value for the output.
- An equation in two variables defines a relation. When we can solve the equation for one variable (dependent) as a single expression of the other variable (independent), the expression defines an explicit function.
- Function mapping notation $x \overset{f}{\mapsto} y$ indicates that y is a function of x and f is the name of the function.
- Function evaluation notation $f(\square)$ uses substitution of whatever appears between the parentheses $(\square)$ in place of the independent variable.

2.1.5 Exercises

For each function defining a map between two variables, interpret the stated function value by indicating the value of each variable.

1. Suppose that P represents the population size in millions and B represents the birth rate in hundreds of births per month. If $f : P \mapsto B$, interpret $f(30) = 12$.

2. Suppose that t represents time in seconds and h represents the height of an object above the ground in meters. If $g : t \mapsto h$, interpret $g(2) = 3$.

3. Suppose that n represents the number of items a company will sell in thousands and p represents the price the company charges per item in dollars. If $h : p \mapsto n$, interpret $h(2) = 5$.

For each function, illustrate how the function maps values between variables. Draw two parallel number lines, with the top number line corresponding to the independent variable. Mark the given domain as shaded segments or points,

as appropriate. Choose three different values in the domain and then indicate with arrows how the function maps these values.

4. $f : [0, 1] \to \mathbb{R}; x \mapsto y = 2x - 5$

5. $g : [-1, 1] \to \mathbb{R}; s \mapsto T = \dfrac{1}{s+2}$

6. $h : 0, 1, 2, 3, 4, 5 \to \mathbb{R}; n \mapsto p = 10 - 2n$

7. $F : \mathbb{N} \to \mathbb{R}; t \mapsto x = 2^t$

Find the natural domain for each function defined by an equation.

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

9. $g(x) = \sqrt{x^2 - 4x + 3}$

10. $h(x) = \log\left(x^2 - 4x + 3\right)$

11. $f(x) = \dfrac{x^2 - 1}{x^2 - 4}$

12. $g(x) = \sqrt{\dfrac{x^2 - 1}{x^2 - 4}}$

13. $h(x) = \log\left(\dfrac{x^2 - 1}{x^2 - 4}\right)$

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

15. $g(x) = \sqrt{\dfrac{2x}{x^2 - x - 6}}$

16. $h(x) = \log\left(\dfrac{2x}{x^2 - x - 6}\right)$

17. A function $f : x \mapsto y$ is defined by a table shown below.

(a) Graph the points represented by the function.

(b) What are the domain and range of f?

(c) What is $f(3)$?

(d) What value or values of x satisfy $f(x) = 4$?

x	y
1	4
2	2
3	1
4	3
5	4

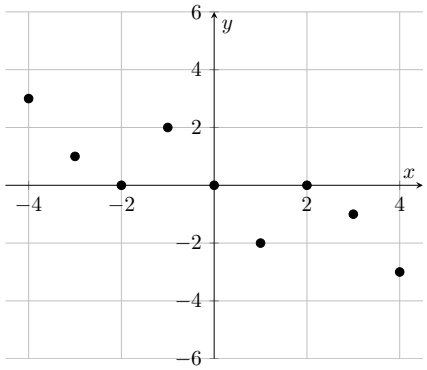
18. A function $g : x \mapsto y$ is defined by a graph shown below.

(a) What are the domain and range of g?

(b) What is $g(4)$?

(c) What value or values of x satisfy $g(x) = 3$?

(d) Find the table representation for g.



Applications

19. Let C be the temperature measured in degrees Celsius, and let F be the temperature measured in degrees Fahrenheit. The function $g(x) = \frac{9}{5}x + 32$ defines the map $g : C \mapsto F$, and $h(x) = \frac{5}{9}(x - 32)$ defines $h : F \mapsto C$.

(a) What is the value and interpretation of $g(30)$?

(b) What is the value and interpretation of $h(70)$?

20. A spring force scale uses the distance a spring is stretched to determine the force that is applied to the spring. We calibrate the scale by using known forces (e.g., weights) and record the corresponding location of the tip on a ruler. Let F be the force (Newtons) applied to the spring and let L be the corresponding location (centimeters). The following table is used for calibration.

F (N)	0	10.0
L (cm)	20.0	42.5

(a) Find a linear equation relating the variables F and L.

(b) Determine functions g and h so that $g : F \mapsto L$ and $h : L \mapsto F$. What are the corresponding equations using evaluation notation?

(c) Suppose a force of 5 N is applied to the spring. What will be the location of the tip of the ruler? Which function was used?

(d) Suppose a force is applied that results in the tip having a location of 28.7 cm. What was the force? Which function was used?

21. The cost C of materials for a project depends on the required area A of materials needed. The unit price is \$3.50 per m^2. The project involves making two squares, each of them having sides with length s (meters).

(a) Find $f : A \mapsto C$.

(b) Find $g : s \mapsto A$.

(c) How much would a project with $s = 4$ cost? How is each function used in order to answer this question?

2.2 Constructing Functions

Overview. We have learned that functions provide a map between two variables of a system. In modeling, the functions are almost always defined by formulas, with the dependent variable being equal to an expression involving only the independent variable. As we analyze these functions with calculus, the rules of computation for limits, derivatives, and integrals will depend on how a function is algebraically put together.

This section focuses on how expressions and functions are constructed. We start by reviewing elementary functions that represent basic operations on the independent variable. These will serve as the building blocks for our functions. We will then consider the basic arithmetic operations of addition, subtraction, multiplication, and division.

2.2.1 Elementary Functions

Every expression defining a function can be interpreted as a combination of various operations. Operations that act on a single expression are functions. Operations that combine multiple expressions include the binary arithmetic operations, particularly addition and multiplication. In order to characterize expressions, we first review the elementary operations that can be considered as elementary functions. We will consider an **elementary operation** to be a single operation on the variable.

The simplest operations are the constant functions and the identity function. As an operation, the constant function ignores the variable and always gives the same value for the output. The identity function, on the other hand, has no net change with the variable and returns an output that matches the input.

Definition 2.2.1 A **constant function** is a function that has the same output value for every input value, $f(x) = c$ for some constant c. ◊

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.2.2 The constant function $f(x) = 3$ as a map $x \overset{f}{\mapsto} 3$.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.2.3 The graph of the constant function $y = f(x) = 3$ in the (x, y) plane.

Definition 2.2.4 The **identity function** is a function where the output value is the same as the input value, $f(x) = x$. ◊

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.2.5 The identity function $f(x) = x$ as a map $x \overset{f}{\mapsto} x$.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.2.6 The graph of the identity function $y = f(x) = x$ in the (x, y) plane.

The four basic arithmetic operations of addition, subtraction, multiplication, and division can be used as functions. Because these binary operations

require two operands (the expressions being acted on), the elementary arithmetic operations will involve the variable and a particular constant.

For example, $x \mapsto x + 4$ is an elementary operation that adds the constant 4 to the independent variable. Similarly, $x \mapsto 4x$ is an elementary operation that multiplies the input by 4. Because subtraction is really addition with an additive inverse (the negation) of a number, an operation like $x \mapsto x - 4$ is equivalent to $x \mapsto x + -4$. Likewise, division is really multiplication with a multiplicative inverse (the reciprocal) of a number, so an operation like $x \mapsto x \div 4$ is equivalent to $x \mapsto \frac{1}{4}x$.

This motivates two new elementary operations: the constant sum and the constant multiple.

Definition 2.2.7 For every real number (constant) c, we can define the **constant sum** operation

$$x \mapsto x + c$$

and the **constant multiple** operation

$$x \mapsto cx.$$

◊

A constant sum represents a mapping that maintains a constant offset between the input and output. For example, the function $x \mapsto x - 3$ has an output that is always 3 units to the left of the input. We can think of the constant sum as a shift or translation. This mapping is illustrated in the following interactive figure.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.2.8 The constant sum $f(x) = x - 3$ as a map $x \overset{f}{\mapsto} x - 3$.

A constant multiple represents a mapping that maintains a constant scaling or ratio between the input and output. For example, the function $x \mapsto 2x$ has an output that is always twice the value of the input. We can think of the constant multiple as stretching or squeezing by a scale. This mapping is illustrated in the following interactive figure.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.2.9 The constant multiple $f(x) = 2x$ as a map $x \overset{f}{\mapsto} 2x$.

There are two more arithmetic operations possible with constants. Taking a constant and subtracting the variable, as in $x \mapsto 4 - x$, is not equivalent to a constant sum because we are not adding something to x. Similarly, dividing a constant by a variable, as in $x \mapsto 4 \div x$, is not equivalent to a constant multiple. These operations each involve two steps. The first step to each, however, introduces a new elementary operation.

Definition 2.2.10 The **negation** or additive inverse operation is the function $x \mapsto -x$, defined for all x. The **reciprocal** or multiplicative inverse operation is the function

$$x \mapsto \div x = \frac{1}{x},$$

defined for all $x \neq 0$. ◊

The negation operation maps a value x to its opposite value. This corresponds to a reflection on the numberline across zero.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.2.11 The negation $f(x) = -x$ as a map $x \overset{f}{\mapsto} -x$.

The reciprocal operation maps a value x to its multiplicative inverse. The product of a number and its inverse always equals 1. We could think of this operation as a multiplicative reflection across 1 for positive values and across -1 for negative values.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.2.12 The reciprocal $f(x) = \div x = 1/x$ as a map $x \overset{f}{\mapsto} \div x$.

The other elementary functions that we have studied can also be considered to be elementary operations. These include the basic powers and roots and exponentials and logarithms.

Definition 2.2.13 The **elementary power function** with power p is the function that raises the variable to a constant power,

$$\text{pow}_p(x) = x^p.$$

Because roots are also powers, roots are also elementary operations,

$$\text{pow}_p^{-1}(x) = \sqrt[p]{x} = x^{(1/p)}.$$

$\Diamond$

Definition 2.2.14 The **elementary exponential function** with base b, where $b > 0$ and $b \neq 1$ is the function that raises a constant base to the power of the variable,

$$\exp_b(x) = b^x.$$

Logarithms, as the inverses of exponentials, are included as elementary functions as well,

$$\exp_b^{-1}(x) = \log_b(x).$$

$\Diamond$

Additional elementary functions that we study later are the trigonometric functions. Trigonometric functions are used in relation to triangles as well as cyclic or periodic behavior. There are two fundamental trigonometric functions, the sine and cosine functions, from which the others are defined. We will study these functions in more depth later, but for the purpose of summary include the following table here.

$\sin(x)$	sine
$\cos(x)$	cosine
$\tan(x) = \dfrac{\sin(x)}{\cos(x)}$	tangent
$\sec(x) = \dfrac{1}{\cos(x)}$	secant
$\cot(x) = \dfrac{\cos(x)}{\sin(x)}$	cotangent
$\csc(x) = \dfrac{1}{\sin(x)}$	cosecant

The trigonometric functions are periodic, which implies that they must not be one-to-one. Inverse trigonometric functions are defined to solve equations

for a limited interval and provide additional elementary functions for our use.

$$\begin{aligned} \sin^{-1}(x) &= \arcsin(x) && \text{arcsine} \\ \cos^{-1}(x) &= \arccos(x) && \text{arccosine} \\ \tan^{-1}(x) &= \arctan(x) && \text{arctangent} \\ \sec^{-1}(x) &= \operatorname{arcsec}(x) && \text{arcsecant} \end{aligned}$$

The arccotangent and arccosecant functions can be defined but are not used in practice.

2.2.2 Algebraic Combinations and Composition

Functions defined by a formula are generally formed by combining these operations and functions into more complicated expressions. One of the most valuable skills in calculus is the ability to recognize how a formula is constructed. Many rules in calculus are named according to which operation forms the expression of interest. The basic operations of combination are the arithmetic operations of addition (a **sum**), subtraction (a **difference**), multiplication (a **product**), and division (a **quotient**) along with the operation of function **composition**.

Composition occurs whenever we apply a function or operation to an expression rather than a simple variable. That is, x^4 is a simple power operation, but $(2x+1)^4$ is a composition because the power acts on the expression $2x+1$. We use the arithmetic operations when we take two expressions and combine them. We use composition when we apply a function or operation to a single expression. The expression on which a composition acts is called the **input expression** or **inner expression**.

Most formulas involve more than one operation. An expression is classified by the *last* operation that would be applied. The order of operations determines the priority with which operations are applied. In algebra, you may have learned the acronym `PEMDAS`, which stands for Parentheses, Exponents, Multiplication, Division, Addition, and Subtraction. Subtraction is really the addition of an inverse, so differences can be classified as sums. The same technically applies for division being multiplication, but this is less frequently used. We will change the meaning of `E` to stand for *Every function*, including powers and exponentials, as all functions have higher precedence than the arithmetic operations.

Example 2.2.15 Classify each function by the last operation that is applied, and then classify each component expression. Make note of when a binary operation involves a constant instead of two variable expressions.

1. $f(x) = x^2 - 3x\sin(x)$
2. $g(x) = (2x+1)(x-3)$
3. $h(x) = (x^2+3)^4$
4. $j(x) = \dfrac{2xy}{\sqrt{3x-1}}$
5. $k(x) = 5e^{2x}$

Solution.

1. The function $f(x) = x^2 - 3x\sin(x)$ is a *difference* of the expressions x^2 and $3x\sin(x)$. The first component expression x^2 is a power function

($p = 2$) of x; the second component expression $3x\sin(x)$ is a product of $3x$ and $\sin(x)$. (We could also have used a sum of x^2 and $-3x\sin(x)$.)

2. The function $g(x) = (2x+1)(x-3)$ is a *product* of the expressions $2x+1$ and $x-3$. The first expression $2x+1$ is the constant sum of $2x$ and 1 while the second expression is the constant sum of x and -3.

3. The function $h(x) = (x^2+3)^4$ has the power ($p = 4$) as its last operation. Because we treat powers as functions, this is a composition. The inner expression is $u = x^2 + 3$, and the operation is the elementary power $\text{pow}_4(u) = u^4$. The inner expression is a sum of x^2 and the constant 3.

4. The function $j(x) = \dfrac{2xe^x}{\sqrt{3x-1}}$ is a *quotient* of expressions $2xe^x$ and $\sqrt{3x-1}$. The first expression $2xe^x$ is a product of $2x$ and e^x; the second expression $\sqrt{3x-1}$ is a square root (a function) of the expression $u = 3x - 1$, meaning this is a composition with the operation would be $\sqrt{u}$. We could also think of the square root as an elementary power function, $\sqrt{u} = \text{pow}_{1/2}(u) = u^{1/2}$.

5. The expression $5e^{2x}$ is a *constant multiple* of 5 with e^{2x}. the expression e^{2x} is a natural exponential function (base e) in composition, e^u, with the expression $u = 2x$.

□

Although binary operations like addition and multiplication are defined in terms of two operands, we often see them in expressions involving more than two terms, such as $a+b+c$ or $3xy$. By convention, the operations are performed left to right as $(a + b) + c$ or $(3x)y$. Because addition and multiplication are commutative and associative, this order doesn't matter; we act as if it were one sum or one product. In calculus, however, all of the rules are based on the binary nature of the operations. When classifying the structure of a formula, we should identify exactly two operands.

One of the most common ways to combine expressions in mathematics is to create a sum of constant multiples of those expressions. Such a combination is called a **linear combination**. The calculus operations of limits, integrals, and derivatives all satisfy a linearity in that they preserve linear combinations. It is therefore useful to recognize them.

Definition 2.2.16 Given a finite set of expressions, $u = (u_1, u_2, \ldots, u_n)$, and the same number of constants, $c = (c_1, c_2, \ldots, c_n)$, the **linear combination** of the expressions u with **coefficients** c is the sum of constant multiples of the expressions

$$c_1u_1 + c_2u_2 + \cdots + c_nu_n.$$

◊

The non-negative integer powers of x are the powers $x^0 = 1$, $x^1 = x$, x^2, x^3, etc. Linear combinations of non-negative integer powers establish a family of functions called polynomials.

Definition 2.2.17 Let n be a non-negative integer. A **polynomial** of **degree** n is a function that can be written in the form

$$f(x) = a_nx^n + \cdots + a_2x^2 + a_1x + a_0,$$

where $a_0, a_1, \ldots, a_n$ are constants called the **coefficients**. The term with the highest power a_nx^n is called the **leading term** and a_n is called the **leading coefficient**. A single term a_kx^k is called a **monomial** of degree k. A

polynomial with exactly two terms is called a **binomial**. ◊

Example 2.2.18 The polynomial $f(x) = 3x^3 + x^2 - 5x + 8$ is a linear combination of the powers $(1, x, x^2, x^3)$. The degree of the polynomial is $n = 3$, and the coefficients are $(c_0, c_1, c_2, c_3) = (8, -5, 1, 3)$. The leading coefficient is $c_3 = 3$. □

Example 2.2.19 Write down the polynomial $f(x)$ of degree $n = 4$ with coefficients $(c_0, c_1, c_2, c_3, c_4) = (16, 0, -8, 0, 1)$.

Solution. Because $c_1 = 0$ and $c_3 = 0$, we skip the terms with powers $x^1 = x$ and x^3. We usually write polynomials in decreasing powers, so we have

$$\begin{aligned} f(x) &= 1x^4 + 0x^3 + -8x^2 + 0x^1 + 16x^0 \\ &= x^4 - 8x^2 + 16. \end{aligned}$$

□

2.2.3 Models From Arithmetic

Understanding how functions are constructed also helps us develop models. When a quantity has contributions from multiple sources, we might create a model for each source and then add the contributions. Multiplication often combines factors that affect a single contribution. Division is used when the quantity of interest is defined as a ratio.

Example 2.2.20 Suppose that a population of an diploid organism has a trait characterized by a single gene. That gene has two alleles, a dominant allele A and a recessive allele a. The dominant trait will be present in two possible ways. Either the individual has two copies of the dominant allele (homozygous dominant) or the individual has one copy of each allele (heterozygous). If the population is subject to random mating that is independent of this trait, then the probability that an individual in the next generation will exhibit the dominant genotype can be calculated knowing the proportion of all alleles that are dominant.

Because there are two distinct ways to exhibit the dominant genotype, the probability of exhibiting the dominant genotype will be the *sum* of the probabilities of being homozygous dominant and heterozygous. This is often described as the sum rule of probability, which states that the probability of some outcome that can be attained through multiple pathways is the sum of the probabilities of each of the possible pathways. To calculate the probability of each pathway, we use a *product* rule associated with sequential events. When a pathway requires that a sequence of random outcomes occur, the probability of that individual pathway is the product of the probabilities of the individual outcomes along the pathway.

We can create a diagram showing all of the pathways by creating a decision tree. An individual receives one allele from each parent. Our tree will consider which allele is received from each parent. Let us call p the proportion of alleles in the current generation with the dominant allele. The remaining alleles must be recessive, and we call $q = 1 - p$ the proportion of alleles that are recessive. The probability associated with receiving an allele from a parent will be equal to the proportion of that allele in the population.

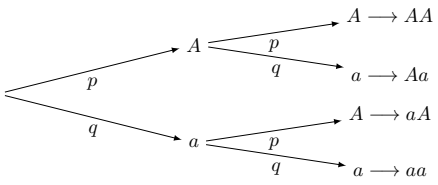


Figure 2.2.21 Tree showing inheritance of alleles from two parents.

There are three pathways that result in the dominant trait: AA, Aa, and aA. The probabilities associated with each pathway are p^2, $pq = p(1-p)$, and $qp = (1-p)p$, respectively. Consequently, the probability that an offspring will have the dominant trait will be

$$f(p) = p^2 + p(1-p) + (1-p)p.$$

The structure of this unsimplified formula reveals a direct relation to the tree. A slightly simplified version,

$$f(p) = p^2 + 2pq = p^2 + 2p(1-p),$$

combines the two pathways resulting in a heterozygous genotype. □

Example 2.2.22 Suppose a population of plants reproduces annually and is subject to density dependence. Density dependence typically results from the effects of competition and crowding with other individuals. The number of seeds each plant can produce is likely to depend on the population density. In addition, the probability that individual seeds will germinate and grow to maturity in the subsequent generation also depends on the population density. If we could characterize these dependencies as functions, then we could create a function that would predict the population size in a subsequent generation.

Let P_0 represent the population size of the current generation. The subscript 0 refers to the number of generations in the future. We wish to create a function $f : P_0 \mapsto P_1$, where P_1 is the population one generation in the future. Suppose that S_0 measures the average number of seeds produced by each plant in the current generation. The function $s : P_0 \mapsto S_0$ characterizes the dependence of seed production on the population size so that $s(P_0)$ gives the average number of seeds per plant in a population of size P_0. Now suppose that $\sigma(P_0)$ is another function that gives the success probability for an individual seed to survive to maturity coming from a population of size P_0.

We can use these elements to construct our function $f(P_0)$. The total number of seeds produced will be the current population size P_0 times the average number of seeds produced per plant. This means that $P_0 s(P_0)$ gives the total number of seeds produced. Not all seeds survive to maturity, so we multiply this by the success probability to give

$$f(P_0) = P_0\, s(P_0)\, \sigma(P_0).$$

Thus, the function used to project the subsequent generation's population size is constructed as a product of terms. If there were other ways that seeds could mature to new plants, we would add similar models for those other terms. □

Example 2.2.23 At the beginning of 2018, the US national debt was 20.493 trillion dollars. At the end of the year, the debt had risen to 21.974 trillion dollars. At the beginning of 2018, the US population was 326.2 million. A year later, the population was 328.2 million.

Develop a model for the per capita debt as a function of time, where per

capita debt is calculated as the ratio of the total debt to the total population size.

Solution. The per capita debt will be the total debt D (trillions of dollars) divided by the total population P (millions of individuals). To create a model, we need to make some modeling choices for $t \mapsto D$ and $t \mapsto P$, where t measures the year.

The simplest model might be to use linear functions for both. For a change in time $\Delta t = 1$ (year), we can see that

$$\Delta D = 21.974 - 20.493 = 1.481$$
$$\Delta P = 328.2 - 326.2 = 2.0$$

which are also slopes (dividing by $\Delta t = 1$ year). Consequently, our linear models for D and P are given by

$$D = 20.493 + 1.481(t - 2018)$$
$$P = 326.2 + 2.0(t - 2018)$$

The per capita debt according to this model will be approximated by

$$f(t) = \frac{D}{P} = \frac{20.493 + 1.481(t - 2018)}{326.2 + 2.0(t - 2018)}.$$

We expect that populations and debt grow exponentially. Consequently, an exponential model for our functions might be more appropriate. Using exponential models, $D = A\,b^t$ and $P = B\,a^t$, we use our data to find equations for the model parameters.

$$\begin{aligned} t = 2018 \quad &\Rightarrow \quad 20.493 = A\,b^{2018} \\ &\Rightarrow \quad 326.2 = B\,a^{2018} \\ t = 2019 \quad &\Rightarrow \quad 21.974 = A\,b^{2019} \\ &\Rightarrow \quad 328.2 = B\,a^{2019} \end{aligned}$$

We might use the 2018 equations to solve for A and B,

$$A = \frac{20.493}{b^{2018}}$$
$$B = \frac{326.2}{a^{2018}}$$

Then we substitute our results into the 2019 equations:

$$21.974 = \frac{20.493}{b^{2018}}\,b^{2019} = 20.493b$$
$$b = \frac{21.974}{20.493}$$
$$328.2 = \frac{326.2}{a^{2018}}\,a^{2019} = 326.2a$$
$$a = \frac{328.2}{326.2}$$

Our models can now be written down:

$$D = A\,b^t = \frac{20.493}{b^{2018}}\,b^t = 20.493\,b^{t-2018}$$
$$= 20.493\left(\frac{21.974}{20.493}\right)^{(t-2018)}$$

$$P = B\,a^t = \frac{326.2}{a^{2018}}\,a^t = 326.2\,a^{t-2018}$$
$$= 326.2\left(\frac{328.2}{326.2}\right)^{(t-2018)}$$

The function for the per capita debt is then calculated as a ratio,

$$f(t) = \frac{D}{P} = \frac{20.493\,b^{t-2018}}{326.2\,a^{t-2018}}$$
$$= \frac{20.493}{326.2}\left(\frac{21.974(326.2)}{20.493(328.2)}\right)^{t-2018}$$
$$\approx 6.2823 \times 10^{-2}\,(1.0657)^{t-2018}$$

Because we modeled the units of the debt as trillions of dollars and of the population as millions of individuals, the units for the per capita debt is in trillions of dollars per millions of individuals. To make sense of the units, it would help to go back to simple units of dollars and individuals. We would need to multiply D by 10^{12} to account for each debt unit representing a trillion dollars. Similarly, we multiply P by 10^6 to account for each population unit representing a million individuals. The per capita debt is the ratio, so we multiply the numerator by 10^{12} and the denominator by 10^6, with a net effect of multiplying by 10^6.

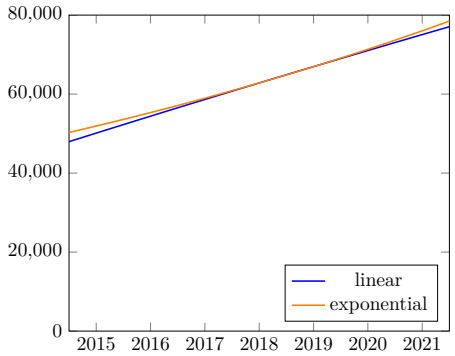


Figure 2.2.24 Models of US Per Capita Debt around 2018 in dollars per person.

□

2.2.4 Piecewise-Defined Functions

It is often the case that we use different models for different parts of the domain. When we introduced restricted domains, we defined functions by stating an inequality that specified the domain. For example, the equation

$$f(x) = x^2, \quad x \geq 0$$

defines a function with a domain $[0, \infty)$ based on the restriction $x \geq 0$. If we wanted a different rule for $x < 0$, say

$$f(x) = -x, \quad x < 0,$$

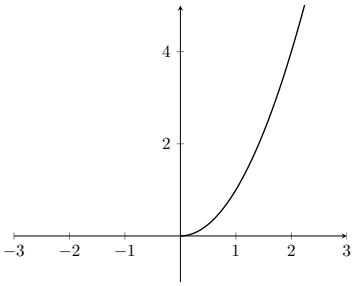
then the function now has domain $(-\infty, 0)$.

Functions that do this are called **piecewise-defined functions**. A piecewise-defined function allows us to specify rules on different parts of the domain. The

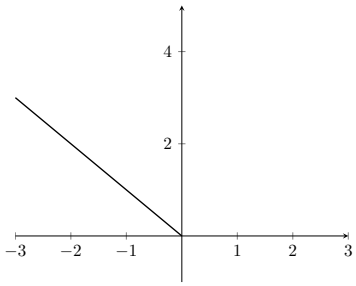
notation is similar to restricted domains, but we group all of the rules with a curly brace. The function

$$f(x) = \begin{cases} x^2, & x \geq 0, \\ -x, & x < 0 \end{cases}$$

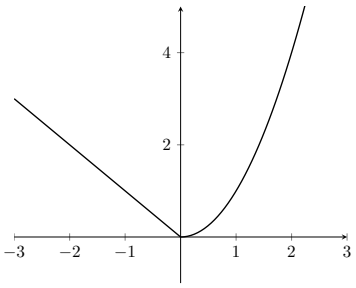
is defined for all values x so that the domain is $(-\infty, \infty)$.



(a) $f(x) = x^2$, $x \geq 0$



(b) $f(x) = -x$, $x < 0$



(c) $f(x) = \begin{cases} x^2, & x \geq 0, \\ -x, & x < 0 \end{cases}$

Figure 2.2.25 Comparison of functions with restricted domains and a piecewise-defined function.

Piecewise functions appear when there is a sudden change in behavior. The income tax structure in the United States is called a graduated tax because the tax rate increases as the amount of taxable income increases.

Example 2.2.26 For 2019, the first three IRS tax brackets for single individuals are as follows:

1. If taxable income is not over \$9700, then the tax is 10% of the taxable income.
2. If taxable income is over \$9700 but not over \$39475, then the tax is \$970 plus 12% of the excess over \$9700.
3. If taxable income is over \$39475 but not over \$84200, then the tax is \$4543 plus 22% of the excess over \$39475.

Create a piecewise function that calculates the tax given the taxable income.

Solution. The taxable income I is the independent variable. The "if" statements describing the taxable income levels describe the intervals of the domain. The first tax bracket is for $0 \leq I \leq 9700$, the second bracket is for $9700 < I \leq 39475$, and the third bracket is for $39475 < I \leq 84200$. Notice how the phrase "not over" is interpreted as including the stated value. The description of the tax amount uses percentages, which we will need to translate as a decimal multiplication. In addition, the phrase "excess over" will be

interpreted as subtraction. Putting the pieces together, we create the function

$$f(I) = \begin{cases} 0.10I, & 0 \le I \le 9700, \\ 970 + 0.12(I - 9700), & 9700 < I \le 39475, \\ 4543 + 0.22(I - 39475), & 39475 < I \le 84200. \end{cases}$$

The function could be extended further if we had additional information for the remaining tax brackets. □

The absolute value function is a particularly important mathematical function defined piecewise. For values that are negative, the absolute value returns the opposite (positive) value. For zero or for values that are already positive, the absolute value returns the original value.

Definition 2.2.27 The absolute value function is defined as

$$\text{abs}(x) = |x| = \begin{cases} -x, & x < 0, \\ x, & x \ge 0. \end{cases}$$

◊

2.2.5 Summary

- Functions defined by formulas are typically constructed from elementary functions: constant functions, the identity function, power functions, exponential functions, logarithms, and trigonometric functions.
- Combinations of expressions can be arithmetic (sum, difference, product, or quotient) or the composition of functions.
- An expression is classified by the *last* operation used to construct that expression.
- Binary operations involving a constant operand are special cases. They can be constructed using only constant sums, constant multiples, and reciprocals.
- A parametrized family of functions is a set of functions that have the same structure with different constants. The constants that can change are called parameters.
- Common parametrized families of functions are linear, exponential, and power functions.

Parametric Formula	Description
$f(x) = mx + b$	linear, slope-intercept
$f(x) = A\,x^p$	power
$f(x) = A\,b^x$	exponential, general base b
$f(x) = A\,e^{kx}$	exponential, natural base e

- A polynomial is a linear combination of simple powers $(1, x, x^2, \ldots, x^n)$, or, in other words, a sum of constant multiples of these powers,

  $$f(x) = a_n x^n + \cdots + a_2 x^2 + a_1 x + a_0.$$

  The constant multiples $(a_0, a_1, \ldots, a_n)$ are called the coefficients. The term $a_n x^n$ is called the leading term.
- A piecewise-defined function uses different rules for different portions of the domain.

2.2.6 Exercises

1. Classify each elementary function.

(a) $f(x) = \pi$

(b) $g(x) = x$

(c) $h(x) = x^{\pi}$

(d) $j(x) = \pi^{x}$

(e) $k(x) = \sin(x)$

2. Classify each function according to the last operation. Then classify the component expressions. Make note if the operation involves a constant expression.

(a) $f(x) = 4x^4$

(b) $g(x) = 2^{3x} + 5$

(c) $h(x) = 3^{5x-1}$

(d) $j(x) = 3\sqrt{x} + \dfrac{1}{x^2}$

(e) $k(x) = 4x^2 e^{3x}$

(f) $m(x) = \dfrac{x^2(3x-1)}{(x^2+1)^4}$

3. For each polynomial, determine the degree and list the coefficients.

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

(b) $f(x) = x^3 - 2x + 8$

(c) $f(x) = x^4 - 1$

(d) $f(x) = x^4 + 4x^3 + 6x^2 + 4x + 1$

Find the equation of the function $x \mapsto y$, if possible, for each of the following parametric models satisfying the states $(x, y) = (0, 3)$ and $(x, y) = (5, 9)$.

4. linear function

5. power function

6. exponential function

7. quadratic function of the form $y = a + bx^2$

8. quadratic function of the form $y = ax + bx^2$

Find the equation of the function $x \mapsto y$, if possible, for each of the following parametric models satisfying the states $(x, y) = (1, 3)$ and $(x, y) = (4, 6)$.

9. linear function

10. power function

11. exponential function

12. quadratic function of the form $y = a + bx^2$

13. quadratic function of the form $y = ax + bx^2$

14. a function the form $y = \dfrac{ax}{x+b}$

Use the description of each relation to create a corresponding piecewise-defined function.

15. The marginal tax rate is the percentage rate applied to the amount of taxable income that falls in the tax bracket. Based on the example, we see the marginal tax rate is 10% for income no greater than \$9700, 12% for income greater than \$9700 and no greater than \$39475, and 22% for income greater than \$39475 and no greater than \$84200. Define the function that takes the taxable income and returns the marginal tax rate for these three brackets.

16. Many bulk supplies are sold at a discount when enough items are purchased at once. An online gem store sells packages with two amethyst beads. If you purchase fewer than 15 packages, each package costs \$10.89. If you purchase at least 15 packages but fewer than 50, each package costs \$8.57. If you purchase at least 50 packages but fewer than 100, each package costs \$6.42. If you purchase at least 100 packages, each package costs \$5.87. Define the function that takes the number of packages ordered and returns the per package cost. Be clear about the domain.

17. For the gem example in Exercise 2.2.6.16, define the function that takes the number of packages ordered and returns the total cost of the order.

18. An electronic scooter can be unlocked for \$1.00 and then you are charged \$0.15 per minute of use. Partial minutes are rounded up to the next minute, so a rental of two minutes and fifteen seconds would be charged for three minutes or \$1.45 total. Define a piecewise function that gives the cost for rental times up to five minutes. What is the domain?

19. A car has a gas tank that holds 12 gallons and drives 35 miles per gallon. The owner starts with a full tank of gas, drives 300 miles, refills the tank, and then drives another 200 miles. Define a piecewise function that gives the amount of gas in the tank as a function of total distance traveled. What is the domain? Are there any ambiguities?

2.3 Chains and Function Composition

Overview. When we studied how formulas are constructed, we introduced the idea of function composition. Composition occurs whenever a function uses an expression rather than a simple variable as its input. In this section, we study the application of composition in terms of creating a chain of relationships between dependent variables. Given a complex formula that involves composition, we will learn to identify these chains. We will also consider applications of chains using formulas, graphs, and tables.

2.3.1 Composition and Chains of Relationships

An algebraic formula typically contains many different operations. If the formula involves a single variable, then that formula could be used to define a function. The function is the map that takes a value for the variable as its input and returns the value of the expression as the output. We could think of the function as a new operation; the formula provides the detailed instructions on how to perform that operation.

Composition occurs whenever the output of one function acts as the input to another function. For example, $f(x) = (\ln(x))^2$ takes the value of x, uses that as the input to ln, the natural logarithm, and then squares the result. This is a composition of the logarithm function with the squaring function. If we introduced the power function $\text{pow}_2(x) = x^2$, then the expression could be rewritten

$$(\ln(x))^2 = \text{pow}_2\big(\ln(x)\big).$$

The parentheses of function notation illustrate that the expression $\ln(x)$ acts as input to the pow_2 function.

Because a function should be interpreted as a map between an independent variable and a dependent variable, we can think of function composition as a chain of relationships between more than two variables. In our working example, suppose we introduce the dependent variable $y = (\ln(x))^2$. The calculation involves two steps: first we compute the logarithm, then we square the result. The result of that first operation is an intermediate variable, most commonly chosen as u, and we say $u = \ln(x)$. The final operation is an action applied to u, namely $y = u^2$. The chain of relationships could be expressed as the system of equations

$$\begin{cases} u = \ln(x), \\ y = u^2. \end{cases}$$

The logarithm is used as the map $x \overset{\ln}{\mapsto} u$ while the squaring function is the map $u \overset{\text{pow}_2}{\mapsto} y$. The overall calculation $f(x) = (\ln(x))^2$ provides the map $x \overset{f}{\mapsto} y$.

When functions used in a composition are named, a composition operation is represented by a small circle between their names. For our example, we would have $f(x) = \text{pow}_2 \circ \ln(x)$. Because the convention for function notation places the input on the right of the function name (and inside parentheses), the function name on the right of the circle is the inner function that defines the intermediate variable. The function name on the left of the circle is the outer function that completes the calculation.

Many formulas can be interpreted as compositions. To identify a composition, you need to be able to identify that the calculation performs some action on the result of an intermediate expression. However, you need to be careful that you are not using the original independent variable except in the intermediate expression. The intermediate expression itself can appear multiple

times. It is often helpful to try to write down a chain of variables to represent the composition. Start by identifying the intermediate expression and assign it to some intermediate variable, such as u. This defines the inner function. Then try to write the original expression only in terms of the new variable, substituting every instance of the intermediate expression by your variable. The resulting expression is the outer function.

Example 2.3.1 Express $y = \dfrac{2}{3(x-2)^2+1}$ as a composition.

Solution. There are multiple ways we could express our relation as a composition. We can interpret the order of operations as a sequence of operations acting on the original input x.

1. Take x.
2. Take the value and subtract 2.
3. Square the result.
4. Multiply the result by 3.
5. Take the result and add 1.
6. Divide 2 by the result.

Because each step in the operation takes the result of all prior steps, we could define the intermediate expression after any of the operations.

Suppose we define the intermediate operation to be all of the steps through squaring the result. Our intermediate variable defines our inner function,

$$u = g(x) = (x-2)^2.$$

The remaining steps in our description describe the outer function.

1. Take u.
2. Multiply the value by 3.
3. Take the result and add 1.
4. Divide 2 by the result.

As an equation, this becomes

$$y = h(u) = \frac{2}{3u+1},$$

which exactly corresponds to replacing the expression $(x-2)^2$ in the original equation with the intermediate variable u. The composition defines the equation

$$y = h \circ g(x).$$

Choosing a different expression for our intermediate variable results in a different choice for the composition. For example, if we had chosen an inner function to be

$$u = p(x) = x - 2,$$

then the outer function would need to be

$$y = r(u) = \frac{2}{3u^2+1}.$$

Thus, we also have $y = r \circ p(x)$. Similarly, if we had chosen

$$u = Q(x) = 3(x-2)^2 + 1$$

then our outer function would be

$$y = S(u) = \frac{2}{u}$$

to give $y = S \circ Q(x)$. □

Did you ever have to learn about an algebra topic **completing the square**? If so, did you find yourself asking the question "Why are we doing this?" One answer could be that a quadratic expression is not written in a way that it can be interpreted as a composition, but after completing the square it is.

Example 2.3.2 Consider the expression $x^2 + 6x - 3$. The expression is not a composition because it involves addition of two unrelated terms that involve the variable x. Completing the square is a strategy where the expression involving only the x^2 and x terms is recognized as matching the corresponding terms of a squared binomial term. In this case, because $6 \div 2 = 3$, we see that $x^2 + 6x - 3$ has the same x^2 and x terms as $(x+3)^2 = x^2 + 6x + 9$. Consequently, because we now know that

$$x^2 + 6x = (x+3)^2 - 9,$$

we can write

$$x^2 + 6x - 3 = (x+3)^2 - 12.$$

This new representation expresses our quadratic as the composition of three simple operations: adding 3, squaring, and subtracting 12. □

When we are given two functions and compute their composition, we use substitution to simplify our work. It is important to think about inputs and outputs of functions. Function notation is about substitution, using whatever expression appears as the input in place of the independent variable. Be careful that you don't think about multiplying by a function—we *apply* a function. Otherwise, you are liable to make algebra errors.

Example 2.3.3 Suppose $f(x) = 2x^2 - 1$ and $g(x) = 2x + 5$. Compute $f \circ g(x)$ and $g \circ f(x)$.

Solution. Because function notation has the input on the right, composition places the inner function to the right and the outer function to the left. We start with $f \circ g(x) = f\big(g(x)\big)$. Using the idea of a chain, we have an intermediate variable $u = g(x) = 2x + 5$. The composition asks for $f(u) = 2u^2 - 1$, substituting the independent variable with u. When we substitute the inner function in place of u, we get

$$f \circ g(x) = f(u) = 2(2x+5)^2 - 1.$$

Notice how the expression replacing u is placed inside parentheses.

Next, we find $g \circ f(x)$. We now have an intermediate variable $u = f(x) = 2x^2 - 1$. The outer function then compute $g(u) = 2u + 5$. Using substitution, this gives

$$g \circ f(x) = g(u) = 2(2x^2 - 1) + 5.$$

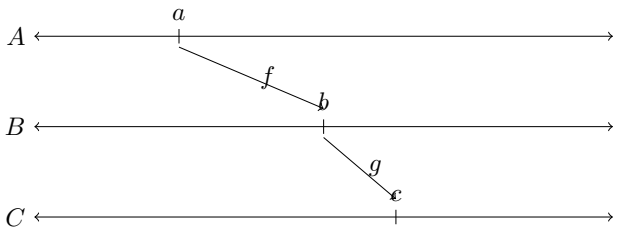
Once we have our expression using substitution, we could expand and simplify the results:

$$f \circ g(x) = 2(2x+5)^2 - 1 = 8x^2 + 40x + 49$$
$$g \circ f(x) = 2(2x^2 - 1) + 5 = 4x^2 + 1$$

This example should make it clear that the order of composition is important. □

2.3.2 Linking Maps through Chains

Composition corresponds to linking functions together, with the output of one function becoming the input to another function. In the context of a physical system, we are considering where the state involves multiple variables, say $(A, B, C, \ldots)$. Suppose we know one function, $f : A \mapsto B$, that determines the value of B knowing the value of A. Then suppose know another function, $g : B \mapsto C$, that predicts the value of C from the value of B. If we link these together in a chain, we can start with a value of A, compute the value of $B = f(A)$, and then use that value of B to compute the value of $C = g(B)$. Together, this composition creates a map $g \circ f : A \mapsto C$. Using substitution, we have $C = g\big(f(A)\big)$, the output of f becoming the input to g.



The following dynamic figure allows us to explore how composition links two functions together in a chain. The first (inner) function or map is $g : x \mapsto u = x + 3$. The second (outer) function is $f : u \mapsto y = u^2$. As you change the value of the input x, you can see where the functions map. The combined action $f \circ g(x) = (x + 3)^2$ represents a single function that is the composition of the steps.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.3.4 $f \circ g(x) = (x + 3)^2$

In the preceding subsection, we looked at composition in terms of formulas. Maps between variables can also be represented in tables and graphs. We can interpret composition by thinking through the relations between variables as we work through the linked maps.

Example 2.3.5 Suppose f and g are functions defined (at least partially) according to the following table. Find each of the following values.

1. $f \circ g(2)$
2. $g \circ f(2)$
3. $f \circ f(4)$
4. $g \circ g(0)$

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

Solution. When evaluating a function with the table, notice that the columns for $f(x)$ and $g(x)$ use an independent variable x. This means that we will find the input value in the x column and then find the corresponding out value in the function's column.

1. To find $f \circ g(2)$, we expand the substitution $f \circ g(2) = f\big(g(2)\big)$. The inner function is evaluated first to find $g(2) = 2$. That is, we find 2 in the column for x (placeholder for the input), and looking in the column of $g(x)$ we find 2 as the output. This output is used in the chain linking the function as the input for f,

$$f\big(g(2)\big) = f(2) = -2.$$

2. To find $g \circ f(2)$, we will expand $g \circ f(2) = g\big(f(2)\big)$. We start with the inner function $f(2) = -2$. We then use the output as the input of the outer function, $g(-2) = -1$. Consequently,

$$g\big(f(2)\big) = g(-2) = -1.$$

3. Continuing this pattern, we have

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

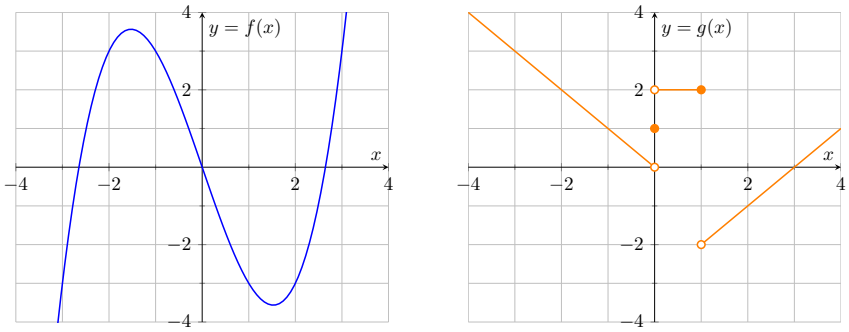
4. Similarly, we have

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

□

Example 2.3.6 Suppose f and g are functions with graphs as shown below. Find each of the following values.

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



Solution. To evaluate a function when given a graph, we find the value of the input along the x-axis. This is analogous to looking in the column of x in a function's table. Once we find the x-value, we imagine a vertical line at that point and find the point that is included in the graph that intersects our line. If there is no point included in the graph, then the value of x is not in the domain. For composition, we will use the original input to evaluate the inner function. Once we know the output value, we will use that result when we evalute the outer function.

1. To evaluate $f \circ g(2) = f\big(g(2)\big)$, we evaluate $g(2)$ using the graph on the right. The vertical line at $x = 2$ intersects our graph at the point $(2, -1)$, so we have $g(2) = -1$. Using that output as the input of f, we find $x = -1$ on the x-axis of the graph on the left. The vertical line intersects that graph at $(-1, 3)$ so that $f(-1) = 3$. Putting this together gives

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

2. When we evaluate $g \circ f(2)$, we repeat the process but with f as the inner function and g as the outer function.

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

3. When we evaluate $f \circ f(2)$, the same function is used as the inner and outer function. This means that when we find the output $f(2) = -3$, we use the same function to evaluate $f(-3) = -3$. Consequently, we have

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

4. When an expression has more than one composition, we proceed through the chain from the inside out. The expression $f \circ g \circ g(0)$ has an innermost function g. The vertical line $x = 0$ intersections the graph of g at the point $(0, 1)$. The open points at $(0, 0)$ and $(0, 2)$ represent end-points of the graph segments immediately to the left and to the right of the point but are not included as actual points. Consequently, $g(0) = 1$ which is now the input of the next g operation. We have

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

□

2.3.3 Applications of Composition

In modeling settings, composition arises in the context of chains of related variables. Whenever we model relationships between variables, we ideally cre-

ate functions to describe these relations. A chain occurs when we know a relation between, say, A and B and another relation between B and C. Each of these relations might be based on observations or experiments. The chain allows us to identify a relation between A and C, even if a direct observation or experiment is not possible or convenient.

Example 2.3.7 The radius r, circumference C, and area A of a circle are all related. The equation $C = 2\pi r$ defines C as a function of r and the equation $A = \pi r^2$ defines A as a function of r. Use composition to define the function $C \mapsto A$.

Solution. The final output should be area A, which we can compute if we know r. We can use the relation between C and r to solve for r as a function of C. That is, we want to create a composition of $C \mapsto r$ and $r \mapsto A$.

$$C = 2\pi r$$
$$\frac{C}{2\pi} = r$$

This equation defines $C \mapsto r$, so that we have a chain

$$r = \frac{C}{2\pi},$$
$$A = \pi r^2.$$

Composition corresponds to substitution of r by its formula,

$$A = \pi r^2 = \pi\Big(\frac{C}{2\pi}\Big)^2 = \frac{\pi C^2}{2^2\pi^2} = \frac{C^2}{4\pi}.$$

□

Example 2.3.8 Suppose you are blowing up a balloon with air. What is the radius of the balloon as a function of time?

Solution. The question is intentionally somewhat vague in order to illustrate the modeling process. Without more information, the question is ill-posed and there is not a clear answer. What simplifying assumptions could we make that will allow us to create a reasonable answer?

1. What shape is the balloon? We could make a simplifying assumption that it is approximately a sphere.

2. How fast is air being added? If we pretend to blow air in a balloon, we can time how long each breath takes. With a quick internet search, we can discover the typical amount of air blown per breath.

3. Keep things steady. It adds complications to the process if we try to account for inhaling between breaths or slowing down because we are tired. Let us replace human breaths blowing up the balloon with a model that would correspond to steady airflow that matches the average rate of filling.

With our assumptions identified, we can start to establish our model using equations. A sphere has a relationship between volume and radius according to the equation $V = \frac{4}{3}\pi r^3$. Because we want the radius r as the final output variable, we need $V \mapsto r$ which we find by solving for r.

$$r = \sqrt[3]{\frac{3V}{4\pi}} = \left(\frac{3V}{4\pi}\right)^{1/3}.$$

Our information about filling the balloon will give us a model for $t \mapsto V$. I found a result showing that there is about 1/2 liter of air exhaled in a breath, which corresponds to $500\,\text{cm}^3$. With a timer, I approximated that each steady blow takes about 5 seconds, including inhalation to prepare for the next breath. This means that the balloon is gaining $100\,\frac{\text{cm}^3}{\text{s}}$. If air flows at a steady rate, the relation $t \mapsto V$ is linear and starts at $V = 0$ when $t = 0$. This gives us our second function in the chain,

$$V = 100t.$$

The composition of the chain $t \mapsto V \mapsto r$ will give us a model for $t \mapsto r$, which we create using substitution:

$$r = \sqrt[3]{\frac{300t}{4\pi}}.$$

□

Our examples have included questions where a function was composed with itself. This actually occurs in practical settings, such as where a function maps from the value of some quantity to the value of the same quantity at a later time.

Example 2.3.9 A population's growth and decline depend on how the population size relates to its carrying capacity. If the population is below its capacity, then abundance of resources will lead the population to grow. If the population is too large, then physical constraints will cause the population to decline. Mathematical ecologists study possible behaviors for populations through the use of **projection functions**. A projection function maps the value of the population size at one time to the next observed population size. That is, knowing the size of the population this year, a projection function allows us to predict the population size next year. Composition of the function with itself then allows us to predict two years away.

Suppose the size of a population (in thousands) has been modeled by an annual projection function $f(x) = 1.6x - 0.32x^2$. If the population is currently 400, what will it be next year? in two years? What is the function that projects the population size two years from the present?

Solution. A population of 400 corresponds to a current population value $x = 0.4$ (thousands). The projection function uses this value to predict one year into the future. When we evaluate the function, we find

$$f(0.4) = 1.6(0.4) - 0.32(0.4)^2 = 0.5888,$$

corresponding to a population prediction of 588.8. If we use the function again with an input $x = 0.5888$, the function will predict one year from next year, or two years away. This gives

$$f(0.5888) = 1.6(0.5888) - 0.32(0.5888)^2 \approx 0.83114.$$

The model therefore predicts approximately 589 individuals in one year and 831 individuals the next year. (The calculation stays exact; the interpretation rounds.)

We found the projected population in two years through composition $f \circ f(0.4) \approx 0.83114$. The process of computation would be the same for any current population value x. Consequently, we can create a function that projects the population size in two years by computing the composition $f \circ f(x)$ using substitution.

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

$$= f(1.6x - 0.32x^2)$$
$$= 1.6(1.6x - 0.32x^2) - 0.32(1.6x - 0.32x^2)^2$$

We have replaced each x in the formula $1.6x - 0.32x^2$ with the expression $1.6x - 0.32x^2$. We can use a computer to help expand and simplify this algebraic formula,

$$f \circ f(x) = 2.56x - 1.3312x^2 + 0.32768x^3 - 0.032768x^4.$$

```
f(x) = 1.6*x - 0.32*x^2
show( f(f(x)).expand().simplify() )
```

□

2.3.4 Summary

- A chain of related variables is where knowing A you can predict B, and knowing B you can predict C, and so on. Composition is using A and the chain to find C.
- Composition $f \circ g$ is evaluation of the outer function f with an input using the output of the inner function g,

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

  As maps, if $g : x \mapsto u$ and $f : u \mapsto y$, then

  $$x \overset{f\circ g}{\mapsto} y \quad = \quad x \overset{g}{\mapsto} u \overset{f}{\mapsto} y.$$

2.3.5 Exercises

Rewrite each function as a nontrivial composition of two functions. (Nontrivial means that neither function should be the identify function $x \mapsto x$.)

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

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

3. $f(x) = 4e^{-x^2}$

4. $f(x) = 2\sin(3x) + 1$

5. $f(x) = \dfrac{2}{(e^x + 1)^2}$

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

7. $f(x) = \sin^2(x) + 4\sin(x) + 3$

8. $f(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$

Using the given functions, compute and simplify the expressions listed.

9. Given $p(x) = x^2 - 1$ and $r(x) = 2x + 1$.

(a) $p \circ r(x)$

(b) $r \circ p(x)$

(c) $p \circ p(x)$

(d) $r \circ r(x)$

10. Given $k(x) = e^x$ and $h(x) = 1 - x^2$.

(a) $k \circ h(x)$

(b) $h \circ k(x)$

(c) $k \circ k(x)$

11. Given $C(x) = \dfrac{2}{x-3}$ and $D(x) = e^{1/x}$.

(a) $C \circ D(x)$

(b) $D \circ C(x)$

(c) $D \circ D(x)$

12. Given f and g defined by the table below and $h(x) = 2x - 1$.

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

(a) $f \circ g(2)$

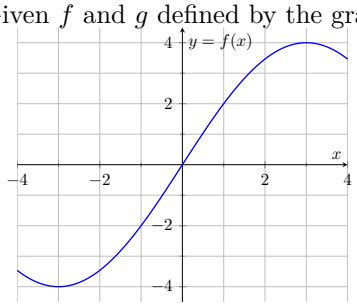
(b) $g \circ f(-2)$

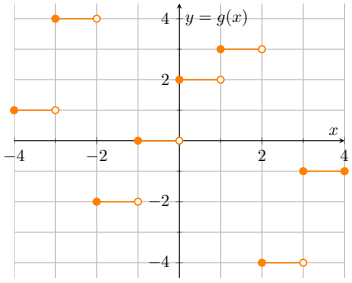
(c) $f \circ f(-2)$

(d) $f \circ h(2)$

(e) $h \circ g(1)$

13. Given f and g defined by the graphs below.





(a) $f \circ g(3)$

(b) $g \circ f(3)$

(c) $f \circ g(1)$

(d) $g \circ f(1)$

(e) $f \circ g(-3.25)$

(f) $g \circ f(-1.25)$

Applications

14. The perimeter P and area A of a square are each functions of the length of the sides s by $P = 4s$ and $A = s^2$. Find perimeter as a function of area, $P \mapsto A$.

15. The volume of a sphere is related to the radius of the sphere by the equation $V = \frac{4}{3}\pi r^3$. Suppose the radius is a function of time defined

by $r = 1 + 2t$. Find the volume as a function of time, $t \mapsto V$.

16. The cost C of materials for a project depends on the required area A of materials needed. The unit price is \$3.50 per m^2. The project involves making two squares, each of them having sides with length s (meters).

(a) Find $A \overset{f}{\mapsto} C$.

(b) Find $s \overset{g}{\mapsto} A$.

(c) Use composition to find $s \mapsto C$. Is this $f \circ g$ or $g \circ f$?

(d) How much would a project with $s = 4$ cost? How much area of materials will be required? What function is used for each calculation?

17. The density of plants (number of plants per square meter) on a plot of land from year to year has been modeled by the projection function $f(x) = 2.8x - 0.18x^2$. The plot in the current year is observed to have 3.50 plants per square meter.

(a) What is the predicted density of plants in one year?

(b) What is the predicted density of plants in two years?

(c) What is the predicted density of plants in three years?

(d) Find the function that predicts the density of plants two years from the present.

2.4 Inverse Functions

Overview. A function defines a map from one variable A to another variable B, $A \mapsto B$. In the context of the function, we call A the independent variable and call B the dependent variable. Knowing A, the function provides the rule to determine B. There are times when knowing B, we wish to find the value of A. This corresponds to using the function to solve an equation. When each value of B results in only a single value of A, the relation defines a new function $B \mapsto A$. In this case, the inverse relation is called the **inverse function**.

This section discusses the general concept of inverse functions. We will learn to compute inverse functions for given functions by solving equations and by interpreting tables and graphs. Because not all functions are defined by equations that can be solved, the definition of an inverse function will need to be more general and will involve function composition. We will identify attributes of functions that indicate if an inverse function exists. The calibration of instruments using standardized measurements will illustrate a practical application of inverse relations.

2.4.1 Finding Inverse Functions

When we think of a function as a map between variables, say $f : A \mapsto B$, we think of f as the rule that goes from an input value on the A number line to a corresponding predicted output value on the B number line. An **inverse function** would be a rule that goes in the reverse direction, $B \mapsto A$. A function and its inverse function allow us to go back and forth between the two variables in either direction.

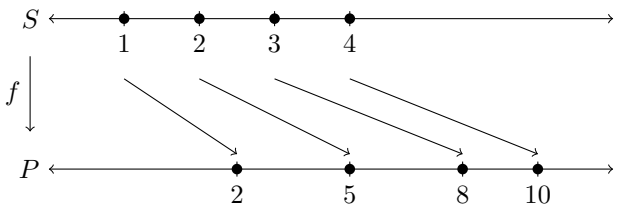
Definition 2.4.1 A function representing a map $f : A \mapsto B$ has an **inverse function**, which we write $f^{-1} : B \mapsto A$, if the equation $f(a) = b$ is equivalent to $f^{-1}(b) = a$ for every state $(A, B) = (a, b)$. $\Diamond$

We first illustrate the idea of an inverse function with a function defined by a simple map and no formula.

Example 2.4.2 Imagine a theater that has a promotional wheel so that the price of a ticket is based on which number you spin. The prices are listed in the table below.

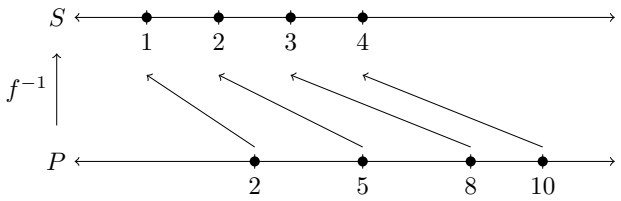
Spin	Price
1	\$2
2	\$5
3	\$8
4	\$10

We introduce variables for the system. Let S represent the result of the spin and let P represent the price of a ticket. The map $f : S \mapsto P$ can be visualized using number lines. It represents the idea that if you know what spin was achieved, then you will be able to know the price of the tickets.



The inverse map $f^{-1} : P \mapsto S$ reverses the direction of the arrows. The inverse indicates that knowing the price of the tickets is enough information to

know the result of the customer's spin.



Based on this system, we see that $f(1) = 2$ because when $S = 1$ we have $P = 2$. The equivalent inverse equation is $f^{-1}(2) = 1$ because a price $P = 2$ comes from $S = 1$. Similarly, $f(2) = 5$ and $f^{-1}(5) = 2$ are equivalent. Because the system is defined by the table and not a formula, $f(5)$ and $f^{-1}(3)$ each have no meaning. In the first case, $f(5)$ has no meaning because $S = 5$ is not a possibile spin. In the second case, $f^{-1}(3)$ has no meaning because $P = 3$ is not a possible ticket price. □

You might have realized a possible problem. What happens if two input values map to the same output value? We wouldn't know which arrow to follow for the reverse mapping. A function that guarantees that different input values always have different output values is called **one-to-one**. A function that is *not* one-to-one has at least one value that is the output to two or more different input values.

Theorem 2.4.3 *If a function $f : A \mapsto C$ is one-to-one, then the inverse $f^{-1} : C \mapsto A$ is also a function. If f is not one-to-one, then the inverse relation is not a function.*

When an algebraic equation defines the relation between the variables, we can attempt to solve the equation for either of the variables. If both variables can successfully be written as dependent variables, the corresponding formulas define the inverse functions.

Example 2.4.4 A rope of length 100 centimeters is cut into exactly five pieces Two of the pieces are of one length, and the other three pieces are of another length. Let d be the length of the ropes in the group of two. Let t be the length of the ropes in the group of three.

Find the functions $f : d \mapsto t$ and $g : t \mapsto d$. Interpret the meaning of $f(10)$ and $g(10)$.

Solution. We start by finding an equation relating d and t. The total length of the five pieces of rope added together must equal the original length of rope. This results in an equation

$$2d + 3t = 100.$$

With this equation, we can solve for each of the state variables in turn. Solving for d, we subtract $3t$ and divide by 2:

$$d = \frac{100 - 3t}{2}.$$

Solving for t, we subtract $2d$ and divide by 3:

$$t = \frac{100 - 2d}{3}.$$

The function f was defined as the map $d \mapsto t$, so we use the equation with t as the dependent variable,

$$t = f(d) = \frac{100 - 2d}{3}.$$

We can find and interpret $f(10)$. With 10 as an input, we find $f(10) = \frac{100-2(10)}{3} = \frac{80}{3} = 26\frac{2}{3}$. To interpret this, we recall that the input represents a value for d. The equation represents the state $d = 10$ and $t = 26\frac{2}{3}$ If the group of two has length 10 centimeters, then the group of three has length $26\frac{2}{3}$ centimeters.

The function g was defined as the map $t \mapsto d$, so we now use the equation with d as the dependent variable,

$$d = g(t) = \frac{100 - 3t}{2}.$$

With $t = 10$ as an input, we find $g(10) = \frac{100-3(10)}{2} = \frac{70}{2} = 35$. The function tells us that $d = 35$ when $t = 10$. The group of two has length 35 centimeters whenever the group of three has length 10 centimeters.

The functions f and g are inverse functions to each other. If we used a placeholder variable instead of the state variables, we would write

$$f(x) = \frac{100 - 2x}{3}$$

with its inverse function

$$f^{-1}(x) = g(x) = \frac{100 - 3x}{2}.$$

Similarly, $g^{-1}(x) = f(x)$. □

A function might be defined through a graph. You may remember something about the graph of the inverse being a reflection of the graph of the original. The following example will help clarify where that idea originates.

Example 2.4.5 Consider the function defined by the following table. Create a table representing the inverse function. Compare the graphs of the function and its inverse.

x	-4	-3	-2	-1	0	1	2	3	4
$f(x)$	4	2	1	0.5	0	-0.5	-1	-2	-4

Solution. Functions represent maps between variables, so let us say that $f : A \mapsto B$. This gives us a physical interpretation of the values in the table. The row for x corresponds to values of the input A. The row for $f(x)$ correspond to values of the output B. If we were to relabel our table with our variables, we would create the following table.

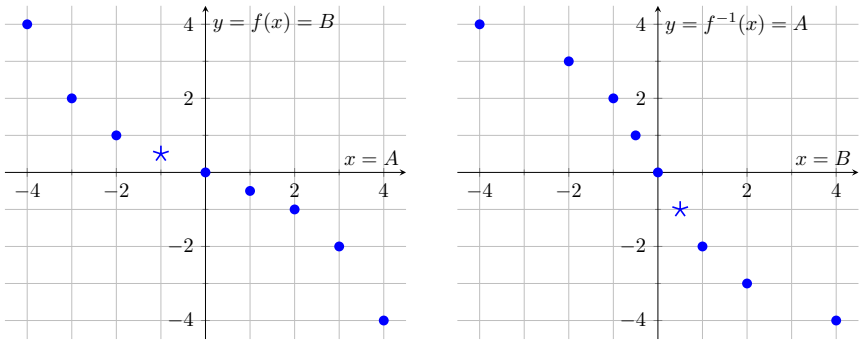
A	-4	-3	-2	-1	0	1	2	3	4
B	4	2	1	0.5	0	-0.5	-1	-2	-4

An inverse function f^{-1} would be the map $B \mapsto A$. The row associated with B now represents the independent variable while the row for A represents the dependent variable. Because we usually sort the values of the independent variable, we would reorder the columns while keeping the states for (A, B) together.

$x = B$	-4	-2	-1	-0.5	0	0.5	1	2	4
$f^{-1}(x) = A$	4	3	2	1	0	-1	-2	-3	-4

A graph of the function is formed by points created from ordered pairs. The graph $y = f(x)$ corresponds to points $(x, y) = (A, B)$, because $f : A \mapsto B$. The graph $y = f^{-1}(x)$ corresponds to points $(x, y) = (B, A)$, because $f^{-1} : B \mapsto A$. For example, the state $(A, B) = (-1, 0.5)$ corresponds to the point $(x, y) = (-1, 0.5)$ on the graph $y = f(x)$ and to the point $(x, y) = (0.5, -1)$ on

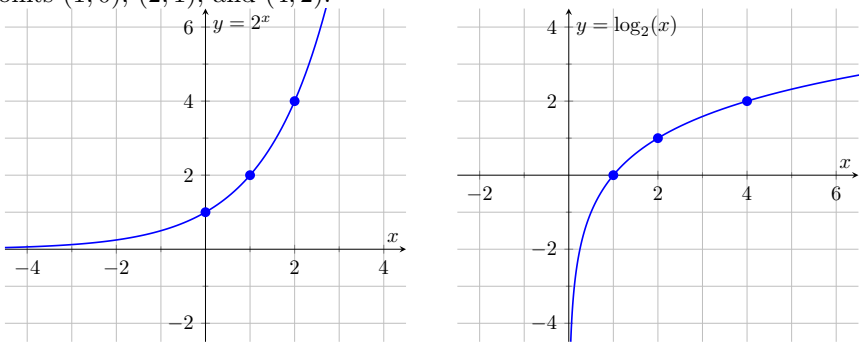
the graph $y = f^{-1}(x)$. This state is highlighted using a star in the graphs of the two functions below.



□

From our example, we see that the graph of an inverse function takes each point on the graph of the original function and reverses the role of the coordinates. Reversing the coordinates of every point in the graph geometrically corresponds to reflecting the graph across the line $y = x$. This result can help us remember graphs of inverse pairs.

Example 2.4.6 The graph of $y = 2^x$ is fairly easy to construct. It is exponential growth that doubles every integer increment of x. So we include, for example, the points $(0, 1)$, $(1, 2)$, and $(2, 4)$. The inverse function of the exponential is the logarithm with base $b = 2$. The graph of the logarithm is the reflection of the exponential graph across $y = x$ and includes the corresponding points $(1, 0)$, $(2, 1)$, and $(4, 2)$.



□

Now consider a function that is defined as a composition of operations. When we solve the equation to find the inverse function, we will discover that the inverse corresponds to applying the inverses of the original operations in the reverse order.

Example 2.4.7 Consider the function $f(x) = \dfrac{5}{3x + 2}$ corresponding to a map

$$a \overset{f}{\mapsto} b = \frac{5}{3a + 2}.$$

The inverse function is found by solving for a in the equation. Cross-multiplying the equation gives

$$b(3a + 2) = 5.$$

Dividing by b then gives

$$3a + 2 = \frac{5}{b}$$

from which we get

$$a = \frac{\frac{5}{b} - 2}{3}.$$

Although we would normally simplify this equation, we have successfully solved for a and so have a formula for the inverse,

$$b \overset{f^{-1}}{\mapsto} a = \frac{\frac{5}{b} - 2}{3}.$$

To simplify the fraction so that there is not a fraction in the numerator, we can multiply top and bottom by the value b giving

$$a = \frac{5 - 2b}{3b} = \frac{5}{3b} - \frac{2}{3}.$$

Using a placeholder variable, like x, instead of the physical variables a and b, the pair of inverse functions are

$$f(x) = \frac{5}{3x + 2},$$
$$f^{-1}(x) = \frac{\frac{5}{x} - 2}{3} = \frac{5}{3x} - \frac{2}{3}.$$

□

Let us consider the operations involved in the previous example. The function $f(x) = \dfrac{5}{3x + 2}$ involved the following sequence of operations:

1. Take the value of x.
2. Multiply by 3.
3. Add 2.
4. Divide 5 by the result.

The inverse function $f^{-1}(x) = \dfrac{\frac{5}{x} - 2}{3}$ did the inverse operations in the reverse order:

1. Take the value of x.
2. Divide 5 by the value.
3. Subtract 2.
4. Divide by 3.

This should make sense. Solving the equation for the original independent variable is accomplished by starting at the end and working backwards. This is summarized in the following theorem. The proof captures the idea of solving the equation by working in reverse.

Theorem 2.4.8 *Suppose $f(x) = g \circ h(x)$ and g and h each have inverse functions. Then $f^{-1}(x) = h^{-1} \circ g^{-1}(x)$.*

Proof. The composition corresponds to a chain. Suppose the independent variable is a and the ultimate dependent variable is c so that $f : a \mapsto c$. Then there is an intermediate variable $b = h(a)$ so that $c = g(b)$. The inverse function f^{-1} is the map going from c to a, $f^{-1} : c \mapsto a$. Because g has an inverse and $c = g(b)$, we can apply $c \mapsto b = g^{-1}(c)$. Then, because h has an inverse and $b = h(a)$, we can similarly apply $b \mapsto a = h^{-1}(b)$. Combining the chain, we

have

$$a = h^{-1}(b) = h^{-1}\big(g^{-1}(c)\big) = h^{-1} \circ g^{-1}(c).$$

The result follows by using a generic independent variable x. ■

2.4.2 Inverse Functions and Composition

When we discussed inverse functions earlier in (((Unresolved xref, reference "subsection-inverse-functions"; check spelling or use "provisional" attribute))) , we thought of them as inverse maps. Given an equation defining the map $x \mapsto y$, if we could solve the equation for the input x as a single expression involving y, then this new equation defined the inverse function. Inverse functions undo one another's operations.

Let us consider the calculations involved in the previous example. The function f took an input and performed the following operations in order:

- Multiply by 3.
- Add 2.

The inverse function f^{-1} took an input and performed related operations:

- Subtract 2.
- Divide by 3.

The functions are inverse because they will exactly undo one another's operations.

Consider what happens if you create a chain and apply f^{-1} immediately after f:

- Multiply by 3.
- Add 2.
- Subtract 2.
- Divide by 3.

The middle two steps cancel one another's effects, so this would be the same as the simpler chain of steps:

- Multiply by 3.
- Divide by 3.

Again, the operations cancel each other out. The output will always be the same as the original input,

$$f^{-1} \circ f(x) = x.$$

The following interactive figure shows this composition as a chain of maps.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.4.9 Composition $y = f^{-1} \circ f(a)$, corresponding to chain $a \overset{f}{\mapsto} b \overset{f^{-1}}{\mapsto} y$. As the functions are inverses, this always yields $y = a$.

A composition in the reverse order, $f \circ f^{-1}(x)$, also results in exact cancellation.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.4.10 Composition $y = f \circ g(b)$, corresponding to chain $b \overset{g}{\mapsto} a \overset{f}{\mapsto} y$. As the functions are inverses, this always yields $y = b$.

Simplifying the composition of inverse functions algebraically reveals the cancellation directly.

Example 2.4.11 For $f(x) = 3x+2$ and $f^{-1}(x) = \dfrac{x-2}{3}$ compute and simplify $f \circ g(x)$ and $g \circ f(x)$.

Solution. Using substitution and algebraic simplification, we find the values requested.

$$\begin{aligned} f \circ g(x) &= f\big(g(x)\big) \\ &= f\Big(\frac{x-2}{3}\Big) && \text{substitute } g(x) \\ &= 3\Big(\frac{x-2}{3}\Big) + 2 && \text{substitute } f(\square) \\ &= x - 2 + 2 = x \end{aligned}$$

$$\begin{aligned} g \circ f(x) &= g\big(f(x)\big) \\ &= g\big(3x+2\big) && \text{substitute } f(x) \\ &= \frac{(3x+2)-2}{3} && \text{substitute } g(\square) \\ &= \frac{3x}{3} = x \end{aligned}$$

□

Inverse functions will always simplify in this way: the composition of inverse functions cancel to just leave the input. Functions are not always defined by an equation, so we shouldn't define inverses through solving equations. Mathematicians actually define inverse functions in terms of the property of composition.

Definition 2.4.12 Two functions f and g are inverses of one another, and we write $g = f^{-1}$ and $f = g^{-1}$, if for every x in the domain of g, we have

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

and for every x in the domain of f, we have

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

◊

It is time for a comment about real variables. In science, variables represent physical measurements and the variables are the objects of study. These variables can be related by functions. However, in mathematics, it is the function itself that is being studied. For simplicity, mathematics textbooks have adopted an approach where x is almost universally the independent variable of every function and y is the dependent variable. This makes it easier to remember the role each variable plays, but it can lead to confusion in actual applications.

Example 2.4.13 An enzyme is a protein that helps catalyze a chemical reaction. For many enzymes, the rate of reaction R and the concentration of the reactant C satisfy a relation called Michaelis-Menten kinetics

$$R = \frac{aC}{C+K},$$

where a and K are parameters that characterize the particular reaction. Physically, we require $C \geq 0$. In mathematics, this relation might be characterized by a function

$$f(x) = \frac{ax}{x+K}.$$

We would then say $R = f(C)$. This is equivalent to mapping notation

$$C \overset{f}{\mapsto} R = \frac{aC}{C+K}.$$

To find the inverse function, most mathematics textbooks say to write $y = f(x)$, switch all x and y and then solve for y. The only reason to switch the variables is to preserve x as the independent variable of the relation. This is an artificial requirement. We might as well just solve for C as a function of R. Start by cross-multiplying to eliminate the denominator in the equation.

$$\begin{aligned} R &= \frac{aC}{C+K} \\ R(C+K) &= aC \\ RC + KR &= aC \end{aligned}$$

Because we are solving for C, we need to collect C terms on one side of the equation and then factor.

$$\begin{aligned} RC - aC &= -KR \\ C(R-a) &= -KR \\ C &= \frac{-KR}{R-a} \end{aligned}$$

Multiplying the numerator and denominator each by -1, we get an equivalent and simpler explicit function

$$R \overset{g}{\mapsto} C = \frac{KR}{a-R}.$$

As the functions come from the same relation, we know $g = f^{-1}$.

This equation shows that C is the dependent variable and is a function of the independent variable R. Mathematically, using x as the independent variable, we would have written

$$f^{-1}(x) = \frac{Kx}{a-x}.$$

However, this equation loses the context of what the input variable x and the output value represent. In applications, it is better to include the variables so that their interpretation can be preserved. □

Recall that the composition of inverse functions should result in the input of the inner function. Consider how that applies in the context of actual variables. Recall the earlier example relating a reaction rate R and the reactant

concentration C. We had inverse functions $C \overset{f}{\mapsto} R$ and $R \overset{g}{\mapsto} C$. Composition applies these operations one immediately after the other, with the inner function applied first. Composition $f \circ g$ applies g to the input followed by f, which would be written in mapping notation with the variables as

$$R \overset{g}{\mapsto} C \overset{f}{\mapsto} R.$$

The original input is the value R and the final output is also the value R. So the comoposition is equal to the original input. Algebra should verify that this actually works.

Example 2.4.14 For the inverse functions of Michaelis-Menten kinetics,

$$C \overset{f}{\mapsto} R = \frac{aC}{C+K},$$
$$R \overset{g}{\mapsto} C = \frac{KR}{a-R},$$

show that the composition of functions cancel.

Solution. To compute $f \circ g(x)$, we use $g(x)$ as the input to f. Using meaningful variables, g takes a reaction rate as input, so we compute $f(g(R))$ and simplify. Recall that function evaluation is just substitution of the input in a formula.

$$\begin{aligned} f\big(g(R)\big) &= f\big(\frac{KR}{a-R}\big) \\ &= \frac{a\big(\frac{KR}{a-R}\big)}{\big(\frac{KR}{a-R}\big)+K} \end{aligned}$$

We replaced the C as input to f with the formula for $g(R)$. To simplify this, we can clear the denominator of the fractions inside the fraction by multiplying numerator and denominator by $(a-R)$.

$$\begin{aligned} f\big(g(R)\big) &= \frac{a\big(\frac{KR}{a-R}\big)(a-R)}{\big(\frac{KR}{a-R}+K\big)(a-R)} \\ &= \frac{aKR}{KR+K(a-R)} \\ &= \frac{aKR}{KR+Ka-KR} \\ &= \frac{aKR}{Ka} \\ &= R \end{aligned}$$

Using the placeholder variable x, we have $f \circ g(x) = x$, as required for inverse functions.

The algebraic verification that g undoes the evaluation of f,

$$C \overset{f}{\mapsto} R \overset{g}{\mapsto} C,$$

follows a similar calculation. To compute $g \circ f(x)$, we use $f(x)$ as the input to g. In context, f takes a reactant concentration C as input, so we compute $g(f(C))$ and simplify.

$$g\big(f(C)\big) = g\big(\frac{aC}{C+K}\big)$$

$$
\begin{aligned}
&= \frac{K\left(\frac{aC}{C+K}\right)}{a - \left(\frac{aC}{C+K}\right)} \\
&= \frac{K\left(\frac{aC}{C+K}\right)(C+K)}{\left(a - \frac{aC}{C+K}\right)(C+K)} \\
&= \frac{aKC}{a(C+K) - aC} \\
&= \frac{aKC}{aC + aK - aC} \\
&= \frac{aKC}{aK} \\
&= C
\end{aligned}
$$

□

2.4.3 Summary

- A function is a relation between an independent variable (input) and a dependent variable (output) such that for each value of the input, there is exactly one value for the output.
- An equation in two variables defines a relation. When we can solve the equation for one variable (dependent) as a single expression of the other variable (independent), the expression defines an explicit function.
- A linear function $x \mapsto y$ is a relationship between variables that have a constant rate of change. The rate of change equals the slope between two states (x_1, y_1) and (x_2, y_2) and is the ratio of the change in the output to the change in the input:

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

- Function mapping notation $x \overset{f}{\mapsto} y$ indicates that y is a function of x and f is the name of the function.
- Function evaluation notation $f(\square)$ uses substitution of whatever appears between the parentheses ($\square$) in place of the independent variable.
- Composition $f \circ g$ is evaluation of the outer function f with an input using the output of the inner function g,

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

  As maps, if $g : x \mapsto u$ and $f : u \mapsto y$, then

  $$x \overset{f\circ g}{\mapsto} y \quad = \quad x \overset{g}{\mapsto} u \overset{f}{\mapsto} y.$$

- Two functions f and g are inverses of one another if $f \circ g(x) = x$ for all x in the domain of g *and* $g \circ f(x) = x$ for all x in the domain of f. This means that inverse functions cancel one another when applied in a chain:

  $$f \circ f^{-1}(x) = x \quad \text{and} \quad f^{-1} \circ f(x) = x.$$

- If an equation can be solved for each variable in terms of the other (e.g., $x \mapsto y$ and $y \mapsto x$), the relation is one-to-one. The two resulting functions are inverse functions.

2.4.4 Exercises

For each equation, determine if the relation defines functions $x \mapsto y$ and $y \mapsto x$ by solving the equation for the dependent variable.

1. For the equation $3x - 5y = 10$, do the following.

(a) Determine if $x \mapsto y$.

(b) Determine if $y \mapsto x$.

2. For the equation $2xy - 6 = 4x - 3y$, do the following.

(a) Determine if $x \mapsto y$.

(b) Determine if $y \mapsto x$.

3. For the equation $6x + 4y - 3xy = 0$, do the following.

(a) Determine if $x \mapsto y$.

(b) Determine if $y \mapsto x$.

4. For the equation $x^2 + 3y = 25$, do the following.

(a) Determine if $x \mapsto y$.

(b) Determine if $y \mapsto x$.

Given a function, compute and simplify the expressions listed.

5. Suppose $f(x) = \frac{2}{3}x + 4$. Simplify each of the following expressions.

(a) $f(5)$

(b) $f(t)$

(c) $f(t^2 - 1)$

(d) $3f(2x) - 8$

6. Suppose $g(x) = \dfrac{4}{x+1}$. Simplify each of the following.

(a) $g(1)$

(b) $g(\frac{1}{x})$

(c) $\dfrac{1}{g(x)}$

(d) $g\left(\frac{1}{x} - 1\right)$

7. Suppose $f(x) = 2x - 5$, $g(x) = \frac{1}{2}x + 5$, and $h(x) = \frac{1}{2}(x + 5)$. Simplify each of the following.

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

(b) $f \circ h(x)$

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

(d) $g \circ h(x)$

What conclusion can be drawn?

8. Suppose $f(x) = \dfrac{3}{x+2}$ and $g(x) = \dfrac{3}{x} - 2$. Simplify each of the follow-

ing.

(a) $f(x-2)$

(b) $g(\frac{1}{x})$

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

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

Is $g = f^{-1}$?

Applications

9. Let C be the temperature measured in degrees Celsius, and let F be the temperature measured in degrees Fahrenheit. The function $g(x) = \frac{9}{5}x + 32$ defines the map $g : C \mapsto F$, and $h(x) = \frac{5}{9}(x - 32)$ defines $h : F \mapsto C$.

(a) Use algebra to verify that g and h are inverse functions.

(b) What is the value and interpretation of $g(30)$?

(c) What is the value and interpretation of $g \circ h(30)$?

10. A spring force scale uses the distance a spring is stretched to determine the force that is applied to the spring. We calibrate the scale by using known forces (e.g., weights) and record the corresponding location of the tip on a ruler. Let F be the force (Newtons) applied to the spring and let L be the corresponding location (centimeters). The following table is used for calibration.

F (N)	0	10.0
L (cm)	20.0	42.5

(a) Find a linear equation relating the variables F and L.

(b) Determine functions g and h so that $F \overset{g}{\mapsto} L$ and $L \overset{h}{\mapsto} F$. What are the corresponding equations using evaluation notation?

(c) Suppose a force of 5 N is applied to the spring. What will be the location of the tip of the ruler? Which function was used?

(d) Suppose a force is applied that results in the tip having a location of 28.7 cm. What was the force? Which function was used?

11. The perimeter P and area A of a square are each functions of the length of the sides s by $P = 4s$ and $A = s^2$. Find perimeter as a function of area, $P \mapsto A$.

12. The volume of a sphere is related to the radius of the sphere by the equation $V = \frac{4}{3}\pi r^3$. Suppose the radius is a function of time defined by $r = 1 + 2t$. Find the volume as a function of time, $t \mapsto V$.

13. The cost C of materials for a project depends on the required area A of materials needed. The unit price is \$3.50 per m^2. The project involves making two squares, each of them having sides with length s (meters).

(a) Find $A \overset{f}{\mapsto} C$.

(b) Find $s \overset{g}{\mapsto} A$.

(c) Use composition to find $s \mapsto C$. Is this $f \circ g$ or $g \circ f$?

(d) How much would a project with $s = 4$ cost? How much area of materials will be required? What function is used for each calculation?

2.5 Transformations of Functions

Overview. As a final section in the chapter on functions, we turn our attention to transformations. We think of functions as a relation between variables, $x \mapsto y$. In general, a transformation maps the state (x, y) to another pair (u, v). Sometimes, we will think of starting with a function $x \mapsto y$ and describe the transformation by describing what happens to each of the coordinates, $x \mapsto u$ and $y \mapsto v$. This is how we might normally think about elementary transformations including translations or shifts, scaling or stretching, and reflections.

In modeling settings, on the other hand, we might think of (u, v) as being physical variables which show a relationship similar to a well-known mathematical relationship. That relationship might seen in a graph as a parabola, as exponential growth or decay, or as periodic cycles that look like a sine wave. Here, we might think of (x, y) as describing the mathematically simple function. In order to understand the function $u \mapsto v$ based on the function $x \mapsto y$, we will more naturally think of the transformation as finding a way to map $u \mapsto x$ and $y \mapsto v$. Elementary transformations that include translations, scaling, and reflections correspond to $u \mapsto x$ and $y \mapsto v$ that are linear functions.

2.5.1 Elementary Transformations

The elementary transformations of a graph include translation, scaling, and reflection. In algebra courses, we are often given a summary of the equations of such transformations.

Elementary Transformations of Graphs.

Suppose we know the graph of a function $y = f(x)$. The following equations define the specified transformations of that graph.

- Vertical translation, shifting the graph c units vertically,
$$y = f(x) + c.$$
- Horizontal translation, shifting the graph c units horizontally,
$$y = f(x - c).$$
- Vertical scaling, stretching or compressing all vertical coordinates by a factor a,
$$y = af(x).$$
- Horizontal scaling, stretching or compressing all horizontal coordinates by a factor a,
$$y = f(\frac{x}{a}).$$
- Vertical reflection across the horizontal axis,
$$y = -f(x).$$
- Horizontal reflection across the vertical axis,
$$y = f(-x).$$

There are some key patterns to these equations of transformation. All of the

vertical transformations occur outside the function, while all of the horizontal transformations occur on the input to the function. Vertical transformations involve arithmetic consistent with the operation. For example, to move the graph up 3 units, you add $+3$ to the output of the function. Horizontal transformations involve arithmetic opposite of the desired operation. To move a graph 3 units to the right, you add -3 to the independent variable.

The following interactive graphs allow you to explore transformations of the graph $y = \sin(x)$ by dragging sliders.

Example 2.5.1 Explore horizontal and vertical translations using the equation

$$y = \sin(x + a) + b$$

using parameters a and b. Notice that because the input to the sine function is $x + a = x - -a$, the direction of translation is opposite the value of a chosen. The values $a = 0$ and $b = 0$ correspond to no transformation.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.5.2

□

Example 2.5.3 Explore horizontal and vertical scaling using the equation

$$y = b\sin(ax)$$

using parameters a and b. Notice that negative multiples result in reflections. The values $a = 1$ and $b = 1$ correspond to no transformation.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 2.5.4

□

We will explore a new approach to understanding transformations that will help us understand more complicated transformations. This approach will also help us understand why horizontal transformations seem to be opposite of what we want.

We start by thinking of the original graph $y = f(x)$ as a relation between two variables. The transformation will define a relation between two other variables, which usually will have some physical interpretation that we want to model. In this section, we will name the physical variables u and v. If our physical variables should be named x and y, then we need a way to distinguish between the original relation and the transformed relation representing physical variables. We might use uppercase X and Y or decorate the variables $\tilde{x}$ and $\tilde{y}$.

In our general approach to transformations, we will describe the transformation as the composition of a chain of mappings. Because we ultimately want a function $u \mapsto v$, we start with the physical variable u and need to map it to the mathematical variable x, $u \overset{d}{\mapsto} x$. The original function f provides the relation $x \overset{f}{\mapsto} y$. Then we need to find a map from the mathematical dependent variable y to the physical dependent variable v, $y \overset{r}{\mapsto} v$. This is summarized by the following notation:

$$u \overset{d}{\mapsto} x \overset{f}{\mapsto} y \overset{r}{\mapsto} v.$$

That is, we are going to think of the original graph as some function representing an operation. Our transformed graph is going to be a sequence of

operations that includes that operation in addition to operations that transform the values in the domain and range. All operations that occur before the function operation represented by the original graph affect the domain values, represented above by the function $d : u \mapsto x$. All operations that occur after the function operation affect the range values, represented above by the function $r : y \mapsto v$.

Example 2.5.5 Consider the function $f(x) = 2(x-3)^2 + 4$. This is a transformation of the squaring function $y = x^2$. Identify the operations affecting the domain and range in the transformation.

Solution. Let $\text{sq}(x) = x^2$ represent the operation that squares the input value. Then our function $f(x)$ can be written as a composition of operations that occur before squaring and after squaring as

$$f(x) = r \circ sq \circ d(x)$$

where $r(y) = 2y + 4$ and $d(x) = x - 3$. That is, the function f performs the following operations:

- Take the input x,
- Subtract 3,
- Square the result,
- Multiply the result by 2 and add 4.

The function d describes steps before squaring, and the function r describes the steps after squaring.

The physical meaning of the variables in the graph of $y = f(x)$ are different from the meaning of the variables in the graph of $y = x^2$, even though we use the same symbols. If we use the decorated variables $(\tilde{x}, \tilde{y})$ for our transformed graph, then we will think of f as a map $\tilde{x} \mapsto \tilde{y}$. We have a chain:

$$\begin{aligned} d : \tilde{x} &\to x = \tilde{x} - 3 \\ \text{sq} : x &\to y = x^2 \\ r : y &\to \tilde{y} = 2y + 4 \end{aligned}$$

□

On the other hand, when we geometrically describe a transformation, we usually describe how we take the original graph in (x, y)-coordinates in order to find the graph for the physical relation in (u, v)-coordinates. Describing the horizontal transformation corresponds to a mapping $x \mapsto u$. Our composition was stated in terms of the inverse operation $d : u \mapsto x$. This means that the geometric description of a transformation of the domain involves the *inverse* of the function used in the actual composition. This explains why horizontal transformations use operations that are the inverse of what we expect.

Example 2.5.6 Describe how the graph of the function $y = f(x) = 2(x-3)^2+4$ is a transformation of the elementary parabola $y = x^2$.

Solution. Based on the work in the previous example, we saw that $f(x) = r \circ \text{sq} \circ d(x)$. The transformation is geometrically described by taking a point (x, y) on the parabola and mapping it to a new point $(\tilde{x}, \tilde{y})$ on our transformed parabola. The domain is transformed using the inverse function for d,

$$x = d(\tilde{x}) = \tilde{x} - 3 \quad \Leftrightarrow \quad \tilde{x} = d^{-1}(x) = x + 3.$$

That is, the graph is translated (shifted) to the right by +3 units. The range is transformed by the function r,

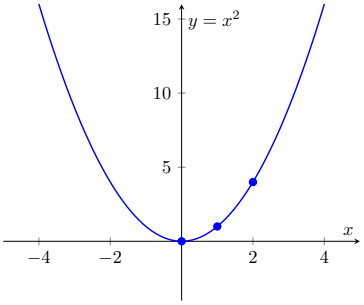
$$\tilde{y} = r(y) = 2y + 4.$$

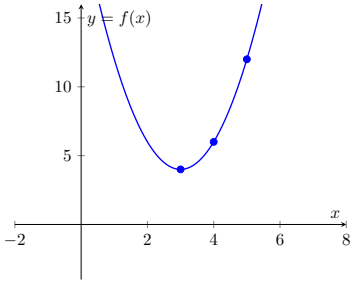
That is, the y-value of every point is first multiplied by 2 and then increased by adding 4.

To illustrate the transformation, $(\tilde{x}, \tilde{y}) = (x + 3, 2y + 4)$, consider some actual points from $y = x^2$.

$$\begin{aligned}(x, y) = (0, 0) &\mapsto (\tilde{x}, \tilde{y}) = (0 + 3, 2(0) + 4) = (3, 4)\\ (x, y) = (1, 1) &\mapsto (\tilde{x}, \tilde{y}) = (1 + 3, 2(1) + 4) = (4, 6)\\ (x, y) = (2, 4) &\mapsto (\tilde{x}, \tilde{y}) = (2 + 3, 2(4) + 4) = (5, 12)\end{aligned}$$

A graph that includes these points is shown below, in comparison with the original parabola.





□

We consider one more example of interpreting the transformation of a known graph.

Example 2.5.7 Describe $y = 2^{-3x} + 5$ as a transformation of the graph $y = 2^x$.

Solution. To make the distinction between variables more clear with two equations, let $(\tilde{x}, \tilde{y})$ be the variables in the new equation

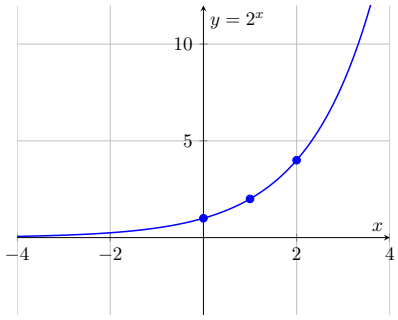
$$\tilde{y} = 2^{-3\tilde{x}} + 5$$

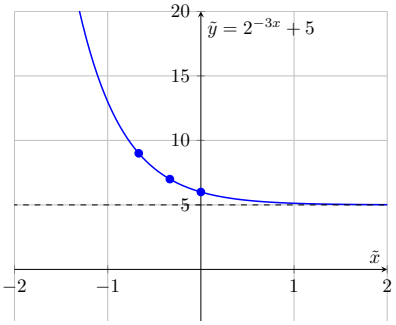
and $y = 2^x$. The steps prior to the exponential base $b = 2$ identify $x = -3\tilde{x}$ and the steps after the exponential give $\tilde{y} = y + 5$. If we solve the first equation for $\tilde{x}$ to obtain $\tilde{x} = -\frac{1}{3}x$, we have our geometric description of the transformation:

$$(x, y) \mapsto (\tilde{x}, \tilde{y}) = (-\frac{1}{3}x, y + 5).$$

That is, the graph is horizontally compressed by a factor of $\frac{1}{3}$ and reflected across the y-axis. Vertically, the graph is shifted up by 5 units.

Examples of points on the original graph $y = 2^x$ include $(0, 1)$, $(1, 2)$, and $(2, 4)$. The transformed graph results in these points being mapped to point $(\tilde{x}, \tilde{y})$ given by $(0, 6)$, $(-\frac{1}{3}, 7)$, and $(-\frac{2}{3}, 9)$. In addition, the original graph includes a horizontal asymptote of $y = 0$. The transformation $\tilde{y} = y + 5$ also applies to the asymptote, which is mapped to $\tilde{y} = 5$.





□

2.5.2 Creating a Model by Transformations

The elementary transformations correspond to coordinates that are mapped using linear functions for $u \mapsto x$ and $y \mapsto v$. The slope affects scaling and reflection and a non-zero intercept corresponds affects translation. If we know where corresponding points are found in the original and the transformed graphs, then we can find linear functions that map the coordinates. We use these functions to find the equation of the transformed graph.

Pay particular attention to the order of the mappings. The horizontal transformation that will be used in the composition is $u \mapsto x$. That is, we map from our physical independent variable (what we model) to a corresponding value of the independent variable in the elementary model. This is the inverse map of the geometric description of the transformation. On the other hand, the vertical transformation matches the geometric description, $y \mapsto v$. We map from the value of the dependent variable in the elementary model to our dependent physical variable.

Example 2.5.8 Find the equation of a parabola whose vertex is at $(2, 3)$ and which has another point at $(4, 5)$ by finding a transformation of $y = x^2$ using the points $(0, 0)$ and $(1, 1)$.

Solution. Our original graph uses coordinates (x, y). For our transformed graph, we will use coordinates (u, v). Based on the description of the problem, we need a geometric transformation of coordinates

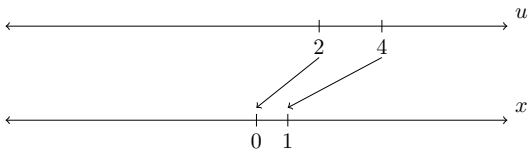
$$(x, y) = (0, 0) \mapsto (u, v) = (2, 3)$$

and

$$(x, y) = (1, 1) \mapsto (u, v) = (4, 5).$$

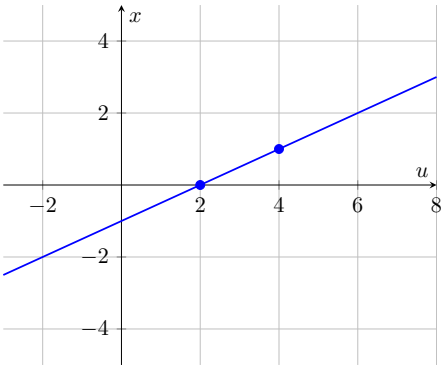
We will work with one coordinate variable at a time.

The transformation of x-coordinates, $u \mapsto x$, is illustrated in the following map. The u-coordinate $u = 2$ of the vertex should map to the x-coordinate $x = 0$, and the u-coordinate $u = 4$ of the second point should map to the $x = 1$.



This map is a linear function. To help reinforce what we are doing in this map, we also illustrate the map using a graph. The input variable for the map is u, which will be on the horizontal axis. The output variable of the map is x, which will be on the vertical axis. From our map, we know that $(u, x) = (2, 0)$

should be on our graph. We also know that $(u, x) = (4, 1)$ should be on the graph. Our linear function corresponds to the equation of the line in this graph.



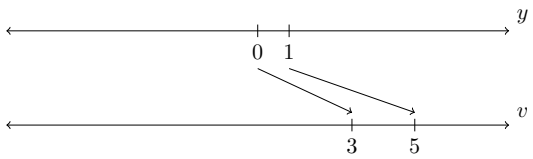
To find the equation of the function, we need the slope,

$$\frac{\Delta x}{\Delta u} = \frac{1-0}{4-2} = \frac{1}{2}.$$

Using the given point $(u, x) = (2, 0)$ and the point–slope equation, we find

$$x = \frac{1}{2}(u-2) = \frac{u-2}{2}.$$

The transformation of y-coordinates is similar, illustrated as another map below. The given points correspond to $y = 0 \mapsto v = 3$ and $y = 1 \mapsto v = 5$.



We can create the equation of the linear function describing this map. We don't need to draw the graph so long as we recognize the points that would be on the line. These two points are $(y, v) = (0, 3)$ and $(y, v) = (1, 5)$. The equation for the map $y \mapsto v$ requires the slope,

$$\frac{\Delta v}{\Delta y} = \frac{5-3}{1-0} = 2,$$

and the known point $(y, v) = (0, 3)$. The transformation is given by

$$v = 2y + 3.$$

We find the equation of the transformation by finding the composition represented by the chain,

$$\begin{cases} x = \dfrac{u-2}{2}, \\ y = f(x) = x^2, \\ v = 2y + 3. \end{cases}$$

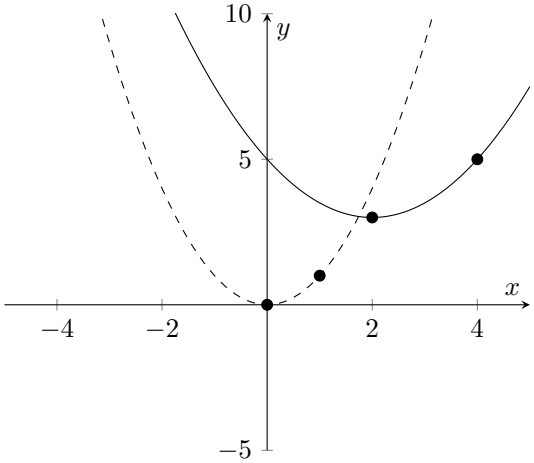
Simplifying this composition gives

$$v = 2\left(\frac{u-2}{2}\right)^2 + 3.$$

If we wanted our transformed variables to be (x, y), then we just substitute those variables and the resulting equation would be

$$y = 2\Big(\frac{x-2}{2}\Big)^2 + 3.$$

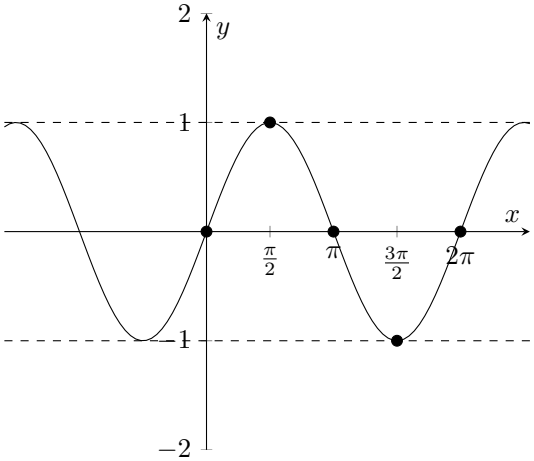
A graph of $y = x^2$ and the transformation are shown in the figure below.



□

To find the equations used in the composition for a transformation, you don't need to use the same points for both maps. All that you need to do is find the linear equation required to transform each coordinate separately. Sometimes, it is more convenient to choose different features to describe $u \mapsto x$ and $y \mapsto y$. The following example illustrates this for a sinusoidal graph.

Example 2.5.9 The sine function $y = \sin(x)$ is a periodic function with a period 2π and range $[-1, 1]$. Key points on the graph include $(0, 0)$, $(\frac{\pi}{2}, 1)$, $(\pi, 0)$, $(\frac{3\pi}{2}, -1)$, and $(2\pi, 0)$. Any graph that is an elementary transformation of the sine function is called **sinusoidal**.



Use transformations of the sine function to model the height H (in cm) of a mass bouncing on a spring as function of time t (in s). The mass completes one cycle every 2 seconds and reaches a maximum height of 10 cm and minimum height of 2 cm. The mass is known to be at its minimum at $t = 0$.

Solution. In order to find our model $t \mapsto H$, we need to determine a compo-

sition of maps

$$t \mapsto x \overset{\sin}{\mapsto} y \mapsto H.$$

We will need to find the functions relating the independent and dependent variables separately.

We start by finding the map corresponding to the independent variables, $t \mapsto x$. Because our mass is at its minimum at $t = 0$ and the sine function is at its minimum at $x = \frac{3\pi}{2}$, we can use the point $(t, x) = (0, \frac{3\pi}{2})$. We use the period to find a second point. The mass will return to its minimum at $t = 2$. The sine function returns to its minimum at $x = \frac{3\pi}{2} + 2\pi$.

We use the points $(t, x) = (0, \frac{3\pi}{2})$ and $(t, x) = (2, \frac{3\pi}{2} + 2\pi)$ to find the transformation $t \mapsto x$. First, we find the slope or rate of change,

$$\frac{\Delta x}{\Delta t} = \frac{2\pi}{2} = \pi.$$

Note that this is the ratio of the period of the elementary model 2π to the period of the oscillating mass $p = 2$. We then write down the equation using the point–slope form of a line,

$$x = \pi(t - 0) + \frac{3\pi}{2} = \pi t + \frac{3\pi}{2} = \pi(t + \tfrac{3}{2}).$$

To find the transformation for the dependent variables $y \mapsto H$, we can use the minimum and maximum values. The sine function has minimum $y = -1$ and maximum $y = 1$. The mass has minimum height $H = 2$ and maximum height $H = 10$. Our map $y \mapsto H$ includes the points $(y, H) = (-1, 2)$ and $(y, H) = (1, 10)$. The equation is based on the slope or rate of change

$$\frac{\Delta H}{\Delta y} = \frac{10 - 2}{1 - -1} = 4$$

and the point–slope equation

$$H = 4(y - 1) + 10 = 4y + 6.$$

We put these together as a chain or composition,

$$t \mapsto x \mapsto y \mapsto H,$$

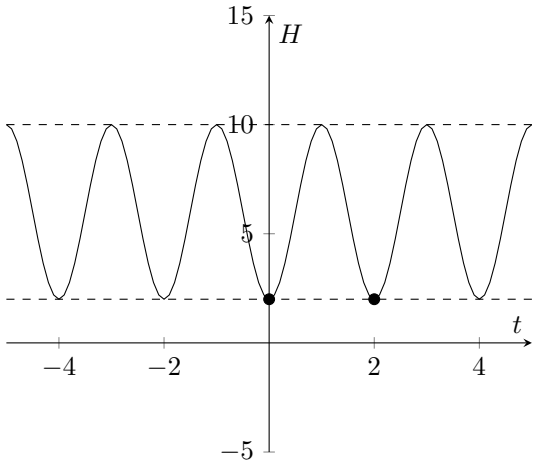
using the individual relations found above,

$$\begin{cases} x = \pi(t + \frac{3}{2}), \\ y = \sin(x), \\ H = 4y + 6. \end{cases}$$

Combining these equations, we find our model equation

$$H = 4 \sin\left(\pi(t + \tfrac{3}{2})\right) + 6.$$

A graph shows that this model matches the description given.



□

In general, a sinusoidal graph can be characterized by a centerline $v = c$, an amplitude A, a period p, and a phase shift $u = \phi$. The standard period of the elementary sine and cosine functions is 2π. The phase shift ϕ is the value of the physical variable $u = \phi$ that corresponds to $x = 0$. These two pieces of information completely determine the map $u \mapsto x$, with slope $\frac{\Delta x}{\Delta u} = \frac{2\pi}{p}$, with a point–slope equation

$$x = \frac{2\pi}{p}(u - \phi).$$

The centerline $v = c$ physically corresponds to the elementary centerline of $y = 0$. The amplitude A corresponds to the elementary amplitude of 1. Consequently, the slope of the transformation map $y \mapsto v$ is $\frac{\Delta v}{\Delta y} = A$ so that the point–slope equation of the map is

$$v = Ay + c.$$

The composition of the chain of mappings with the sine function results in

$$v = A \sin\left(\frac{2\pi}{p}(u - \phi)\right) + c,$$

while a composition with the cosine function results in

$$v = A \cos\left(\frac{2\pi}{p}(u - \phi)\right) + c.$$

We use the sine function when the phase shift corresponds to where the graph crosses the centerline $v = c$; the cosine function is used when the phase shift corresponds to the locations of maximum values at $v = c + A$.

2.5.3 Summary

- Elementary transformations of a graph of a function include translation, scaling, and reflection.
- Vertical transformations are applied to the output of the function directly.
- Horizontal transformations are applied to the input of the function as inverse operations.

- Elementary transformations can be found using linear functions that map the original coordinates (x, y) to the transformed coordinates (X, Y) with $x \overset{T_x}{\mapsto} X$ and $y \overset{T_y}{\mapsto} Y$ using composition,

$$X \overset{T_x^{-1}}{\mapsto} x \mapsto y \overset{T_y}{\mapsto} Y.$$

  - Addition by a constant (intercept) corresponds to translation.
  - Multiplication by a constant (slope) corresponds to rescaling.
  - A negative multiple (slope) corresponds to reflection.

- Nonlinear transformations using invertible maps can also be useful. A common transformation is the logarithm, corresponding to viewing a graph with logarithmic scales.

- Data (x, y) that appear linear on a semi-log plot (with the y-axis in logarithmic scale) means that $(x, \ln(y))$ will have a linear relation. Subsequently, $x \mapsto y$ will be an exponential model.

- Data (x, y) that appear linear on a log-log plot (with both axes in logarithmic scale) means that $(\ln(x), \ln(y))$ will have a linear relation. Subsequently, $x \mapsto y$ will be a power function model.

2.5.4 Exercises

1. Find the equation of a parabola with a vertex at $(3, 2)$ and a second point at $(6, 0)$.

2. View a graph of $y = |x|$, which forms the shape of a "V". Find the equation of a transformation that moves the vertex to $(-3, 2)$, opens downward, and has a second point at $(0, 0)$.

3. The function $f(x) = \dfrac{x}{1+x}$ is a simple increasing, concave down function that passes through $f(0) = 0$ and has a horizontal asymptote $y = 1$ and half-saturation constant $x = 1$. This basic function is often used to model the reaction rate of enzyme-catalyzed reactions.

Suppose that C is the concentration of a reactant in an enzyme-catalyzed reaction and V is the rate of reaction. Use transformations of $y = \frac{x}{1+x}$ to find a model for $C \mapsto V$ such that the $(C, V) = (0, 0)$ is a possible state, the saturating rate is $V = 50$, and the half-saturation occurs at $C = 80$.

4. The Gaussian function $f(x) = e^{-\frac{1}{2}x^2}$ is symmetric about $x = 0$ with a maximum $f(0) = 1$, has a horizontal asymptote $y = 0$ as $x \to \pm\infty$, and has inflection points at $x = \pm 1$. This function is often used in statistics to describe normally distributed data.

The height of individuals in a population were recorded and observed to have a normal distribution. A histogram plot showing the number of individuals N with the same height H, rounded to the nearest centimeter, could be modeled $H \mapsto N$ as a transformation of f. The maximum in the histogram (line of symmetry) is at $H = 152$ cm with $N = 250$. The inflection points were observed to be at $H = 144$ cm and $H = 160$ cm. The horizontal asymptote is still $N = 0$. Find a Gaussian curve to model the distribution.

5. Find an equation for a sinusoidal graph (x, y) with a period 10, a center line (midpoint between the maximum and minimum) of $y = 8$, and has a maximum at $(x, y) = (2, 11)$.

6. The number of hours of daylight H is a periodic function of the time D, measured in days after the year begins. The cycle repeats every 365 days. The longest day is on the summer solstice, which occurs at $D = 172.25$, with $H = 14.85$. The shortest day is the winter solstice, which occurs at $D = 354.75$, with $H = 9.48$. Find a sinusoidal model for $D \mapsto H$. Use your model to determine the number of hours of daylight on *pi*-day, $D = 72$.

7. The cosine is another trigonometric function with period 2π and range $[-1, 1]$ and has the identical graph shape as sine. However, the maximum for cosine occurs at $\cos(0) = 1$ and the minimum occurs at $\cos(\pi) = -1$. Find the equation of cosine in terms of the sine function by using transformations.

8. The square of the cosine function, $\cos^2(x)$, has the same shape as the sine and cosine graphs, except that the period is π and the minimum and maximum values are 0 and 1, respectively. We know $\cos^2(0) = 1$ and $\cos^2(\frac{\pi}{2}) = 0$. Find the equation of $\cos^2(x)$ in terms of the sine or cosine function by using transformations.

9. The square of the sine function, $\sin^2(x)$, has the same shape as the sine and cosine graphs, except that the period is π and the minimum and maximum values are 0 and 1, respectively. We know $\sin^2(0) = 0$ and $\sin^2(\frac{\pi}{2}) = 1$. Find the equation of $\sin^2(x)$ in terms of the sine or cosine function by using transformations.

10. A population P grows exponentially in time t such that $P = 500$ when $t = 2$ and triples every 12 years. Use the semi-log transform to find a linear model $t \mapsto \ln(P)$ and then find the model $t \mapsto P$. What was the population when $t = 0$?

11. In a simple electrical circuit, the voltage V on a capacitor decays exponentially as a function of time t. After $t = 5$ seconds, we find $V = 8$ volts; after another 5 seconds, we find $V = 6.4$ volts. Use the semi-log transform to find a linear model $t \mapsto \ln(V)$ and then find the exponential model $t \mapsto V$. When will the circuit reach $V = 1$ volt?

12. During childhood development, the head grows at a different rate than the rest of the body. This is why the heads of young children look larger proportional to their body than older children and adults. Such growth is called **allometry** and is often observed to follow a power law.

According to the World Health Organization's statistics for child development, the median circumference C of the head for a one-year-old girl is 45 cm and the median height H is 74.5 cm. For a five-year-old, the median head size is $C = 50$ cm and the median height is $H = 109.5$ cm. Use a log-log transform to find a linear model $\ln(H) \mapsto \ln(C)$ and then find the power function model $H \mapsto C$. If this pattern continues, predict the head circumference for a ten-year-old where the median height is $H = 140$ cm.

13. Chemical reactions generally occur at a rate R that is proportional to a power of the reactant concentration C. Such a reaction will have a graph (C, R) that appears linear in a log-log plot. Suppose you have a reaction such that $R = 0.5$ when $C = 0.2$ and $R = 1.5$ when $C = 0.4$. Use a log-log transform to find a linear model $\ln(C) \mapsto \ln(R)$ and then find the power function model $C \mapsto R$.

Chapter 3

Accumulation and Rates of Change

3.1 Describing the Behavior of Functions

Overview. We have been learning about how functions are constructed and how they are defined. In many instances, before we construct a formula for a function, we need to identify what behavior we are attempting to model. At other times, we have a formula and we need to know what behavior that predicts. We need specific language that we can use to describe behavior.

In this section, we will focus on three types of behavior: monotonicity, concavity, and end behavior. Monotonicity will describe where a function is increasing or decreasing. Concavity will describe where the slope ro rate of change of a function is increasing or decreasing. In a graph, concavity describes whether the curve is bending up or bending down. We also discuss simple end behavior including unbounded growth (tending to infinity) and horizontal asymptotes.

Our emphasis is in learning the language of behavior, describing graphs using this language, and creating graphs based on a description of a function. As our study of calculus develops, we will learn mathematical tools that will allow us to determine function behavior more precisely.

3.1.1 Functions Have Shapes

We often describe functions according to the shape of their graphs. The different possible shapes we see in graphs correspond to specific behaviors of the functions. We will focus on two aspects of a graph: monotonicity and concavity.

3.1.1.1 Monotonicity

The **monotonicity** of a function deals with whether the function is increasing or decreasing. We start with the mathematical definitions of increasing and decreasing functions. We will explore the ideas graphically in terms of maps and then graphs.

Definition 3.1.1 Monotonicity. A function f is **increasing** on a subset S of the domain (usually an interval) if for every $x_1, x_2 \in S$,

$$x_1 < x_2 \qquad \text{implies} \qquad f(x_1) < f(x_2).$$

A function f is **decreasing** on a subset S of the domain (usually an interval) if for every $x_1, x_2 \in S$,

$$x_1 < x_2 \qquad \text{implies} \qquad f(x_1) > f(x_2).$$

$\Diamond$

One way to think of monotonicity is that the function retains an ordering of the sets. An increasing function preserves the order, so that if two inputs are in a particular order, $x_1 < x_2$, then the resulting outputs have the same order, $f(x_1) < f(x_2)$. A decreasing function reverses the order, so that if inputs have an order, $x_1 < x_2$, then the outputs must have the opposite order $f(x_1) > f(x_2)$. A function that is not monotone (neither increasing or decreasing) does not maintain a sense of order uniformly over the set. Sometimes the outputs might have the same order as the inputs, and sometimes the outputs might have the opposite order.

Example 3.1.2 The function $f(x) = 2x + 1$ is a linear function with positive slope $m = 2$. We can show that f is an increasing function. Suppose $x_0 < x_1$. Multiplying both sides of an inequality by a positive number *preserves* the ordering, as does adding the same value to both sides:

$$\begin{aligned} x_0 &< x_1 \\ 2x_0 &< 2x_1 \\ 2x_0 + 1 &< 2x_1 + 1 \\ f(x_0) &< f(x_1) \end{aligned}$$

This is visualized in the following figure.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 3.1.3 Dynamic illustration of the function $f(x) = 2x + 1$ as a map $x \mapsto y$ showing that f is increasing.

Thinking of the map dynamically, we see that as we increase the input, the output also increases. This is captured in the graph of the function in the (x, y) plane. The graph shows y-values increasing as viewed from left to right, which is corresponding to x-values increasing.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 3.1.4 The graph of the function $y = f(x) = 2x + 1$ in the (x, y) plane.

□

Example 3.1.5 The function $f(x) = -2x+3$ is a linear function with negative slope $m = -2$. We can show that f is an decreasing function. Suppose $x_0 < x_1$. Multiplying both sides of an inequality by a negative number *reverses* the ordering, while adding the same value to both sides preserves the order:

$$\begin{aligned} x_0 &< x_1 \\ -2x_0 &> -2x_1 \\ -2x_0 + 3 &> -2x_1 + 3 \\ f(x_0) &> f(x_1) \end{aligned}$$

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 3.1.6 Dynamic illustration of the function $f(x) = -2x + 3$ as a map $x \mapsto y$ showing that f is decreasing.

The map shows that the order of outputs is always opposite to the order of the inputs. Thinking of the map dynamically, we see that as we increase the input, the output decreases. The graph of the function in the (x, y) plane captures the same information. Viewing the graph from left to right (as x increases), the y-values decrease.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 3.1.7 The graph of the function $y = f(x) = -2x + 3$ in the (x, y) plane.

□

Some functions are not monotone because the map does not retain the ordering of the sets. Dynamically, this is because the output will sometimes increase and sometimes decrease as the input is increased.

Example 3.1.8 The function $f(x) = x^2$ is not a linear function and is not monotone. We can show this by illustrating that the function is inconsistent in ordering the output values relative to the input values. Consider $x_0 = -2$ and $x_1 = -1$. We have $f(x_0) = 4 > f(x_1) = 1$, so for these inputs the order is reversed. However, for $x_0 = 1$ and $x_1 = 2$, we have $f(x_0) = 1 > f(x_1) = 4$ and the order is preserved. This function is not increasing or decreasing, but is a combination.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 3.1.9 Dynamic illustration of the function $f(x) = x^2$ as a map $x \mapsto y$ showing that f is not monotone.

We can see graphically that f is decreasing on $(-\infty, 0]$ because for any two inputs in this interval, the order of the outputs is reversed. We can also see that f is increasing on $[0, \infty)$ because for any two inputs in that interval, the order of the outputs is preserved. This point where monotonicity switches corresponds to the vertex of the parabola $y = x^2$.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 3.1.10 Dynamic illustration of the function $f(x) = x^2$ as a map $x \mapsto y$ showing that f is not monotone.

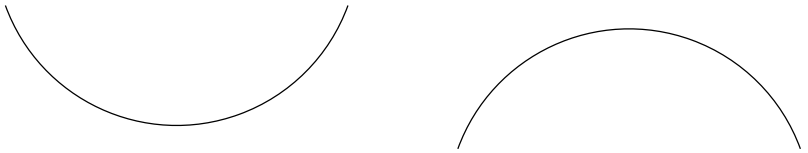
□

One of our goals in calculus will be to develop a method to determine the intervals on which a function is increasing or decreasing. When we motivated monotonicity with linear functions, we saw that a positive slope implied an increasing function and a negative slope implied a decreasing function. Calculus will develop a more general sense of the slope of a function using the derivative such that we will describe monotonicity based on the signs of the derivative.

Note 3.1.11 When listing intervals on which a function is increasing or decreasing, it is important *not* to use a union of the intervals. The reason is that we are saying that the function is increasing on *each* of the intervals individually and not on the set formed by the union. If listing multiple intervals, simply form a comma-separated list.

3.1.1.2 Concavity

Concavity describes how the graph of a function in the (x, y) plane bends. If the graph bends upward, we say the function is **concave up**. If the graph bends downward, we say the function is **concave down**.



(a) Concave up **(b)** Concave down

Figure 3.1.12 Comparison of concave up and concave down graphs

As with monotonicity, these attributes of functions apply over intervals rather than at individual points. When a graph changes concavity at a point, for example switching from bending up to bending down, the function has an **inflection point**. A technical definition of concavity that depends on the concept of a derivative will be provided later.

However, we can capture the essential idea by thinking about how the slope is changing between points. A function that has an increasing slope or rate of change over an interval is concave up on the interval. A function that has decreasing slope or rate of change is concave down.

Definition 3.1.13 Concavity. A function f is **concave up** on a subset S of the domain (usually an interval) if for every $x_1, x_2, x_3 \in S$, the slope or rate of change is increasing,

$$x_1 < x_2 < x_3 \qquad \text{implies} \qquad \frac{f(x_2) - f(x_1)}{x_2 - x_1} < \frac{f(x_3) - f(x_2)}{x_3 - x_2}.$$

A function f is **concave down** on a subset S of the domain (usually an interval) if for every $x_1, x_2, x_3 \in S$, the slope or rate of change is decreasing,

$$x_1 < x_2 < x_3 \qquad \text{implies} \qquad \frac{f(x_2) - f(x_1)}{x_2 - x_1} > \frac{f(x_3) - f(x_2)}{x_3 - x_2}.$$

$\Diamond$

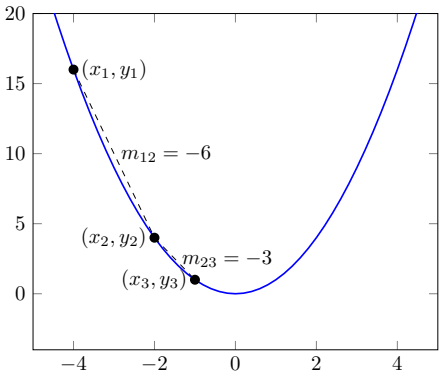
This definition is not very easy to use directly. When we learn more about derivatives to describe the slope at individual points, we will have a much better method known as the second derivative test for concavity. However, the following examples will illustrate what is happening.

Example 3.1.14 The function $f(x) = x^2$ is concave up on $(-\infty, \infty)$ (the entire domain). We will not *prove* that this is true because this is too difficult without derivatives. But we can illustrate the idea.

Consider the graph $y = f(x) = x^2$ and the particular values $x_1 = -4$, $x_2 = -2$, and $x_3 = -1$. We will calculate the slope or rate of change between $(x_1, y_1) = (-4, 16)$ and $(x_2, y_2) = (-2, 4)$ and between (x_2, y_2) and $(x_3, y_3) = (-1, 1)$.

$$m_{12} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{4 - 16}{-2 - -4} = \frac{-12}{2} = -6$$
$$m_{23} = \frac{y_3 - y_2}{x_3 - x_2} = \frac{1 - 4}{-1 - -2} = \frac{-3}{1} = -3$$

We can see that the slope or rate of change is increasing, $m_{12} < m_{23}$. These slopes are illustrated in the following figure.



This is not a proof of concavity because we only illustrated the order for three specific points. Use the following dynamic figure to convince yourself that for *any* three points we might choose, the slopes increase from left to right.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 3.1.15

The reason that f has an inflection point at $x = 0$ is that point is where f has the steepest negative slope. To the left, $x < 0$, the slope decreases; to the right, $x > 0$, the slope increases. □

Example 3.1.16 The function $f(x) = x^3 - 3x$ changes concavity at $x = 0$. f is concave down on $(-\infty, 0]$ and concave up on $[0, \infty)$. When three points are chosen with $x \in (-\infty, 0]$, the slope is decreasing. When the three points are chosen with $x \in [0, \infty)$, the slope is increasing. This can be verified in the following dynamic figure. However, the three points must all be in either $(-\infty, 0]$ or in $[0, \infty)$.

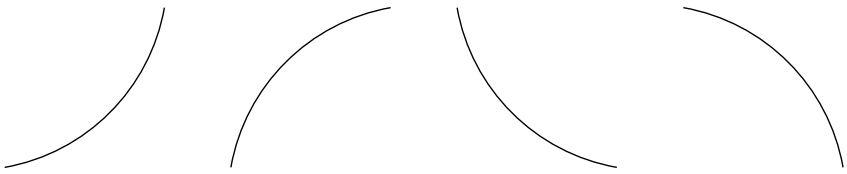
A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 3.1.17

□

3.1.1.3 Combining Monotonicity and Concavity

The shape of a graph of a function is often defined in terms of the monotonicity and concavity combined. There are four basic shapes that correspond to the four quadrants of a circle, illustrated in the figure below. A curve that has a positive and increasing slope is increasing and concave up. A curve that has a positive but decreasing slope is increasing and concave down. A curve that has a negative but increasing (becoming less negative) slope is decreasing and concave up. A curve that has a negative and decreasing (becoming more negative) slope is decreasing and concave down.



(a) Increasing, Concave up

(b) Increasing, Concave Down

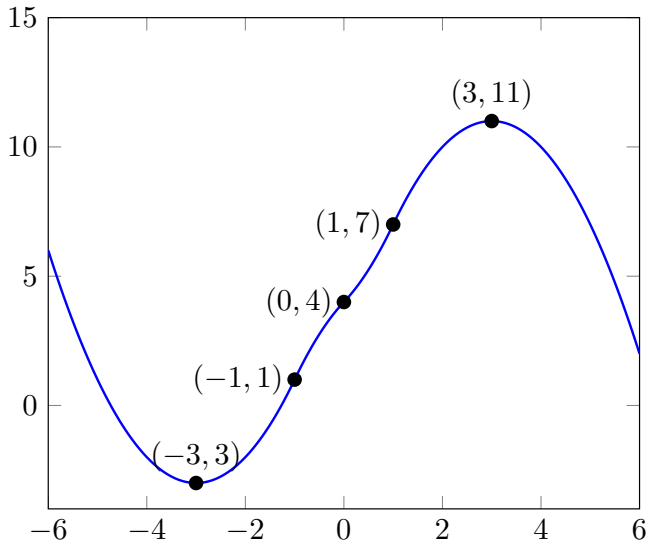
(c) Decreasing, Concave up

(d) Decreasing, Concave Down

Figure 3.1.18 Basic shapes defined by monotonicity and concavity.

We can describe the shape of a graph by stating intervals on which the function satisfies each of the possible behaviors. The intervals are separated by points where the graph reaches either a maximum or minimum value (changes in monotonicity) or where the slope of the graph reaches an extreme and begins to bend the other direction (changes in concavity or points of inflection).

Example 3.1.19 The graph of a function $y = f(x)$ is shown below, with labeled extreme points and inflection points. Describe the shape of the graph by giving intervals of monotonicity and concavity.



Solution. Intervals for monotonicity are based on the function increasing or decreasing. The end-points of these intervals are the extreme points for the function. When the graph extends beyond the frame of the figure, we assume the function behavior continues as shown. Intervals always are read from left to right. The end-point of an interval is included (closed) if the behavior extends up to and including that point.

The function f is decreasing on $(-\infty, -3]$, increasing on $[-3, 3]$, and decreasing on $[3, \infty)$. Notice that the extremes at $x = -3$ and $x = 3$ are included in two intervals. The continuous function is decreasing on $(-\infty, -3)$ as an open interval. Because f decreases up to and including $x = -3$, we include the end-point.

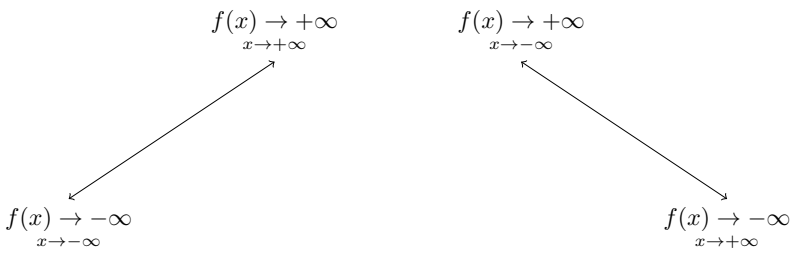
Intervals for concavity are based on where the slope is increasing or decreasing. Intervals on which the graph bends upward, f is concave up. Intervals on which the graph bends downward, f is concave down. Notice our graph has inflection points (where the concavity changes) at $x = -1$, $x = 0$, and $x = 1$. At these points, the graph starts to bend in the opposite direction.

The function f is concave up on $(-\infty, -1]$, concave down on $[-1, 0]$, concave up on $[0, 1]$, and concave down on $[1, \infty)$. We include the inflection points as the end points of the intervals (closed) because the slope is increasing or decreasing up to and including those points. □

3.1.2 End Behavior

End-behavior of a function describes what happens to a function as the size of the input grows. Consider the possibilities of a linear function, $y = f(x) = mx + b$. So long as the slope is non-zero, the function is *unbounded*, meaning that the graph eventually goes above every level and eventually goes below every level (on opposite sides of the graph).

If the slope is positive, $m > 0$, then the function is increasing. We say $f(x) \to +\infty$ as $x \to +\infty$, which we read as "the value of $f(x)$ tends to positive infinity as the value of x goes to positive infinity". This is because the y-values will eventually rise *above* any level on the right side of the graph (for sufficiently large positive values x). We also say $f(x) \to -\infty$ as $x \to -\infty$ because the y-values are *below* any specified value on the left side of the graph (for sufficiently large negative values x). When the slope is negative, $m < 0$, the unbounded behavior is reversed.



(a) Positive slope ($m > 0$) **(b)** Negative slope ($m < 0$)

Figure 3.1.20 Unbounded behavior of linear functions with positive and negative slopes.

For consistency in notation to describe the tendency of a function (as opposed to the value of a function), we use limits to describe unbounded behavior.

$$\begin{aligned}
\lim_{x\to-\infty} f(x) &= -\infty && \text{means } f(x) \to -\infty \text{ as } x \to -\infty \\
\lim_{x\to-\infty} f(x) &= \infty && \text{means } f(x) \to +\infty \text{ as } x \to -\infty \\
\lim_{x\to\infty} f(x) &= -\infty && \text{means } f(x) \to -\infty \text{ as } x \to +\infty \\
\lim_{x\to\infty} f(x) &= \infty && \text{means } f(x) \to +\infty \text{ as } x \to +\infty
\end{aligned}$$

When the graph of a function f behaves more and more like a constant function (horizontal line) for larger and larger values of the independent variable, we say f has a **horizontal asymptote**. A horizontal asymptote $y = L$ on the right side (large, positive values for x) uses the limit statement

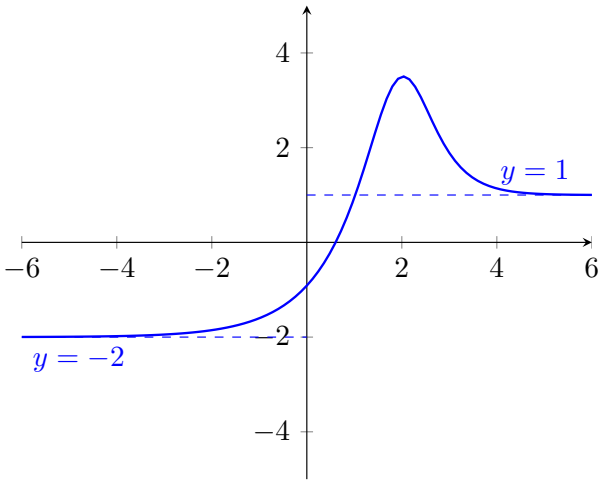
$$\lim_{x\to\infty} f(x) = L,$$

which means that the value of $f(x)$ approaches the constant value L as $x \to +\infty$. When f has a horizontal asymptote $y = L$ on the left side (large, negative values of x), we use the limit statement

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

Example 3.1.21 The graph of a function $y = f(x)$ is shown below. This function has two horizontal asymptotes: $y = -2$ as $x \to -\infty$ and $y = 1$ as $x \to +\infty$. We write

$$\begin{aligned}
\lim_{x\to-\infty} f(x) &= -2, \\
\lim_{x\to+\infty} f(x) &= 1.
\end{aligned}$$



□

A function can also have unbounded behavior near a particular input value, say at $x = a$. Using limit notation, this means that at least one of the following must be true.

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

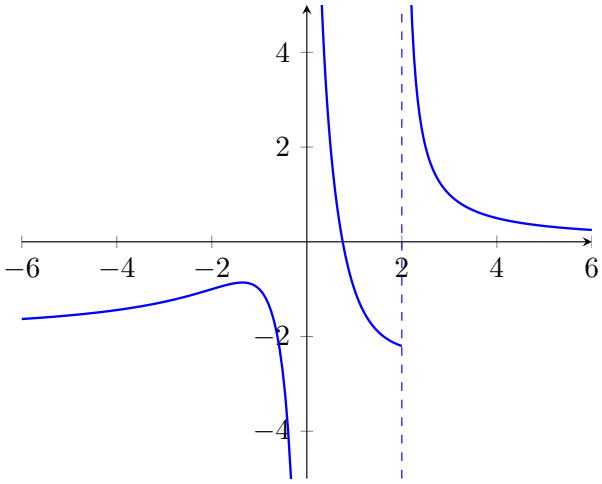
The graph has a **vertical asymptote** at $x = a$, meaning that the graph of the function approaches closer and closer to this vertical line.

Example 3.1.22 The graph of a function $y = f(x)$ is shown below with two vertical asymptotes. The vertical asymptote at $x = 0$ corresponds to left- and right-limits

$$\lim_{x\to 0^-} f(x) = -\infty,$$
$$\lim_{x\to 0^+} f(x) = +\infty,$$

The vertical asymptote at $x = 2$ only corresponds to the right-limit

$$\lim_{x\to 2^+} f(x) = +\infty.$$



It is hard to tell from a graph alone where a vertical asymptote occurs. Using only the limited graph window, it is not obvious that the vertical asymptote is at exactly $x = 0$ since the graph is still fairly far away from that vertical line from this perspective. □

Note 3.1.23 A common false impression about horizontal asymptotes is that the graph of a function can not cross the asymptote. A function can not cross a vertical asymptote, but that is only because a function can not intersect a vertical line at more than one point. An asymptote only requires that the graph behaves more and more like the line.

When a function physically relates two variables, $x \mapsto y$, a horizontal asymptote indicates that for sufficiently large values of the independent variable, the dependent variable is essentially a constant. A common description in physical settings for this constant is a **saturation value**. We think of the quantity measured by the independent variable as a control variable. The dependent variable can be thought of as a response. As the control variable is increased, the response will pass through some of its range of values. However, there will come a point where even though you continue to increase the control variable, the response is no longer able to change very much at all. That is, the response has saturated.

Example 3.1.24 An enzyme is a protein that helps catalyze a chemical reaction. The rate or velocity of reaction V depends on the concentration of the reactant C. Commonly, the function $C \mapsto V$ is increasing, concave down, and has a horizontal asymptote, known as Michaelis–Menten reaction kinetics. The physical domain is $C \in [0, \infty)$. Because the relation is increasing, we know that adding more reactants will raise the reaction rate. Because the relation is concave down, we know that the degree to which the rate increases slows down as more reactants are added. The horizontal asymptote means that this increase in the reaction rate saturates to some maximum rate $V_{\max}$,

$$\lim_{C \to \infty} V = V_{\max}.$$

The reactant concentration where the reaction rate is halfway to the maximum value is called the half-saturation value, and is usually represented with a constant K.

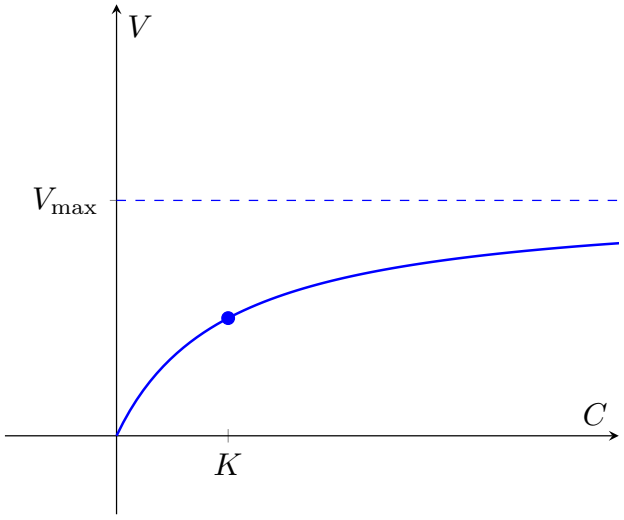


Figure 3.1.25 Michaelis–Menten reaction kinetics with saturating rate $V_{\max}$ and half-saturation constant K.

□

Example 3.1.26 Imagine a crop of plants growing in a field. The total biomass harvested B depends on the number of seeds S that are sown. If very few seeds are sown, the biomass harvested will be small. For more seeds sown, we expect the biomass would increase. However, if too many seeds are sown, then the crop will be overcrowded, resulting in a lower harvest. We expect that there might be an optimal number of seeds S^* for which the biomass is at a maximum.

Describe the behavior of the function $S \mapsto B$ and sketch a possible graph.

Solution. The function $S \mapsto B$ will have a physical domain of $S \in [0, \infty)$. Because B is a maximum at $S = S^*$, the function is increasing on $[0, S^*]$ and decreasing on $[S^*, \infty)$. The simplest assumption for concavity would be that the function starts concave down. However, a concave down and decreasing function will eventually approach $-\infty$, which is not physically possible for our physical scenario. Therefore, the function must change concavity at some inflection point after S^*, say at $S = S^\dagger$. Our function would be concave down on $[0, S^\dagger]$ and concave up on $[S^\dagger, \infty)$. Continuing to increase the number of seeds will result in ever smaller biomass due to overcrowding until it approaches some saturating biomass B_∞,

$$\lim_{S \to \infty} B = B_\infty.$$

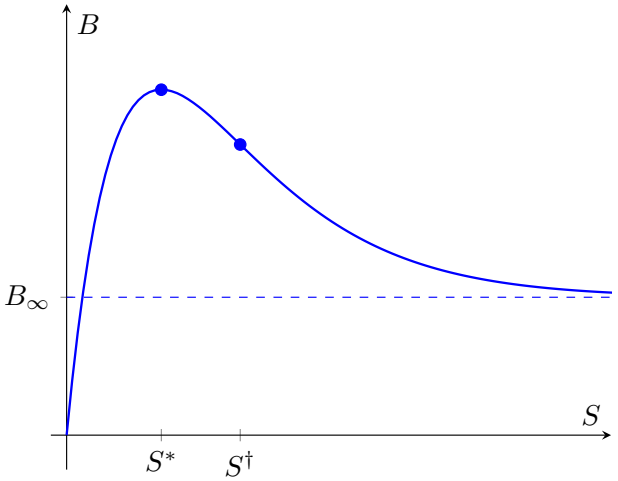


Figure 3.1.27 Possible graph of (S, B) with maximum at $S = S^*$ and inflection point at $S = S^\dagger$.

Note: The asterisk and dagger are decorations so that the symbols S^* and $S^\dagger$ represent general constants. We don't know actual values for the maximum and inflection point, so we can't use numbers. The symbols are place-holders for values that would be determined experimentally. Similarly, the symbol B_∞ represents the value for the biomass harvested when the number of seeds sown saturates the system. □

3.1.3 Summary

- Describing the **monotonicity** of a function is determining intervals on which the function is increasing or decreasing.
- A function f is **increasing** on a set S if the function is order preserving: For all $x_1, x_2 \in S$, we must have

$$x_1 < x_2 \quad \Rightarrow \quad f(x_1) < f(x_2).$$

This corresponds to a graph that is rising left to right (positive slopes).

A function f is **decreasing** on a set S if the function is order reversing: For all $x_1, x_2 \in S$, we must have

$$x_1 < x_2 \quad \Rightarrow \quad f(x_1) > f(x_2).$$

This corresponds to a graph that is falling left to right (negative slopes).

- Describing the concavity of a function is determining intervals on which the function is concave up or concave down.
- A function f is **concave up** on a set S if the slope or rate of change is increasing on S: For all $x_1, x_2, x_3 \in S$, we must have

$$x_1 < x_2 < x_2 \quad \Rightarrow \quad \frac{f(x_2) - f(x_1)}{x_2 - x_1} < \frac{f(x_3) - f(x_2)}{x_3 - x_2}.$$

The graph will be bending upward.

A function f is **concave down** on a set S if the slope or rate of change is decreasing on S: For all $x_1, x_2, x_3 \in S$, we must have

$$x_1 < x_2 < x_2 \quad \Rightarrow \quad \frac{f(x_2) - f(x_1)}{x_2 - x_1} > \frac{f(x_3) - f(x_2)}{x_3 - x_2}.$$

The graph will be bending downward.

- A **point of inflection** is a point where a function is continuous and changes concavity.
- Lists of intervals of monotonicity and concavity should be separated by commas and not joined by unions.
- Limits as $x \to \pm\infty$ describe end behavior.
  - To say $f(x) \to +\infty$ means values of $f(x)$ eventually rise above *any* possible value.
  - To say $f(x) \to -\infty$ means values of $f(x)$ eventually fall below *any* possible value.
  - To say $f(x) \to L$ means values of $f(x)$ eventually approaches a horizontal asymptote $y = L$.

3.1.4 Exercises

Each of the following problems asks you to prove that the given function is either increasing or decreasing on a particular interval.

1. Prove that $f(x) = 5x - 12$ is an increasing function by showing that whenever $x_1 < x_2$, we have $f(x_1) < f(x_2)$.

2. Prove that $f(x) = -3x - 2$ is a decreasing function by showing that whenever $x_1 < x_2$, we have $f(x_1) > f(x_2)$.

3. Prove that $f(x) = x^2$ is an increasing function on $[0, \infty)$ by showing that whenever $0 < x_1 < x_2$, we have $f(x_1) > f(x_2)$.
Hint: Show that $f(x_2) - f(x_1) > 0$ by factoring and determining the signs of the factors.

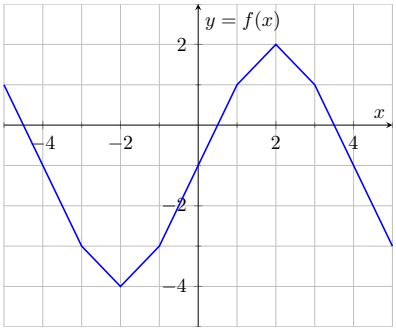
4. Prove that $f(x) = x^2$ is a decreasing function on $(-\infty, 0]$ by showing that whenever $x_1 < x_2 < 0$, we have $f(x_1) > f(x_2)$ or $f(x_2) - f(x_1) <$

0.

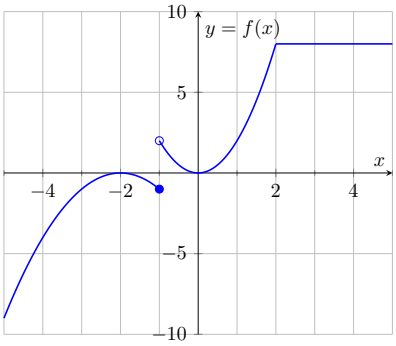
Hint: Show that $f(x_2) - f(x_1) < 0$ by factoring and determining the signs of the factors.

Consider each of the following graphs of functions. Use the graph to determine the intervals of monotonicity for that function.

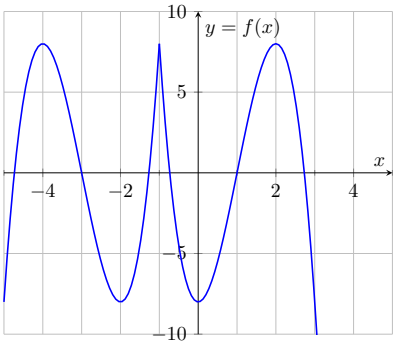
5.



6.



7.



Each of the following problems asks you to illustrate the concavity of the given function.

8. Illustrate that $f(x) = \dfrac{1}{x}$ is concave up on $(0, \infty)$ by showing that the slope is increasing for the sequential points $x_1 = \frac{1}{2}$, $x_2 = 1$, and $x_3 = 2$.

9. Illustrate that $f(x) = \dfrac{1}{x}$ is concave down on $(-\infty, 0)$ by showing that the slope is decreasing for the sequential points $x_1 = -2$, $x_2 = -1$, and $x_3 = -\frac{1}{2}$.

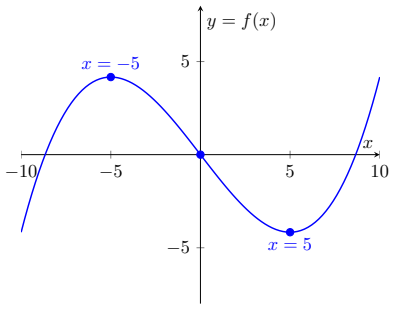
10. Illustrate that $f(x) = 2^x$ is concave up on $(-\infty, \infty)$ by showing that the slope is increasing for the sequential points $x_1 = -1$, $x_2 = 0$, and

$x_3 = 1$.

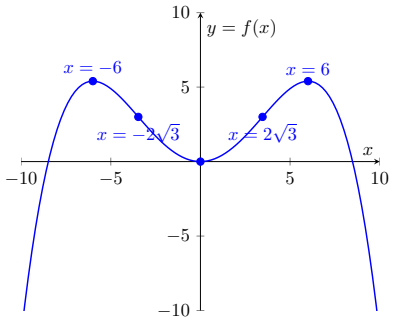
11. Illustrate that $f(x) = 2^{-x}$ is concave up on $(-\infty, \infty)$ by showing that the slope is increasing for the sequential points $x_1 = -1$, $x_2 = 0$, and $x_3 = 1$.

Consider each of the following graphs of functions, which includes turning points and inflection points. Use the graph to determine the intervals of monotonicity and concavity for that function.

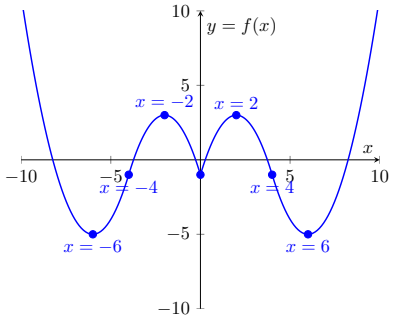
12.



13.

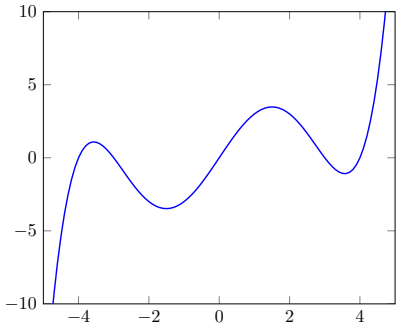


14.



Use the graphs to answer the questions about limits. Assume that the behavior of the graph shown in the window continues outside the window.

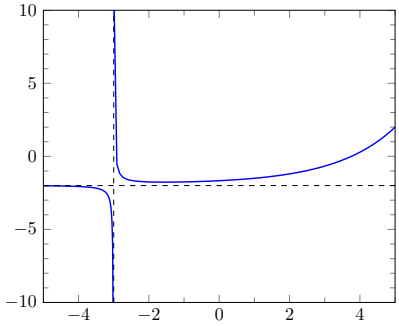
15.



(a) $\lim_{x \to -\infty} f(x)$

(b) $\lim_{x \to +\infty} f(x)$

16.



(a) $\lim_{x \to 3^-} f(x)$

(b) $\lim_{x \to 3^+} f(x)$

(c) $\lim_{x \to -\infty} f(x)$

(d) $\lim_{x \to +\infty} f(x)$

3.2 Rate of Accumulation and the Derivative

3.2.1 Overview

For functions which are defined as the accumulation of a given rate,

$$f(x) = f_0 + \int_{x_0}^{x} f'(z)\,dz, \tag{3.2.1}$$

we can describe their monotonicity and concavity using the signs and monotonicity of their corresponding rates of accumulation $f'(x)$. Of course, most functions are not written using an accumulation formula representation. For polynomials, where we know simple accumulation formulas, we know how to calculation the corresponding rate of accumulation $f'(x)$. What about other functions?

This raises a central question of calculus: Which functions *can* be expressed as an accumulation? And if a function can be expressed as an accumulation, how do we find the formula for the rate of accumulation that will be the integrand in that representation?

This question will be partially resolved through the definition of the derivative. The derivative will provide a new interpretation of the concept of rate of change that is not directly connected to accumulation and definite integrals. At that point, we will have two potential concepts of the rate—the rate of accumulation and the derivative. The connection between these two rates as representing the same thing will ultimately be established through the Fundamental Theorem of Calculus. Anticipating this eventual equality, we will adopt the name **derivative** as being equivalent to the rate of accumulation.

In this section, we will use known elementary accumulation functions and their corresponding rates to compute the rate of accumulation for simple polynomials. The process of finding a rate of change inherits the linearity properties of integration. Using elementary formulas and linearity, we will learn to identify its rate of change or derivative of any polynomial. Once we have a rate of change, we can express the polynomial as an accumulation function. We can also classify the monotonicity and concavity of the polynomial.

3.2.2 The Rate of Accumulation

When a function is defined as an accumulation function in terms of a definite integral (3.2.1), it is easy enough to determine the rate of accumulation or derivative by identifying the function in the integrand. We just need to express the function as a constant (the initial value) plus an integral from the initial point. The function inside the integral, called the **integrand**, will be the rate of accumulation.

Example 3.2.1 If $f(x)$ is defined as $f(x) = 3 + \int_1^x z^2 - 5z\,dz$, then the integrand $z^2 - 5z$ must be $f'(z)$. Changing variables means $f'(x) = x^2 - 5x$ is the rate of accumulation or derivative. □

Example 3.2.2 Find $G'(x)$ for $G(x) = 4 - 3\int_1^x \frac{1}{z}\,dz$.

Solution. Because $G(x)$ is not yet written as a constant *plus* an integral, we need to use properties of integrals to put it in the standard form of an accumulation function. The (((Unresolved xref, reference "thm-definite-integral-constant-multiple"; check spelling or use "provisional" attribute)))constant mul-

tiple rule allows us to treat -3 as a constant multiplied inside the integral,

$$G(x) = 4 + \int_1^x \frac{-3}{z}\, dz.$$

Now that $G(x)$ is an accumulation function, we can find the rate to be

$$G'(x) = \frac{-3}{x}.$$

□

The elementary accumulation formulas for simple powers 3.4.3 can be interpreted as complementary rules used to find the rate of accumulation or derivative for simple powers. As an example, consider the known accumulation formula

$$\int_0^x z^2\, dz = \frac{1}{3}x^3.$$

If we multiply both sides of the equation by 3 to clear the fraction and move the constant inside the integral, we have an equivalent statement

$$x^3 = \int_0^x 3z^2\, dz.$$

That is, we have just found that for the function $f(x) = x^3$ the derivative is the rate of accumulation $f'(x) = 3x^2$. Every accumulation formula that we know provides a corresponding rate of accumulation for simple powers.

Theorem 3.2.3 The Power Rule for the Rate of Accumulation. *The elementary accumulation formulas lead to the following elementary rates of accumulation for powers of the independent variable.*

- *If* $f(x) = x$, *then* $f'(x) = 1$.
- *If* $f(x) = x^2$, *then* $f'(x) = 2x$.
- *If* $f(x) = x^3$, *then* $f'(x) = 3x^2$.
- *If* $f(x) = x^4$, *then* $f'(x) = 4x^3$.

Proof. Each formula follows by applying the same technique described above for $f(x) = x^3$ on the corresponding accumulation formula. ■

You might have noticed a pattern in these formulas, that the rate of accumulation involves a power that has been reduced by one from the accumulation function and that the power for the accumulation has become a constant multiple. Without creating more accumulation formulas, we can not prove that this pattern will always be true. In mathematics, we call a pattern that we believe might be true a **conjecture**.

Conjecture 3.2.4 Power Rule for Rate of Accumulation. *For any constant power* n, *the function* $f(x) = x^n$ *has a rate of accumulation* $f'(x) = nx^{n-1}$.

This conjecture happens to be true, but we will need to wait to develop the process of differentiation to prove it.

Example 3.2.5 For $f(x) = x^{10}$, what does the conjecture about the power rule predict will be $f'(x)$?

Solution. The function f is an elementary power function with $n = 10$. The power rule shown in the conjecture states that

$$f'(x) = 10x^9.$$

The reason this is a conjecture right now is that we would currently need to write x^{10} as an accumulation function

$$x^1 0 = \int_0^x 10z^9,$$

but we don't currently have a rule for a definite integral with a power that high. That would require knowing the limit of a Riemann sum

$$\int_0^x 10z^9 = \lim_{n\to\infty} \sum_{k=1}^{n} 10(\frac{kx}{n})^9 \frac{x}{n}.$$

Without knowing a sum accumulation for $\sum_{k=1}^{n} k^9$, we can't justify this integral. □

We will need one more simple rate of accumulation to deal with constant terms. A constant function sees no change, so the accumulation must always be zero, and that will come from a rate of accumulation that is zero.

Theorem 3.2.6 Rate of Accumulation for Constants. *If $f(x) = c$ for some constant c (a constant function), then $f'(x) = 0$.*

Proof. Because $\int_a^b 0\,dx = 0$, we can write $f(x) = c + \int_a^x 0\,dz$ for any value a. Thus, the rate of accumulation $f'(x) = 0$. ■

The linearity properties of definite integrals imply that rates of accumulation also satisfy the same linearity properties.

Theorem 3.2.7 Linearity of Rates of Accumulation. *If $f(x)$ has a rate of accumulation $f'(x)$ and $g(x)$ has a rate of accumulation $g'(x)$, then the function $h(x) = c_1 f(x) + c_2 g(x)$ with constants c_1 and c_2 has a rate of accumulation $h'(x) = c_1 f'(x) + c_2 g'(x)$.*

Proof. Using a common initial point at $x = a$, we can write

$$f(x) = f(a) + \int_a^x f'(z)\,dz,$$
$$g(x) = g(a) + \int_a^x g'(z)\,dz.$$

Because $h(x) = c_1 f(x) + c_2 g(x)$, we can use the linearity properties of definite integrals to rewrite

$$\begin{aligned} h(x) &= c_1\Big(f(a) + \int_a^x f'(z)\,dz\Big) + c_2\Big(g(a) + \int_a^x g'(z)\,dz\Big) \\ &= c_1 f(a) + c_2 g(a) + c_1 \int_a^x f'(z)\,dz + c_2 \int_a^x g'(z)\,dz \\ &= (c_1 f(a) + c_2 g(a)) + \int_a^x (c_1 f'(z) + c_2 g'(z))\,dz \end{aligned}$$

Since $h(a) = c_1 f(a) + c_2 g(a)$, we can see that $h(x)$ has been written in the form of an accumulation with a rate of accumulation $h'(x) = c_1 f'(x) + c_2 g'(x)$. ■

A polynomial is a linear combination of simple powers. Consequently, the derivative of a polynomial will be the same linear combination of the derivatives of those powers.

Example 3.2.8 Find $f'(x)$ for $f(x) = x^2 - 6x + 5$.

Solution. We look at $f(x)$ as a sum of three terms:

$$f(x) = \underbrace{x^2}_{f_1(x)} + \underbrace{-6x}_{f_2(x)} + \underbrace{5}_{f_3(x)}.$$

First, $f_1(x) = x^2$ is an elementary power for which we know $f_1'(x) = 2x$. Next, $f_2(x) = -6x$ is a constant -6 times x, so $f_2'(x)$ is the same constant times 1, $f_2'(x) = -6 \cdot 1 = -6$. Finally, $f_3(x) = 5$ is a constant function so that $f_3'(x) = 0$. The linearity for rates of accumulation implies

$$f'(x) = \underbrace{2x}_{f_1'(x)} + \underbrace{-6}_{f_2'(x)} + \underbrace{0}_{f_3'(x)} = 2x - 6.$$

□

3.2.3 Using the Rate of Accumulation

Now that we know how to find the rate of accumulation for simple polynomials, we can express a polynomial as an accumulation with that rate. Although we know the rate of accumulation, we also need to be careful that the initial value matches the function of interest.

Example 3.2.9 Express $f(x) = x^3 - 6x^2 - 4$ as an accumulation from $x = 1$.

Solution. Start by finding the rate of accumulation.

$$f'(x) = 3x^2 - 6(2x) + 0 = 3x^2 - 12x$$

The rate of accumulation becomes the integrand of the accumulation using $f'(z)$. Because the integral will start at $x = 1$, the initial value will be

$$f(1) = 1^3 - 6 \cdot 1^2 - 4 = -9.$$

We can now write $f(x)$ as an accumulation from $x = 1$:

$$\begin{aligned} f(x) &= f(1) + \int_1^x f'(z)\,dz \\ &= -9 + \int_1^x 3z^2 - 12z\,dz. \end{aligned}$$

□

The sign of the rate of accumulation determines whether an accumulation is increasing or decreasing. Furthermore, the monotonicity of the rate of accumulation determines the concavity of the accumulation. If we determine the rate of accumulation for a polynomial $f(x)$, then we can use sign analysis on the resulting formula $f'(x)$ to characterize monotonicity of $f(x)$.

Because the rate of accumulation $f'(x)$ will itself be a new polynomial, we can describe its monotonicity using the rate of accumulation for that rate of accumulation. The rate of accumulation of the rate of accumulation is called the **second derivative** and is named $f''(x)$. Because the monotonicity of the rate $f'(x)$ determines concavity of f, we can use sign analysis of $f''(x)$ to characterize the concavity f.

Example 3.2.10 We found that $f(x) = x^3 - 6x^2 - 4$ has corresponding rate $f'(x) = 3x^2 - 12x$. Describe the monotonicity and concavity of $f(x)$.

Solution. Sign analysis of the rate of accumulation $f'(x)$ will be used to

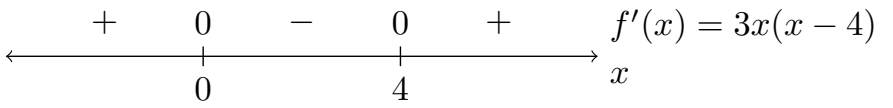
describe the monotonicity of $f(x)$. We factor

$$f'(x) = 3x^2 - 12x = 3x(x-4).$$

The zeros (x-intercepts) occur at $x = 0$ and $x = 4$ and $f'(x)$ is continuous. We need to test the sign of $f'(x)$ in the intervals $(-\infty, 0)$, $(0, 4)$, and $(4, \infty)$. Only the sign matters, so using the factored formula, we can count how many factors are positive and negative.

- $f'(-1) = 3(-1)(-5) = +$ (2 negative factors)
- $f'(1) = 3(1)(-3) = -$ (1 negative factors)
- $f'(5) = 3(5)(1) = +$ (0 negative factors)

When doing such a problem, we just use a sign analysis number line to record our results.

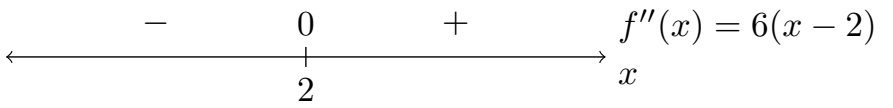


We interpret the signs of $f'(x)$ to deduce the monotonicity of $f(x)$. The function $f'(x)$ is positive on the intervals $(-\infty, 0)$ and $(4, \infty)$ and negative on the interval $(0, 4)$. Consequently, $f(x)$ must be increasing on the intervals $(-\infty, 0)$ and $(4, \infty)$ and decreasing on the interval $(0, 4)$. Because $f(x)$ is continuous (all polynomials are continuous), we can extend each of these intervals to include the end-points at $x = 0$ and $x = 4$.

Concavity of $f(x)$ depends on the monotonicity of $f'(x)$. Because $f'(x) = 3x^2 - 12x$ is itself a polynomial, we can find its rate of accumulation $f''(x)$ and perform sign analysis to deduce that monotonicity.

$$f''(x) = 6x - 12 = 6(x-2)$$

The only zero of $f''(x)$ is at $x = 2$. We test the intervals $(-\infty, 2)$ and $(2, \infty)$ using the signs of $f''(x)$, summarized by the following number line.



Because $f''(x) < 0$ on the interval $(-\infty, 2)$, we know that $f'(x)$ is decreasing on $(-\infty, 2)$. Consequently, $f(x)$ is concave down on $(-\infty, 2)$. Because $f''(x) > 0$ on the interval $(2, \infty)$, we know that $f'(x)$ is increasing on $(2, \infty)$. Consequently, $f(x)$ is concave up on $(2, \infty)$. Again, because $f(x)$ is continuous, intervals of concavity can be extended to include the end-point at $x = 2$.

Notice how the graph of $y = f(x)$ reflects the monotonicity and concavity that we have determined.

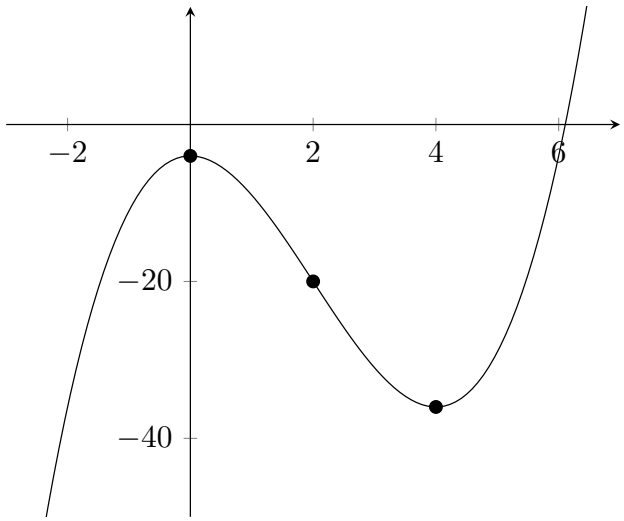


Figure 3.2.11 Graph of $y = f(x) = x^3 - 6x^2 - 4$, including turning points at $x = 0$ and $x = 4$ and an inflection point at $x = 2$.

□

3.2.4 Summary

- The rate of accumulation $f'(x)$ is the integrand function of an accumulation function $f(x)$,

$$f(x) = f(a) + \int_a^x f'(z)\,dz$$

- The rate of accumulation for a function is equivalent to what we will later define as the derivative of the function. Thus, we often just call the rate of accumulation $f'(x)$ the derivative of $f(x)$. Showing this equivalence will be the goal of the Fundamental Theorem of Calculus.

- Our known accumulation formulas have a complementary interpretation as derivatives:

  - $f(x) = x$ has derivative $f'(x) = 1$
  - $f(x) = x^2$ has derivative $f'(x) = 2x$
  - $f(x) = x^3$ has derivative $f'(x) = 3x^2$
  - $f(x) = x^4$ has derivative $f'(x) = 4x^3$

  A pattern in these derivatives suggest a conjecture (which we later prove) that any power $f(x) = x^n$ has derivative $f'(x) = nx^{n-1}$.

  In addition, any constant function $f(x) = c$ has derivative $f'(x) = 0$.

- Rates of accumulation (derivatives) satisfy the linear properties of a sum rule and a constant multiple rule.

- For simple polynomials $f(x)$, we compute $f'(x)$ as a related polynomial in order to answer questions about monotonicity. Because $f'(x)$ is a polynomial, it also has its own derivative $f''(x)$ (called the second derivative of $f(x)$). The sign of $f''(x)$ determines the monotonicity of $f'(x)$, which in turn determines the concavity of $f(x)$.

3.2.5 Exercises

For each function defined in terms of an integral, identify the rate of accumulation.

1. $f(x) = -5 + \int_1^x ze^{-z}\,dz$

2. $Q(x) = \int_0^x \frac{s}{s^2+1}\,ds$

3. $A(x) = 4\int_{-2}^x t + 3t^3\,dt$

4. $R(x) = 1 + 2\int_1^x \sin(z)\,dz - 3\int_1^x \cos(z)\,dz$

Find the derivative of each polynomial.

5. $p(x) = x^2 - 5x$

6. $q(x) = x^3 - 6x^2$

7. $r(x) = x^4 + 2x^3 - 5$

Write each polynomial as an accumulation function from the indicated starting point.

8. $p(x) = x^2 - 5x$ from $x = 1$

9. $p(x) = x^2 - 5x$ from $x = -2$

10. $q(x) = x^3 - 6x^2$ from $x = 2$

11. $q(x) = x^3 - 6x^2$ from $x = 0$

12. $r(x) = x^4 + 2x^3 - 5$ from $x = -1$

13. $r(x) = x^4 + 2x^3 - 5$ from $x = 2$

Determine the monotonicity and concavity of each polynomial.

14. $f(x) = 3x^2 - 48x + 5$

15. $g(x) = 100 + 80x - 2x^2$

16. $h(x) = 9x^2 - x^3$

17. $w(x) = x^3 - 9x^2 + 15x + 4$

18. $u(x) = x^4 - 6x^2 + 15$

19. $M(x) = 15 + 4x^3 - x^4$

3.3 Accumulation of Change

Overview. Many students learn a basic rule relating distance d, speed r and time t: $d = rt$ or "distance equals rate times time." This statement is really only true when the rate is unchanging. If the speed is constant at a rate of r_1 for a time t_1 and then instantly changes to a new constant speed at rate r_2 for another time t_2, then the total distance d traveled over the total time is

$$d = r_1 t_1 + r_2 t_2.$$

This generalizes to any number of intervals of constant rate, that the total change in position (displacement) equals the sum of the products of the rate times the increment of time at that rate.

The **definite integral** is the mathematical generalization of the idea that we just described. Given any rate of change $r(x)$ for a quantity Q with respect to an independent variable x, the definite integral's purpose is to compute the increment of change in Q when the independent variable changes from one value, $x = a$, to another value, $x = b$. We write

$$Q(b) - Q(a) = \int_a^b r(x)dx \qquad \Leftrightarrow \qquad Q(b) = Q(a) + \int_a^b r(x)dx.$$

This section introduces the idea of the definite integral for special functions for which we can compute the increment of change without knowing any additional calculus rules. We start with simple functions, which means the functions are constant on intervals. These functions motivate the basic properties which we then apply graphically and numerically. We will learn the rules later.

3.3.1 Rate of Change

Suppose that we have two variables that are related by a function. In mathematics, we often think of the prototypical variables x and y with some function $f : x \mapsto y$. But in physical situations, we are often considering changes in time so that we use the independent variable t for time. The official definition for **rate of change** is as the **derivative**. In the present context, we will not need to know how to compute derivatives. We only need to consider that there is a function that physically measures a rate of change.

For example, a speedometer measures speed which is a rate of change of distance with respect to time. As another example, we can physically measure the rate at which water flows through a pipe which represents a rate of change of a reservoir (e.g., a tub or a pool) that is being filled or drained. An electrical analog of water flow is electrical current which measures the rate of change of electrical charge along an electrical path. In biology, the rate of change of a population is physically measured through birth, death and migration rates.

When any of these rates are constant over an interval $t \in [a, b]$, the net change in the quantity of interest Q is equals the rate times the increment of time. The following definition makes this clear.

Definition 3.3.1 Constant Rate of Change. Given a quantity Q that is a function of independent variable t, say $t \mapsto Q(t)$, we say that Q has a constant rate of change r on an interval $[a, b]$ if for any t_1, t_2 satisfying $a \le t_1 < t_2 \le b$,

$$Q(t_2) - Q(t_1) = r \cdot (t_2 - t_1),$$

which is often written $\Delta Q = r \cdot \Delta t$. $\Diamond$

A quantity that has a constant rate of change satisfies a linear equation on the given interval and the rate of change corresponds to the slope of that line. In particular, if c is any value for t in the interval, $c \in [a, b]$, then the accumulation $Q(t)$ is a linear function of t,

$$Q(t) = Q(c) + r(t - c),$$

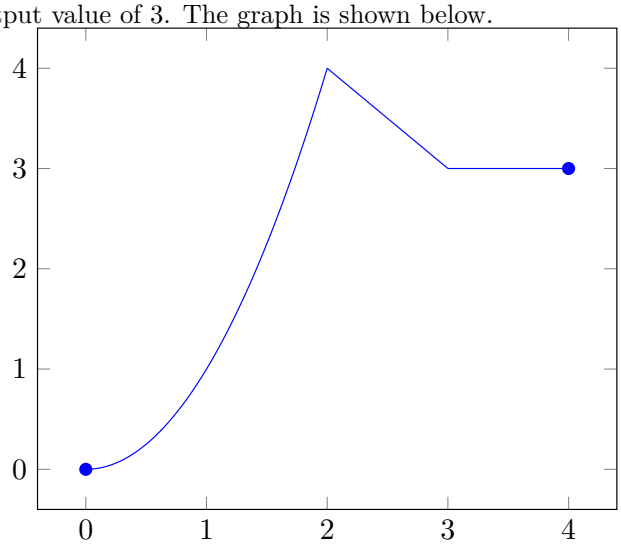
using the point–slope equation of a line. The value $Q(c)$ represents the initial value while $r(t - c)$ represents the increment of change in Q when the independent variable goes from c to the value t.

In preparation for extending the idea of rate of change, we need to recall the concept of piecewise functions 5.2.3. A piecewise function considers its domain as consisting of a collection of disjoint (non-overlapping) intervals. On each such interval, the function has a separate formula or rule of calculation.

Example 3.3.2 The function f is defined by the equation

$$f(x) = \begin{cases} x^2, & 0 \le x < 2, \\ 6 - x, & 2 \le x \le 3 \\ 3, & 3 < x \le 4 \end{cases}.$$

The domain of f is the union of disjoint intervals $[0, 2)$, $[2, 3]$ and $(3, 4]$ which corresponds to $[0, 4]$. The notation states that for input values x between 0 and 2, including 0, the function will square the input to give the output. Between 2 and 3, inclusively, the function will subtract the input from 6 for the output. For input values greater than 3 but less than or equal to 4, the function has a constant output value of 3. The graph is shown below.



□

Using piecewise functions, we can define something called a **simple function**. Such a function is piecewise constant, meaning that the domain is formed as a union of disjoint intervals and the function has a constant value on each interval. To define these intervals, we first introduce the idea of a **partition** which will be used to define the end points of these subintervals.

Definition 3.3.3 Partition. A **partition** of size n of an interval $[a, b]$ is an increasing, finite sequence of numbers $P = (x_0, x_1, \ldots, x_n)$ such that $x_0 = a$, $x_n = b$ and $x_j < x_{j+1}$. The **increments** of the partition correspond to the widths of subintervals, with

$$\nabla x_j = x_j - x_{j-1}$$

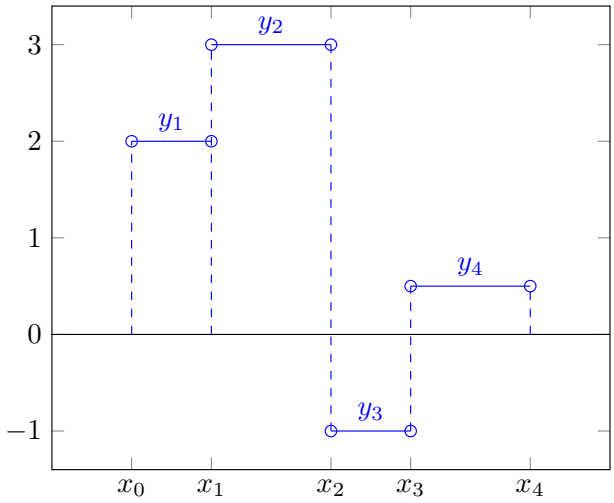
being the width of the subinterval $[x_{j-1}, x_j]$ for $j = 1, \ldots, n$. ◊

Definition 3.3.4 Simple Function. Given a partition P of size n of an interval $[a, b]$, a function f is a **simple function** on the partition P with values $(y_1, \ldots, y_n)$ if

$$f(x) = \begin{cases} y_1, & x_0 < x < x_1, \\ y_2, & x_1 < x < x_2, \\ \vdots & \\ y_n, & x_{n-1} < x < x_n. \end{cases}$$

◊

The figure below illustrates a simple function defined with a partition of size $n = 4$. Open circles are used on the edges of the segments because we did not define the value at the actual partition points, only on the intervals between those points. That is because when a rate changes instantaneously between two values, the rate can not be properly defined at the instant itself.



We can use a simple function to represent a special case of a varying rate of change, namely a rate of change that is constant on subintervals but which changes instantly (not physically possible in most situations) at the points of a partition. Given a simple rate function, $r(x)$, on a partition P of size n of the interval $[a, b]$ with values $(r_1, r_2, \ldots, r_n)$, we can define an **accumulation function** that is piecewise linear on the same partition having initial value $(a, f(a))$:

$$f(x) = \begin{cases} f(a) + r_1(x - x_0), & x_0 \le x < x_1, \\ f(a) + r_1 \nabla x_1 + r_2(x - x_1), & x_1 \le x < x_2, \\ f(a) + \sum_{k=1}^{2} r_k \nabla x_k + r_3(x - x_2), & x_2 \le x < x_3, \\ f(a) + \sum_{k=1}^{3} r_k \nabla x_k + r_4(x - x_3), & x_3 \le x < x_4, \\ \vdots & \\ f(a) + \sum_{k=1}^{n-1} r_k \nabla x_k + r_n(x - x_{n-1}), & x_{n-1} \le x \le x_n. \end{cases}$$

The purpose of the summation is to represent the accumulation of change on all previous subintervals of the partition in order to make the accumulation function $f(x)$ continuous on the full interval $[a, b]$.

Definition 3.3.5 Definite Integral of Simple Function. Given a simple rate function, $r(x)$, on a partition P of size n of the interval $[a, b]$ with values $(r_1, r_2, \ldots, r_n)$, the total accumulated change associated with this rate is the definite integral represented by is given by

$$\int_a^b r(x)\,dx = \sum_{k=1}^{n} r_k\,\nabla x_k.$$

$\Diamond$

It is common that the increments ∇x_k be instead written Δx_k. However, this is a notational abuse because Δx_k technically represents the forward difference $\Delta x_k = x_{k+1} - x_k$ with $k = 0, \ldots, n-1$. Ignoring this complaint, the total accumulation of change is often written

$$f(b) - f(a) = \sum_{k=1}^{n} r_k\,\Delta x_k.$$

(The complaint can formally be resolved by shifting the index values from $k = 1, \ldots, n$ to $k = 0, \ldots, n-1$.)

Example 3.3.6 A storage reservoir starts with 100 gallons of water. Over the next 20 minutes, water is added to the reservoir at a rate of 5 gal/min. Then water is pumped out at a rate of 12 gal/min for 10 minutes. For the next 30 minutes, water is added at a rate of 3 gal/min. Find a piecewise linear function describing the amount of water in the reservoir as a function of time (in minutes).

Solution. The rate function is a simple function using the partition P that starts at $x_0 = 0$ and has increments of $\nabla x_1 = 20$, $\nabla x_2 = 10$ and $\nabla x_3 = 30$. That is, the partition is given by $P = \{0, 20, 30, 60\}$. The rate of change of water is constant on the subintervals defined by this partition:

$$R(t) = \begin{cases} 5, & 0 < t < 20, \\ -12, & 20 < t < 30, \\ 3, & 30 < t < 60. \end{cases}$$

The amount of water in the reservoir is also a function of time, say $W(t)$, and is defined as an accumulation using the rate of change $R(t)$ found above. Because the reservoir begins with $W(0) = 100$, our initial value, we can write $W(t)$ as a piecewise linear function that accumulates the change in water over each of the subintervals. Consider first the total accumulation of change in water on each of the subintervals, which is equal to the rate of change times the increment of time for that subinterval.

$$\begin{aligned} W(20) - W(0) &= 5(20) = 100 \\ W(30) - W(20) &= -12(10) = -120 \\ W(60) - W(30) &= 3(30) = 90 \end{aligned}$$

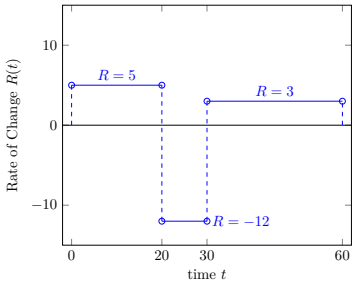
Notice that the total change in water volume over the entire interval $[0, 60]$ is the sum of these increment,

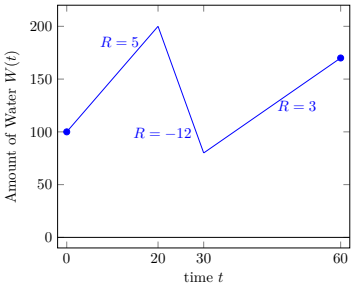
$$W(60) - W(0) = 100 + -120 + 90 = 70.$$

The accumulation function $W(t)$, which has an initial value $W(0) = 100$, is therefore defined by

$$W(t) = \begin{cases} 100 + 5(t-0) = 100 + 5t, & 0 \le t < 20, \\ 100 + 100 - 12(t-20) = 200 - 12(t-20), & 20 \le t < 30, \\ 200 - 120 + 3(t-30) = 80 + 3(t-30), & 30 \le t \le 60. \end{cases}$$

The graphs of the rate function $R(t)$ and the water level $W(t)$ are shown below. Notice that although the rate function is not defined at the partition points, the water level function $W(t)$ is defined and continuous at those points. It is continuous because the accumulations are designed to start on the next interval exactly where it stops from the previous interval with no sudden jumps.





□

There is an important geometric interpretation of accumulation in terms of area on the graph. Recall that the area of a rectangle is defined as the product of the height and the width. Mathematically, this is the same operation as when we calculate an increment as the product of a rate and the increment of the independent variable, except that a rate can be negative. Consequently, we introduce the idea of **signed area**.

Definition 3.3.7 Signed Area (Informal). Suppose we have the graph of a function $y = f(x)$ that is continuous on an interval (a, b) and is either entirely above the axis, $f(x) > 0$ for all $x \in (a, b)$, or entirely below the axis, $f(x) < 0$ for all $x \in (a, b)$. Then we can define the **signed area** of the graph by considering the area A (area itself is always positive) of the region between the curve $y = f(x)$ and the axis $y = 0$ and between the vertical lines $x = a$ and $x = b$. If $f(x) > 0$ (above the axis), then we say that we have **positive area** A; if $f(x) < 0$ (below the axis), then we say that we have **negative area** $-A$.

If the graph $y = f(x)$ on an interval (a, b) has a finite number of discontinuities or crosses the axis so that sometimes the graph is above the axis and sometimes below, then we can consider a partition of $[a, b]$ using the x-values of the discontinuities and zeros of f. Then on every subinterval from this partition, the earlier definition applies and we have a signed area for each subinterval. The signed area for the entire graph is the sum of the signed areas of the subintervals, adding areas that are above the axis and subtracting areas that are below the axis. ◊

Given any simple rate function $f(x)$, the signed area of the graph $y = f(x)$ on the interval $[a, b]$ consists of the sum of signed areas of rectangles. This exactly matches the definite integral

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} r_k \nabla x_k.$$

Therefore, we adopt the definite integral as our formal definition of signed area.

Definition 3.3.8 Signed Area and Accumulated Change (Formal). Suppose we have a function $y = f(x)$ that is bounded and piecewise continuous on an interval (a, b) $(a < b)$. The signed area of f on the interval (a, b) is defined as the definite integral

$$\int_a^b f(x)\,dx.$$

If $f(x)$ gives the rate of change of a quantity Q with respect to the independent variable x, then the definite integral also gives the increment of change in Q:

$$Q(b) - Q(a) = \int_a^b f(x)\,dx.$$

The function $f(x)$ is called the **integrand** and the variable x is called the **variable of integration**. The values a and b are called the **limits of integration**. ◊

The notation of the definite integral uses an elongated "S" called the integral symbol $\int$ that should remind you of the idea of summing increments of signed area. The limits of integration a (lower) and b (upper) represent the starting and ending points of integration, respectively. The increments of signed area are represented by the formula $f(x)\,dx$ which represents a strip of signed area with signed height $f(x)$ (generalizing the constant height of a simple rate function) and infinitesimally small width dx (generalizing the increments of a partition ∇x).

Although we have presented these ideas as definitions, they are really important consequences of the development of calculus. In particular, the statement that the increment of change $Q(b) - Q(a)$ is equal to the definite integral of the rate of change of Q is so most important that this result is called the (((Unresolved xref, reference "fundamental-theorem-calculus"; check spelling or use "provisional" attribute)))Fundamental Theorem of Calculus . One of the primary goals of learning calculus is to understand what this theorem means and why it is really true.

3.3.2 Interpretation of Definite Integrals as Signed Areas

We will learn integration methods later. For now, we will explore examples, including simple functions, where knowing the interpretation of a definite integral allows us to determine results using only the ideas of signed area. When exact area calculations can not be found, then approximations of signed area can allow us to estimate the value of the definite integral. We start by revisiting our earlier examples using the context of definite integrals.

Example 3.3.9 Integral of Constant Rate. Consider the case of a constant function $R(t) = r$, representing a constant rate of change for some quantity Q with respect to time t. Earlier we noted that for constant rate of change, the increment of change in Q as t changes from $t = a$ to $t = b$ is equal to

$$Q(b) - Q(a) = r(b - a).$$

Using the idea of a definite integral to represent the accumulated increment of change, we can rewrite this (for constant rate) as

$$Q(b) - Q(a) = \int_a^b r\,dt.$$

If we rewrite our increment of change in Q so that it is solved for $Q(b)$, we

find

$$Q(b) = Q(a) + \int_a^b r\,dt = Q(a) + r(b-a).$$

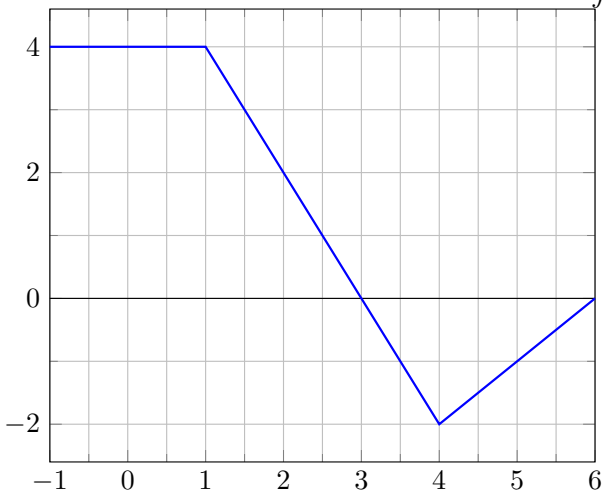
This is read as saying that $Q(b)$ equals the initial value $Q(a)$ plus the total increment of change in Q as t goes from a to b. Because this equation is true for any value of b, we can replace it with a variable and obtain

$$Q(x) = Q(a) + \int_a^x r\,dt = Q(a) + r(x-a).$$

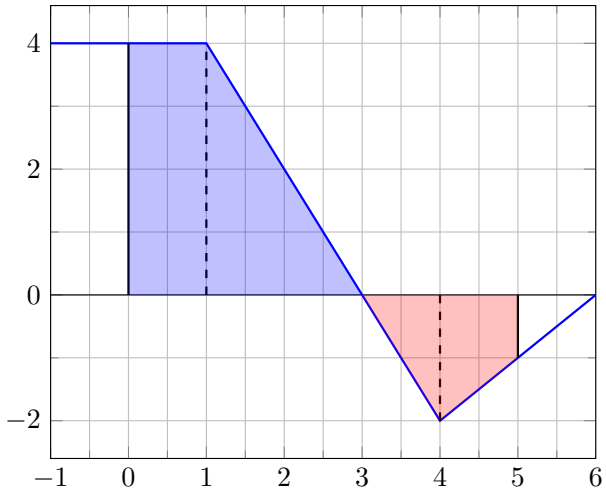
That is, we should recognize the point–slope equation of a line as a special case of an initial value plus an increment of change. □

Knowing the area of regions of a graph can allow us to compute some definite integrals.

Example 3.3.10 The graph of $y = f(x)$ is shown below. Find $\int_0^5 f(x)\,dx$.



Solution. If we consider vertical lines at $x = 0$ and $x = 5$ and then look at the regions between these lines, the graph and the x-axis, we can identify the areas that need to be calculated. Regions above the axis are shaded in blue and represent positive signed area while regions below the axis are shaded in red and represent negative signed area. In addition, dashed lines have been included to represent convenient places to split the region into simple geometric shapes.



Consider the geometric regions contained in the figure. There is a rectangle on the interval $[0, 1]$ with a height $f(x) = 4$. Since the area of the rectangle is 4 and the region is above the axis, we know

$$\int_0^1 f(x)dx = 4.$$

Next, we have a triangle on the interval $[1, 3]$ with a horizontal base given by the increment $\Delta x = 3 - 1 = 2$ and a vertical height of 4. This region is also above the axis so that

$$\int_1^3 f(x)dx = \frac{1}{2}(2)(4) = 4.$$

Combining these regions, we know

$$\int_0^3 f(x)dx = 4 + 4 = 8.$$

The area of the region below the axis can be found in several ways. One way is to identify a triangle on interval $[3, 4]$ and a trapezoid on $[4, 5]$. The triangle has horizontal width $\Delta x = 4 - 3 = 1$ and height 2 for a total area of $\frac{1}{2}(1)(2) = 1$. Since the region is below the axis, we have

$$\int_3^4 f(x)dx = -1.$$

A trapezoid is a geometric shape consisting of two parallel sides and its area is the average length of the parallel sides times the perpendicular length between those sides. The area of the trapezoid on $[4, 5]$ uses parallel lengths of 2 and 1 with a perpendicular distance $\Delta x = 5-4 = 1$ for a total area of $\frac{1}{2}(2+1)(1) = \frac{3}{2}$. As a negative signed area, we have

$$\int_4^5 f(x)dx = -\frac{3}{2}.$$

Combining these two shapes for total signed area on $[3, 5]$, we have

$$\int_3^5 f(x)dx = -1 + -\frac{3}{2} = -\frac{5}{2}.$$

Another way to find this area would be to consider a larger triangle coming from the interval $[3, 6]$ and then subtract the area of the smaller triangle on the interval $[5, 6]$ that should not be included:

$$\int_3^5 f(x)dx = -\big(\frac{1}{2}(6-3)(2) - \frac{1}{2}(6-5)(1)\big) = -\frac{5}{2}.$$

Finding the total signed area for the graph, we find the definite integral desired,

$$\int_0^5 f(x)dx = \int_0^3 f(x)dx + \int_3^5 f(x)dx = 8 + -\frac{5}{2} = \frac{11}{2}.$$

□

3.3.3 Finding Definite Integrals with Technology

When we do not have easy tricks to compute a definite integral, we can get high accuracy estimates using technology. There are free websites that can compute integrals such as WolframAlpha or SageMath described previously. Most graphing calculators have the ability to compute a definite integral and therefore the ability to compute accumulated change or signed areas. In general, you will apply the following steps on a calculator.

1. Identify the integrand function (i.e., the rate of change or the function defining signed area) and the limits of integration.
2. Use the graphing feature of your calculator so that the function is graphed and the interval of interest is showing. You may need to change the window of you graph.
3. Use the menu system to find the integral. You will need to select your function and input the end points of the interval of interest.

Example 3.3.11 We wish to compute $\int_2^5 (2^x - 8)\,dx$. First, steps are given for evaluating this using a TI-83/84 graphing calculator. This is followed by a call to WolframAlpha.com and a SageMath script that compute the same integral.

Solution.

- Find the integral using a TI-83/84 graphing calculator.
  1. Identify the function. In this example, the formula $f(x) = 2^x - 8$ represents the integrand while the interval is based on the limits of integration $[2, 5]$.
  2. Graph the function over this interval. Each step is given as a separate line.

     ```
     Y=
     Y1= 2^x-8
     WINDOW
     Xmin= 2
     Xmax= 5
     GRAPH
     ZOOM: ZoomFit
     ```
  3. Compute the definite integral.

     `CALC:` $\int f(x)dx$

`Lower Limit?` 2

`Upper Limit?` 5

The calculator reports back $\int f(x)dx = 16.3954611$.

- Find the integral using WolframAlpha.

  In the prompt box, just type "`integrate 2^x-8 from 2 to 5`". An exact and approximate answer is given

$$\int_2^5 (2^x - 8)dx = \frac{28}{\log(2)} - 24 \approx 16.3955.$$

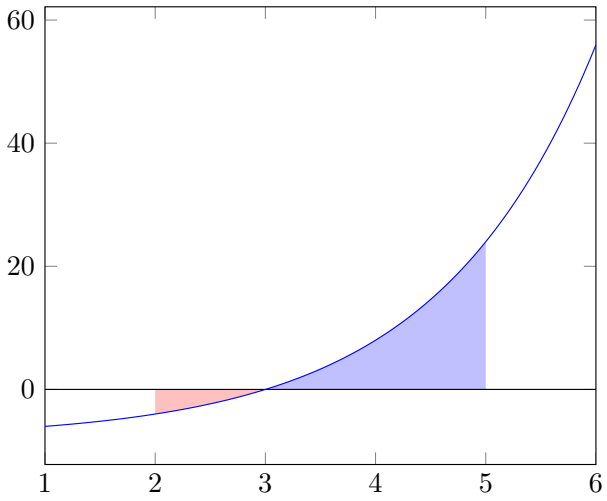
  Note that log on WolframAlpha refers to the natural logarithm ln.

- The SageMath script to compute the integral is just as similar. We need a separate command to show our result as a decimal approximation.

```
var("x")
value = integrate(2^x-8, [x,2,5])
show(value)
show(value.n())
```

□

The graph for the previous example, with the signed areas shaded, is shown below. Notice that the graph has two regions, one of which is negative (red) and one of which is positive (blue). We can find the point where the sign switches by solving $2^x - 8 = 0$ which is $x = 3$. In the next example, we will find the signed area of each interval separately and relate the values to the overall signed area.



Example 3.3.12 Compute $\int_2^3 (2^x - 8)dx$ and $\int_3^5 (2^x - 8)dx$.

Solution. As long as the interval of interest is graphed, we can compute the definite integral to get the signed area. Since we already have our integrand $f(x) = 2^x - 8$ in our calculator, we can just go to the compute steps.

- `CALC:` $\int f(x)dx$

  `Lower Limit?` 2

  `Upper Limit?` 3

The calculator reports back $\int f(x)dx = -2.22922$.

- CALC: $\int f(x)dx$
  Lower Limit? 3
  Upper Limit? 5

  The calculator reports back $\int f(x)dx = 18.624681$.

We interpret these results. The first integral was

$$\int_2^3 (2^x - 8)dx \approx -2.22922$$

and means that the area of the first region from interval $(2, 3)$ is 2.22922. Because the graph is below the axis, the integral counts as negative area. The second integral was

$$\int_3^5 (2^x - 8)dx \approx 18.624681$$

which means that the second region on interval $(3, 5)$ has area 18.624681 and counts as positive area. The total signed area is the sum of these parts,

$$\begin{aligned}\int_2^5 (2^x - 8)dx &= \int_2^3 (2^x - 8)dx + \int_3^5 (2^x - 8)dx \\ &= -2.22922 + 18.624681 = 16.395461.\end{aligned}$$

Notice that there is a slight numerical discrepancy between our two methods. This is because numerical calculation of definite integrals involves an approximation. Approximations of necessity comes with some unavoidable errors. □

We can solve some applications about change by identifying an appropriate accumulation of a rate of change. For example, velocity is a rate of change of position. Consequently, the change in position (displacement) can be computed as an accumulation of change using velocity with a definite integral.

Example 3.3.13 Suppose a hovercraft starts 40 meters away from the shore. If the velocity (meters per second) of the hovercraft is a function of time (seconds) $v(t) = t(t-3)(t-5)$ (a positive velocity is moving away from shore, increasing the distance). What is the position of the hovercraft after 3 seconds and again at 5 seconds?

Solution. Let $x(t)$ measure the position of the hovercraft (meters from shore) as a function time (seconds). By the principle of accumulation of change, the change in position over 3 seconds is equal to the definite integral of the rate of change,

$$x(3) - x(0) = \int_0^3 v(t)dt = \int_0^3 [t(t-3)(t-5)]dt.$$

Using technology, we find the accumulated change, such as asking WolframAlpha "integral of $t(t-3)(t-5)$ from 0 to 3",

$$\int_0^3 [t(t-3)(t-5)]dt = \frac{63}{4} = 15.75.$$

Since the hovercraft started 40 meters from shore and has moved a net amount 15.75 meters (away from shore), the hovercraft is at a position 55.75 meters after 3 seconds.

We can compute the change in position over the next two seconds using

$$x(5) - x(3) = \int_3^5 v(t)dt = \int_3^5 [t(t-3)(t-5)]dt.$$

Using technology, we find the accumulated change, such as asking WolframAlpha "integral of $t(t-3)(t-5)$ from 3 to 5",

$$\int_3^5 [t(t-3)(t-5)]dt = \frac{-16}{3} \approx -5.333.$$

The hovercraft has moved 5.333 meters back toward the shore during the last two seconds. Since the craft was at 55.75 meters after 3 seconds, after 5 seconds it is at 50.417 meters away from the shore.

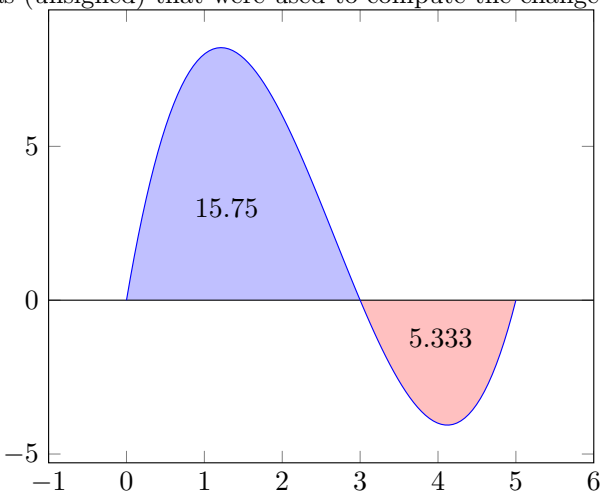
Alternatively, we could do a single integral finding the total change over all 5 seconds,

$$x(5) - x(0) = \int_0^5 v(t)dt = \int_0^5 [t(t-3)(t-5)]dt.$$

We find

$$\int_0^5 [t(t-3)(t-5)]dt = \frac{125}{12} \approx 10.417,$$

which implies that the hovercraft has moved from $x(0) = 40$ meters from shore to a new position of $x(5) = 50.417$ meters from shore, in agreement to our earlier calculations. The figure below shows a graph of the velocity function with the areas (unsigned) that were used to compute the change in position.



□

3.3.4 Summary

- Pending further edits.

3.3.5 Exercises

1. Pending.

3.4 Functions Defined by Accumulation

3.4.1 Overview

When a function $f(x)$ is integrable on an interval I and $a \in I$ is any value in that interval, then for any other value $b \in I$, we can compute the definite integral of $f(x)$ from a to b. Because the value depends on the value of b, we can think of this definite integral as a function of the upper limit b. We call such a function the accumulation function of $f(x)$ relative to $x = a$.

This section introduces how to describe accumulation functions according to the properties of the integrand (rate) function. We learn the definitions for increasing and decreasing functions as well as the definition of concavity. Using the Mean Value Theorem for Integrals, we will be able to classify the behavior of accumulation functions according to the behavior of its corresponding integrand (rate) function.

3.4.2 Accumulation Functions

In our introduction, we used the variable b as the upper limit of the accumulation function. Because we usually think of x as our default independent variable, we would like to use x as the upper limit of the integral. But then we would have the variable x playing two different roles — the upper limit of the integral and the variable of integration. To keep a single role for the variable, we always require that when using a variable in a limit of integration, the integration variable must be chosen to be a dummy variable that does not have another contextual meaning.

Definition 3.4.1 Accumulation Function. Let $f(x)$ represent a rate of change or rate of accumulation with independent variable x. The **accumulation function** $A(x)$ relative to $x = a$ with an initial value $A(a) = A_0$ is defined as

$$A(x) = A_0 + \int_a^x f(z)\,dz,$$

where z can be replaced with any other dummy variable (but not x). The function is defined so long as f is integrable on the interval containing both x and a. ◊

Once an accumulation function is defined, it can be used to evaluate particular definite integrals, even if the starting limit does not match the point used to define the accumulation.

Theorem 3.4.2 Integration as the Difference in Accumulation. *Suppose $f(x)$ is integrable on an interval that contains a, b and c. If $A(x)$ is an accumulation of $f(x)$ relative to $x = c$,*

$$A(x) = A_0 + \int_c^x f(z)\,dz,$$

then

$$\int_a^b f(x)\,dx = A(b) - A(a).$$

Notice that in the theorem, the definite integral used the independent variable x as the variable of integration. This is acceptable because the integral is a specific definite integral and the variable x plays no role other than the integration variable. We could have written using another dummy variable to

get

$$\int_a^b f(z)\,dz = A(b) - A(a),$$

but the result would have been exactly the same.

The integrand function f for an accumulation A is the rate of accumulation or rate of change. When we learn about derivatives in the next chapter, we will learn a different conception of rate of change that is called the derivative, written $A'(x)$. Fortunately, the (((Unresolved xref, reference "fundamental-theorem-calculus"; check spelling or use "provisional" attribute)))Fundamental Theorem of Calculus will show that the two different conceptions agree with one another. That is, for any accumulation function

$$A(x) = A_0 + \int_c^x f(z)\,dz,$$

the derivative $A'(x)$ and the rate of accumulation $f(x)$ are the same. For consistency of discussion later, we will call $f(x)$ the derivative of the accumulation function $A(x)$.

Some accumulation functions can be expressed simply using other well known formulas. For example, we previously discovered the following rules which we can now identify as accumulation functions.

Theorem 3.4.3 Elementary Accumulation Functions.

$$\int_0^x 1\,dz = x \tag{3.4.1}$$

$$\int_0^x z\,dz = \frac{1}{2}x^2 \tag{3.4.2}$$

$$\int_0^x z^2\,dz = \frac{1}{3}x^3 \tag{3.4.3}$$

$$\int_0^x z^3\,dz = \frac{1}{4}x^4 \tag{3.4.4}$$

That is, for a constant rate $A'(x) = 1$, the accumulated change relative to $x = 0$ is $A(x) = x$. Similarly, for the rate $A'(x) = x$, the accumulated change relative to $x = 0$ is $A(x) = \frac{1}{2}x^2$.

Example 3.4.4 Suppose $A(x)$ has a rate $A'(x) = 2x^2 - 3$ and initial value $A(0) = 4$. Express $A(x)$ in terms of a definite integral. Then apply the properties of integrals and the elementary accumulation functions to find an algebraic formula for $A(x)$.

Solution. Because we are given $A(0) = 4$, we will write

$$A(x) = A(0) + \int_0^x f(z)\,dz$$

where $f(x) = A'(x)$ is the desired rate of accumulation. That is,

$$A(x) = 4 + \int_0^x 2z^2 - 3\,dz.$$

To find the algebraic formula for $A(x)$, we will rewrite the definite integral as a linear combination of the elementary rates z^2 and 1. That is, $f(z) = 2z^2 - 3 = 2 \cdot z^2 - 3 \cdot 1$ so that the linearity property of integrals allows us to use the elementary accumulation functions.

$$A(x) = 4 + 2 \cdot \int_0^x z^2\,dz - 3 \cdot \int_0^x 1\,dz$$

$$= 4 + 2 \cdot \left(\frac{1}{3}x^3\right) - 3 \cdot (x)$$
$$= \frac{2}{3}x^3 - 3x + 4$$

□

3.4.3 Monotonicity and Concavity

We first learned to describe the monotonicity of functions in Section 3.1. Recall from Definition 3.1.1 that a function f is **increasing** on a set S if for every $x_1, x_2 \in S$,

$$x_1 < x_2 \quad \Longrightarrow \quad f(x_1) < f(x_2),$$

and decreasing if

$$x_1 < x_2 \quad \Longrightarrow \quad f(x_1) > f(x_2).$$

We can rewrite these inequalities in terms of the *increment of change* of f:

$$f(x_1) < f(x_2) \quad \Leftrightarrow \quad f(x_2) - f(x_1) > 0,$$
$$f(x_1) > f(x_2) \quad \Leftrightarrow \quad f(x_2) - f(x_1) < 0.$$

That is, an increasing function is associated with positive increments of change and a decreasing function is associated with negative increments of change.

When a function $f(x)$ is described as an accumulation, it can be written as the integral of its rate of accumulation or derivative $f'(x)$. Thus,

$$f(x_2) - f(x_1) = \int_{x_1}^{x_2} f'(x)\,dx.$$

Knowing the sign of the rate of accumulation can then be used to determine intervals of monotonicity.

Theorem 3.4.5 Monotonicity Test for Accumulation Functions. *Suppose that $f(x)$ is an accumulation function with corresponding rate function $f('x)$, and suppose that $f'(x)$ is continuous on (a, b) with limits at the endpoints.*

- *If $f'(x) = 0$ for all $x \in (a, b)$, then $f(x)$ is constant on $[a, b]$.*
- *If $f'(x) > 0$ for all $x \in (a, b)$, then $f(x)$ is increasing on $[a, b]$.*
- *If $f'(x) < 0$ for all $x \in (a, b)$, then $f(x)$ is decreasing on $[a, b]$.*

Proof. Let $x_1, x_2 \in [a, b]$ satisfy $x_1 < x_2$. Because f is an accumulation of f', we know that

$$f(x_2) - f(x_1) = \int_{x_1}^{x_2} f'(x)\,dx.$$

We now treat each case individually.

If $f'(x) = 0$ for all $x \in (a, b)$, then

$$f(x_2) - f(x_1) = \int_{x_1}^{x_2} 0\,dx = 0$$

and $f(x_2) = f(x_1)$. Because x_1 and x_2 were arbitrary, $f(x)$ must have the same value for any $x \in [a, b]$.

If $f'(x) > 0$ for all $x \in (a, b)$, then (((Unresolved xref, reference "thm-integral-inequality"; check spelling or use "provisional" attribute))) allows us to form a bound

$$\int_{x_1}^{x_2} f'(x)\,dx > \int_{x_1}^{x_2} 0\,dx = 0.$$

This guarantees that $f(x_2) - f(x_1) > 0$ so that f is increasing on $[a, b]$.

If $f'(x) < 0$ for all $x \in (a, b)$, then again (((Unresolved xref, reference "thm-integral-inequality"; check spelling or use "provisional" attribute))) allows us to form a bound

$$\int_{x_1}^{x_2} f'(x)\,dx < \int_{x_1}^{x_2} 0\,dx = 0,$$

so that $f(x_2) - f(x_1) > 0$. Thus, f is decreasing on $[a, b]$. ■

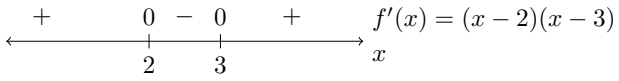
Example 3.4.6 Suppose $f(x) = \int_1^x t^2 - 5t + 6\,dt$. Describe the monotonicity of f.

Solution. The rate of accumulation for f is given by $f'(x) = x^2 - 5x + 6$. (Recall the integral uses a dummy variable; we change it back to x for analysis.) Because f' is continuous everywhere, the domain for f is $(-\infty, \infty)$. The starting location $x = 1$ simply gives the initial value with $f(1) = 0$.

We find the signs of f' by first solving $f'(x) = 0$ and then testing the resulting intervals.

$$\begin{gathered} x^2 - 5x + 6 = 0 \\ (x-2)(x-3) = 0 \\ x - 2 = 0 \text{ or } x - 3 = 0 \\ x = 2 \text{ or } x = 3 \end{gathered}$$

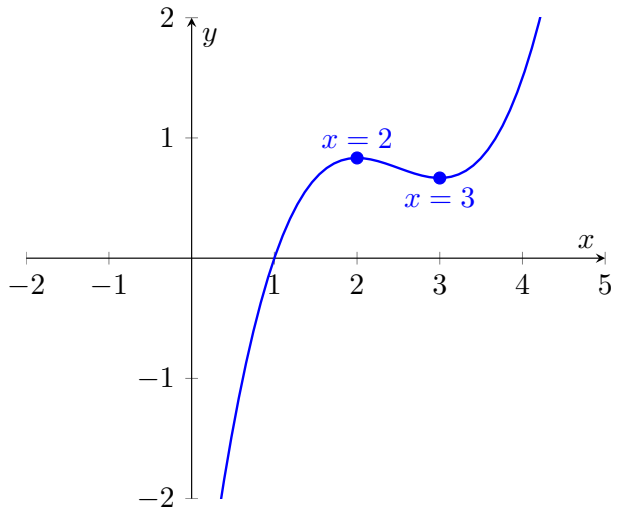
The intervals to test are $(-\infty, 2)$, $(2, 3)$, and $(3, \infty)$. We can test actual values or consider the signs of the factors on each interval. The results are summarized on the number-line summary.



We can now interpret the sign analysis of $f'(x)$.

- $f'(x) > 0$ on $(-\infty, 2)$ implies that $f(x)$ is *increasing* on $(-\infty, 2]$.
- $f'(x) < 0$ on $(2, 3)$ implies that $f(x)$ is *decreasing* on $[2, 3]$.
- $f'(x) > 0$ on $(3, \infty)$ implies that $f(x)$ is *increasing* on $[3, \infty)$.

A graph of $y = f(x)$ is shown below consistent with this analysis and the initial value $f(1) = 0$.



□

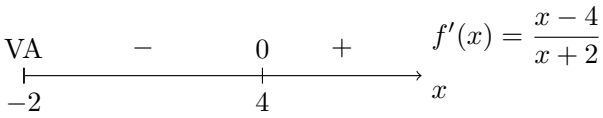
Example 3.4.7 Suppose $g(x) = \int_1^x \frac{t-4}{t+2}\,dt$. Describe the monotonicity of g.

Solution. The rate of accumulation for g is given by $g'(x) = \frac{x-4}{x+2}$. The function g' has a discontinuity at $x = -2$ which corresponds to a vertical asymptote. The domain of the accumulation function is the interval containing $x = 1$ up to this discontinuity, which is the interval $(-2, \infty)$.

To determine monotonicity, we need to find the sign of $g'(x)$. Intervals are determined by the roots and discontinuities. The root is the solution to $g'(x) = \frac{x-4}{x+2} = 0$ which occurs when $x - 4 = 0$ or $x = 4$. Using this root and the discontinuity at $x = -2$, the intervals to test are $(-2, 4)$ and $(4, \infty)$.

$$g'(0) = \frac{0-4}{0+2} = -2$$
$$g'(6) = \frac{6-4}{6+2} = \frac{1}{4}$$

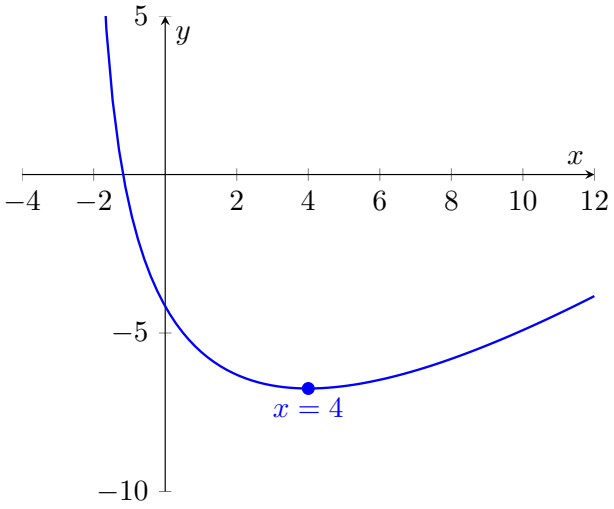
This could be summarized on a number-line as shown below.



We can now interpret the sign analysis of $g'(x)$.

- $g'(x) < 0$ on $(-2, 4)$ implies $g(x)$ is *decreasing* on $(-2, 4]$. (We can not include $x = -2$ because of the vertical asymptote.)
- $g'(x) > 0$ on $(4, \infty)$ implies $g(x)$ is *increasing* on $[4, \infty)$.

A graph of $g(x)$ is shown below that is consistent with this analysis.



□

Concavity was introduced as a way to describe how a function bends. However, our original definition of concavity 3.1.13 also involved inequalities, stating that the rate of change itself was increasing or decreasing. For an accumulation function, the rate of accumulation will control concavity. The following theorem suggests that we try to think of the rate of accumulation as being, on

its own, another accumulation of another function which we call the **second derivative** of the accumulation.

Theorem 3.4.8 Concavity Test for Accumulation Functions. *Suppose that $f(x)$ is an accumulation function with corresponding rate function $f'(x)$ and that $f'(x)$ is itself an accumulation function with its rate function $f''(x)$. Suppose that $f''(x)$ is continuous on (a,b) with limits at the endpoints.*

- *If $f''(x) = 0$ for all $x \in (a,b)$, then $f(x)$ is linear on $[a,b]$.*
- *If $f''(x) > 0$ for all $x \in (a,b)$, then $f(x)$ is concave up on $[a,b]$.*
- *If $f''(x) < 0$ for all $x \in (a,b)$, then $f(x)$ is concave down on $[a,b]$.*

Proof. We will prove that $f'(x)$ is constant, increasing, or decreasing, in each of the respective cases. In fact, this is often adopted as the de facto definition for concavity. Because f' is an accumulation with rate f'', we only need to apply Theorem 3.4.5. The relationship between the monotonicity of f' and the changes of the average rates of change given in the original definition of concavity requires the Mean Value Theorem. ■

At this point, we have not learned how to find the rate so that a function can be written as an accumulation. This requires computing derivatives. However, we can use technology to help us out.

Example 3.4.9 Use technology to find derivatives in order to describe the monotonicity and concavity of $f(x) = x^3 - 4x$.

Solution. We start by writing $f(x)$ as an accumulation. The rate of accumulation $f'(x)$ is a derivative, which we find using technology.

```
f(x) = x^3-4*x
Df(x) = derivative(f(x),x)
show(Df(x))
```

```
3*x^2-4
```

Knowing the rate $f'(x) = 3x^2 - 4$ and an initial value, say $f(0) = 0^3 - 4(0) = 0$, we can write

$$f(x) = 0 + \int_0^x 3z^2 - 4\,dz = \int_0^x 3z^2 - 4\,dz.$$

Monotonicity is determined by the signs of $f'(x) = 3x^2 - 4$.

$$\begin{aligned} 3x^2 - 4 &= 0 \\ 3x^2 &= 4 \\ x^2 &= \frac{4}{3} \\ x &= \pm\sqrt{\frac{4}{3}} \\ x &= \pm\frac{2}{\sqrt{3}} \end{aligned}$$

We can test the sign of $f'(x)$ in each resulting interval and summarize the results on a number line.

$$\underset{-\frac{2}{\sqrt{3}}}{+} \quad \underset{-\frac{2}{\sqrt{3}}}{0} \quad - \quad \underset{+\frac{2}{\sqrt{3}}}{0} \quad + \qquad f'(x) = 3x^2 - 4$$
$$\longleftrightarrow \qquad x$$

Interpreting the sign analysis tells us that $f(x) = x^3 - 4x$ is *increasing* on the interval $(-\infty, -\frac{2}{\sqrt{3}}]$, *decreasing* on the interval $[-\frac{2}{\sqrt{3}}, \frac{2}{\sqrt{3}}]$, and *increasing* on the interval $[\frac{2}{\sqrt{3}}, \infty)$.

To analyze concavity, we need to write $f'(x) = 3x^2 - 4$ as an accumulation function. Technology helps us find the derivative, which requires one additional line.

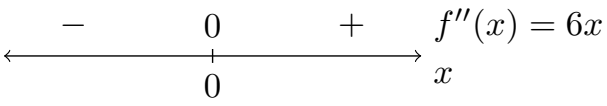
```
f(x) = x^3-4*x
Df(x) = derivative(f(x),x)
D2f(x) = derivative(Df(x),x)
show(D2f(x))
```

```
6*x
```

Now that we know $f''(x) = 6x$, and we have an initial value $f'(0) = 3(0)^2 - 4 = -4$, we can write

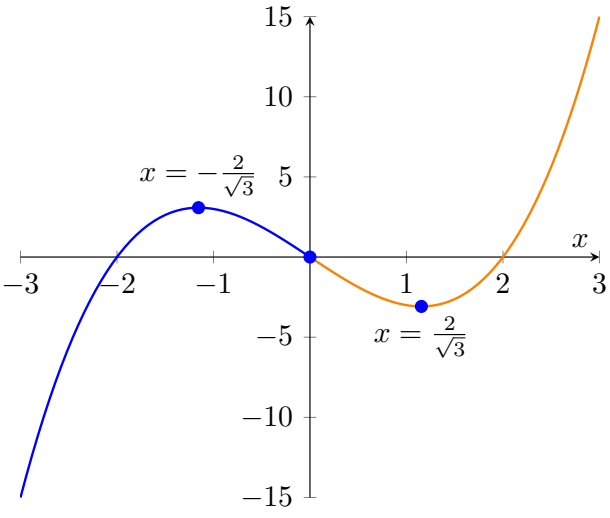
$$f'(x) = -4 + \int_0^x 6z\,dz.$$

The signs of $f''(x) = 6x$ change at $x = 0$, summarized by the sign analysis below.



Interpreting the sign analysis of the second derivative, we describe the concavity. The function $f(x)$ is concave down on $(-\infty, 0]$ and concave up on $[0, \infty)$.

A graph illustrating these features is shown below. The curve is colored differently, depending on concavity. To the left of $x = 0$ (blue), the curve is concave down. To the right of $x = 0$ (orange), the curve is concave up. The local extremes at $x = \pm\frac{2}{\sqrt{3}}$ are also labeled.



□

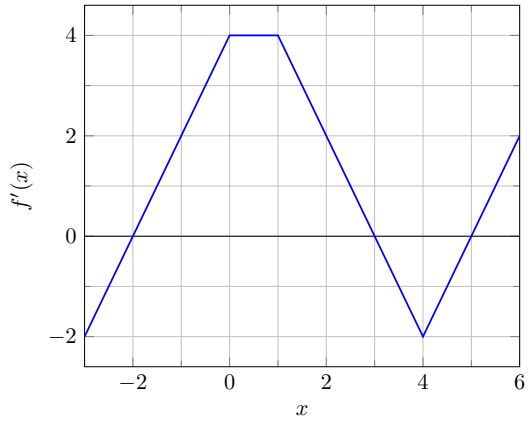
A point where the concavity of a function changes from concave up to concave down or vice versa is called a **point of inflection**, or more simply an **inflection point**. We require that an inflection point only occurs at points where the function is continuous.

Definition 3.4.10 Suppose that $f(x)$ is a function that is continuous at $x = c$ and that there are intervals so that f is concave up on (a, c) and concave down on (c, b), or the reverse, concave down on (a, c) and concave up on (c, b). We say that f has a **point of inflection** at $x = c$. ◊

An inflection point occurs at the points where $f''(x)$ changes sign. On a graph, this is where the curve transitions between bending upward and bending downward. Inflection points are significant because they represent points where the rate of change $f'(x)$ reaches its extreme values.

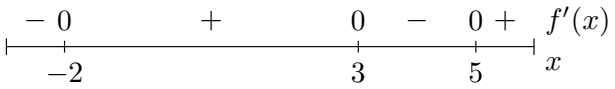
When we can view a graph of the rate of accumulation (the derivative), we can interpret the features of that rate to describe the behavior of the accumulation function itself.

Example 3.4.11 The graph of the accumulation rate $f'(x)$ for a function $f(x)$ is shown in the figure below. Describe the monotonicity and concavity and sketch a graph of the accumulation function $y = f(x)$ with initial value $f(0) = 3$.



Solution. The graph of $f'(x)$ can be used to determine the signs of $f'(x)$ that are used to find the monotonicity of $f(x)$ while the monotonicity of $f'(x)$ can be used to find the concavity of $f(x)$. Computing the signed area of the graph can be used to determine the actual increments of change.

The signs of $f'(x)$ based on the graph are summarized on the following number-line summary.



We interpret this to make the following conclusions about monotonicity: $f(x)$ is

- increasing on intervals $[-2, 3]$ and $[5, 6]$,
- decreasing on intervals $[-3, -2]$ and $[3, 5]$.

From the graph, we can not determine what happens beyond the visible window.

In addition, the graph allows us to identify the monotonicity of f'. We see that $f'(x)$ is

- increasing on intervals $[-3, 0]$ and $[4, 6]$,
- constant on the interval $[0, 1]$,
- decreasing on the interval $[1, 4]$.

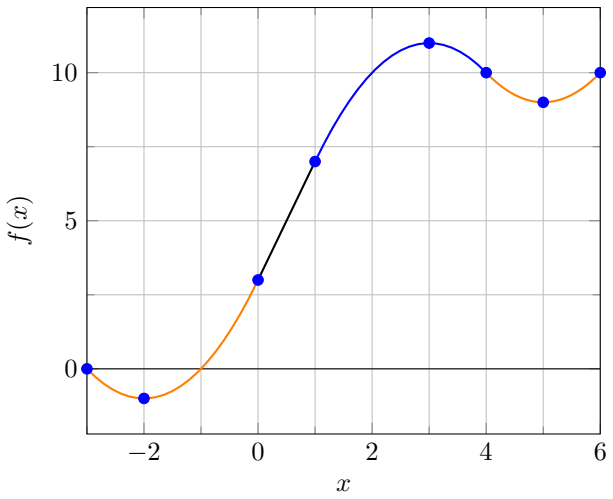
We interpret this to give us concavity: $f(x)$ is

- concave up on intervals $[-3, 0]$ and $[4, 6]$,
- linear on the interval $[0, 1]$ with slope $f'(x) = 4$,
- concave down on the interval $[1, 4]$.

The definite integral of $f'(x)$ over an interval, which computes the signed area, determines the increment of change in $f(x)$. The graph of $f'(x)$ is made of straight line segments, so we can compute the integrals using simple geometric formulas for the areas of triangles, rectangles, and trapezoids.

$$\begin{aligned}
f(-2) - f(-3) &= \int_{-3}^{-2} f'(x)\,dx \\
&= -\tfrac{1}{2}(1)(2) = -1 \\
f(0) - f(-2) &= \int_{-2}^{0} f'(x)\,dx \\
&= +\tfrac{1}{2}(2)(4) = 4 \\
f(1) - f(0) &= \int_{0}^{1} f'(x)\,dx \\
&= +(1)(4) = 4 \\
f(3) - f(1) &= \int_{1}^{3} f'(x)\,dx \\
&= +\tfrac{1}{2}(2)(4) = 4 \\
f(4) - f(3) &= \int_{3}^{4} f'(x)\,dx \\
&= -\tfrac{1}{2}(1)(2) = -1 \\
f(5) - f(4) &= \int_{4}^{5} f'(x)\,dx \\
&= -\tfrac{1}{2}(1)(2) = -1 \\
f(6) - f(5) &= \int_{5}^{6} f'(x)\,dx \\
&= +\tfrac{1}{2}(1)(2) = 1
\end{aligned}$$

The initial value $f(0) = 3$ gives us a starting point for the graph. We can use the increments computed from the definite integrals to find the values of $f(x)$ at several specific points. For example, because $f(1) - f(0) = 4$, we know that $f(1) = 7$. If we start by plotting these points, we can sketch the graph of $y = f(x)$ by including shapes consistent with the monotonicity and concavity of f. Where the monotonicity changes, the graph of $f(x)$ reaches an extreme value. Where the concavity changes, the graph of $f(x)$ has an inflection point. To emphasize concavity, different concavity regions are colored differently—orange for concave up, blue for concave down, and black for linear.



□

As we conclude this section, make note of the relationship between our understanding the behavior of sequences in terms of increments and accumulation sequences with the behavior of accumulation functions in terms of the derivative or rate of accumulation. Where the behavior of a sequence is described in terms of a range of index values, the behavior of a function is described in terms of an interval.

3.4.4 Summary

- An accumulation function $A(x)$ is a function defined using a definite integral in order to have a given rate of accumulation $f(x)$ and initial value $A(x_0) = A_0$,

$$A(x) = A_0 + \int_{x_0}^{x} f(z)\,dz.$$

  The integration variable z is a dummy variable and must be different from the independent variable.

- The accumulation rate $f(x)$ will later be shown (Fundamental Theorem of Calculus) to be the derivative of the accumulation function $A(x)$ so that we will write $A'(x) = f(x)$.

- Knowing an accumulation function can be used to compute definite integrals of the accumulation rate,

$$\int_a^b f(x)\,dx = A(b) - A(a).$$

  See Theorem 3.4.2.

- A function $f(x)$ that can be written as an accumulation with rate (derivative) $f'(x)$ has a monotonicity determined by the sign of $f'(x)$ on intervals.

  - $f'(x) > 0$ on (a, b) implies $f(x)$ is increasing on $[a, b]$
  - $f'(x) = 0$ on (a, b) implies $f(x)$ is constant on $[a, b]$
  - $f'(x) < 0$ on (a, b) implies $f(x)$ is decreasing on $[a, b]$

- A function $f(x)$ that can be written as an accumulation with rate (derivative) $f'(x)$ which itself can also be written as an accumulation $f''(x)$ (second derivative of f) has concavity determined by the sign of $f''(x)$ on intervals.
  - $f''(x) > 0$ on (a, b) implies $f'(x)$ is increasing on $[a, b]$ and $f(x)$ is concave up on $[a, b]$
  - $f''(x) = 0$ on (a, b) implies $f'(x)$ is constant on $[a, b]$ and $f(x)$ is linear on $[a, b]$
  - $f''(x) < 0$ on (a, b) implies $f'(x)$ is decreasing on $[a, b]$ and $f(x)$ is concave down on $[a, b]$
- A point where $f(x)$ changes concavity is called a **point of inflection** or **inflection point**. An inflection point represents where the rate of accumulation reaches an extreme value.

3.4.5 Exercises

Express each accumulation function with its given rate of accumulation and initial value as a formula involving a definite integral. Then, using the properties of definite integrals and the elementary accumulation functions, find the algebraic formula.

1. Find $f(x)$ with $f'(x) = 2x + 5$ and $f(0) = 2$.

2. Find $g(x)$ with $g'(x) = x^2 - 4x$ and $g(0) = -5$.

3. Find $A(x)$ with $A'(x) = x$ and $A(2) = 0$.

4. Find $P(t)$ with $P'(t) = 2t + 5$ and $P(2) = 5$.

5. Find $Q(t)$ with $Q'(t) = t^2 + 5t$ and $Q(1) = 2$.

For each accumulation function, describe the monotonicity.

6. $f(x) = \displaystyle\int_2^x 4z - 7\,dz$

7. $g(t) = \displaystyle\int_1^t 5 - 2x\,dx$

8. $A(x) = 2 + \displaystyle\int_{-3}^x 9 - u^2\,du$

9. $Q(x) = -3 + \displaystyle\int_0^x \frac{2z}{z^2 - 4}\,dz$

10. $R(t) = \displaystyle\int_{-1}^x \frac{2z}{3z - 1}\,dz$

Use technology to find derivatives in order to describe the monotonicity and concavity of each function. Compare your results to a graph of $y = f(x)$.

11. $f(x) = x^2 - 12x + 32$

12. $f(x) = x^3 - 12x + 4$

13. $f(x) = x^3 + 6x^2 - 15x$

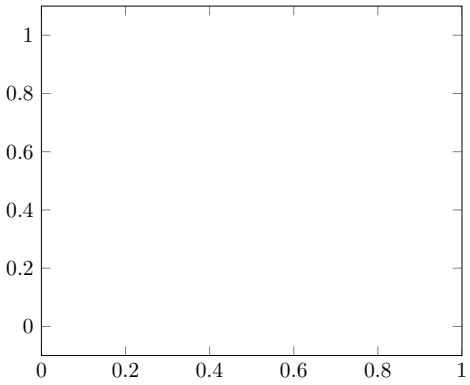
14. $f(x) = e^{2x} - 4x$

15. $f(x) = 5xe^{-3x}$

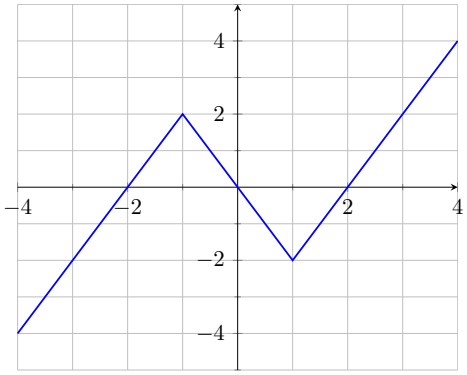
16. $f(x) = \dfrac{\ln(x)}{x}$

Each figure represents the graph of the derivative or rate of accumulation. Describe the monotonicity and concavity of the corresponding accumulation function and sketch a graph consistent with the given initial value.

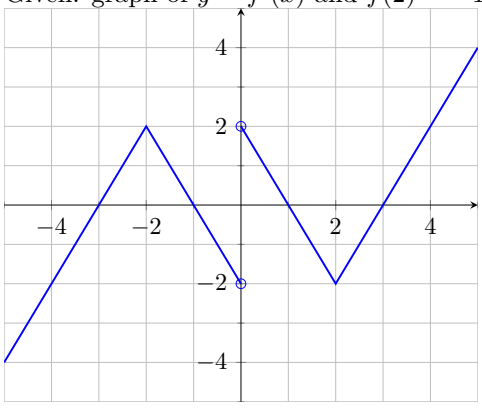
17.



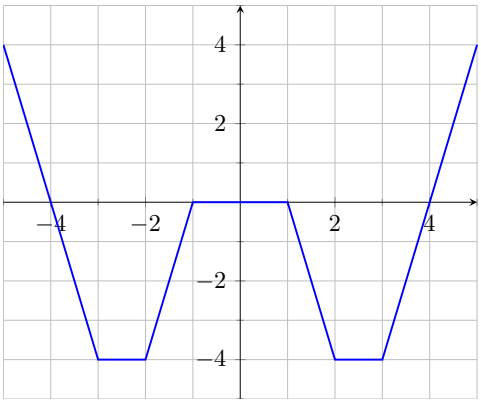
Given: graph of $y = f'(x)$ and $f(0) = 4$.



18. Given: graph of $y = f'(x)$ and $f(2) = -1$.



19. Given: graph of $y = f'(x)$ and $f(1) = 3$.


4
2
−4
−2
2
4
−2
−4

Chapter 4

Sequences and Accumulation

4.1 Introduction to Sequences

Overview. Sequences are often introduced to us as examples for finding basic numerical patterns. We are shown the start to a list of numbers and asked if we can identify the next few numbers in the list or are asked to identify the rule being used to generate the sequence.

$$1, 5, 9, 13, \ldots$$
$$2, 6, 18, 54, \ldots$$

Do you see the patterns?

You probably recognized that in the first sequence, the next number would be 17 because the pattern involved adding 4 to the previous number. In the second sequence, you probably saw that we were multiplying by the value 3, so that the next number would have been 162. Not all sequences follow patterns. However, we use examples such as these to motivate the mathematical definition of a sequence.

We study sequences because they illustrate a number of ideas we will use in calculus. We eventually want to describe functions as dynamic models. Dynamic models for sequences are easier to illustrate than for general functions.

This section introduces the basic terminology for sequences. It explains how a sequence is a special type of function, where the domain is a set of integers. We will learn about explicit formulas for a sequence and recursive formulas for a sequence, using arithmetic and geometric sequences as our original motivation.

Later in this chapter, we will explore the dynamic ideas that will motivate calculus. Sequences that converge to a single value will be used to introduce the concept of limits. Recursive formulas for sequences will be used to introduce the ideas of accumulation which ultimately motivates the concept of integration. The dynamic behavior of a sequence will be analyzed in terms of its increment sequence which will motivate the calculus concept of the derivative.

4.1.1 Basic Terminology and Notation

A **sequence** is an ordered collection of numbers. The idea of being ordered is that we can say what the first number is, what the second number is, and so forth. To emphasize that the number have assigned positions, a sequence can be written as an **ordered list** using parentheses. The entire sequence can be assigned a symbol, just like a variable, so that a sequence assigned a symbol x and given by the values 1, 5, 9, 13, etc., would be written

$$x = (1, 5, 9, 13, \ldots).$$

Because the sequence has a specific order, we use an **index** as a way of counting through the sequence. For a given sequence, the term with index 1 is the first number of the sequence, the term with index 2 is the second number, the term with index 3 is the third number, and so forth. We use subscripts on a sequence to refer to an indexed value. So x_1 is the first value of sequence x and x_5 refers to the value of the sequence x with index 5.

Example 4.1.1 For the sequence $x = (1, 5, 9, 13, \ldots)$ and assuming the pattern continues, find each of the following values: x_1, x_3, and x_5.

Solution. The ordering of the list of values in the sequence can be made explicit with a table.

ordinal position	index	sequence value
first	1	1
second	2	5
third	3	9
fourth	4	13
fifth	5	17

Of course, you probably thought through the ordering in your head rather than make a table. From this ordering, we know that $x_1 = 1$, $x_3 = 9$, and $x_5 = 17$. □

The table in the solution for the previous example illustrates an explicit association between the index and the sequence value. We could reorganize this table to create a **mapping** between values.

n	1	2	3	4	5	6
↓						
x_n	1	5	9	13	17	21

The mapping can be illustrated using two number lines, one for the index and the other for the sequence values, with arrows drawn from the index to the corresponding sequence value. We write $n \mapsto x_n$ to indicate that we have a mapping that goes from a value of n to a value x_n.

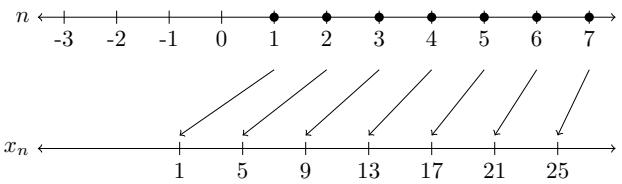


Figure 4.1.2 Illustration of the example sequence as a map $n \mapsto x_n$.

Another name for a mapping is a **function**. Sequences are functions whose domains correspond to an interval of the integers. The **domain** for a sequence is the set of possible values for the index. An interval of integers corresponds to a subset of integers with no gaps. The interval could be finite, as in $\{4, \ldots, 10\}$, or it could be infinite, as in $\{4, \ldots, \infty\}$. The usual domain for sequences is the set of natural numbers $D = \mathbb{N} = \{1, 2, 3, \ldots, \infty\}$. We often also want to include an initial value corresponding to an index value $n = 0$, in which case our domain is the extended natural numbers $D = \mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$.

Definition 4.1.3 Sequence. A **sequence** x is a function with a domain D that is an interval of integers and values that are real numbers. We can write this in symbols using mapping notation,

$$x : n \in D \mapsto x_n \in \mathbb{R}.$$

♢

The mapping notation used in the definition of a sequence is a symbolic representation of the statement that a sequence is a function or a map. In particular, it says there is a map ($\mapsto$) named x that takes a value n from the set D ($n \in D$) and returns a value x_n from the set of real numbers $\mathbb{R}$ ($x_n \in \mathbb{R}$).

Defining a sequence as a function allows us more flexibility in what we include as sequences. The new definition allows us to have our first index value start at a value other than 1. It also allows us to use other variables for our index. The variable used for an index is most often a letter from the middle of the alphabet.

Example 4.1.4 Interpret the statement $u : k \in \{0, \ldots, 10\} \mapsto u_k \in \mathbb{R}$.

Solution. The map u defines a sequence with index values k going from $k = 0$ to $k = 10$. Because we do not have more information, we do not yet know the values of this sequence. □

Although mapping notation is useful to remind us that a sequence is a map or function, it can be a little cumbersome to use all the time. Mathematicians developed a more concise representation that reminds us of an ordered list. If we consider an interval of integers $\{n_i, \ldots, n_f\}$ (n_i is the *initial* value in the interval and n_f is the *final* value in the interval), then the mapping notation

$$x : n \in \{n_i, \ldots, n_f\} \mapsto x_n \in \mathbb{R}$$

is equivalent to the more compact sequence notation

$$x = (x_n)_{n=n_i}^{n_f}.$$

Example 4.1.5 Rewrite $u : k \in \{0, \ldots, 10\} \mapsto u_k \in \mathbb{R}$ using sequence notation.

Solution. We would write $u = (u_k)_{k=0}^{10}$. □

Sequence notation can be coupled with an ordered list of values to define a sequence that follows a pattern with an index that starts at a value other than 1.

Example 4.1.6 Interpret

$$w = (w_j)_{j=4}^{\infty} = (1, 2, 4, 8, 16, \ldots),$$

assuming the sequence follows a simple pattern.

Solution. This sequence notation tells us that w is a sequence, the index variable is j, and the interval of integers used for the index starts at $j = 4$ and continues through all integers greater than 4. The first few terms in a table showing the mapping are given below.

j	4	5	6	7	8	9
$\downarrow$						
w_j	1	2	4	8	16	32

For this sequence, w_1, w_2, and w_3 are not defined because the values 1, 2, and 3 are not in the domain interval. □

Example 4.1.7 Interpret the sequence

$$u = (u_k)_{k=-1}^{\infty} = (8, 5, 2, -1, -4, \ldots),$$

assuming the sequence follows a simple pattern.

Solution. We have defined a sequence u with an index variable k. The first index value is $k = -1$. The sequence has the following values:

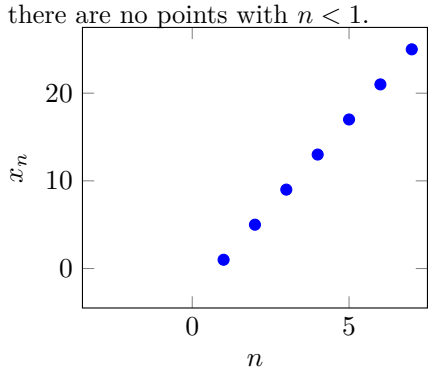
$$\begin{aligned} u_{-1} &= 8, \\ u_0 &= 5, \\ u_1 &= 2, \\ u_2 &= -1, \\ u_3 &= -4, \\ u_4 &= -7. \end{aligned}$$

□

4.1.2 Graphs of Sequences

We can create a graph anytime we can find a relation between two variables. For a sequence x, there is a natural choice for the two variables—the index n and the value x_n. The natural graph for a sequence x consists of the points (n, x_n). Because the index comes from a domain that is an interval of integers, the graph will be a collection of isolated points. This is why a sequence is called a discrete model.

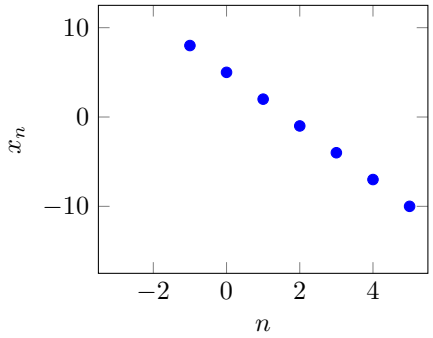
Example 4.1.8 The graph of the sequence $x = (1, 5, 9, 13, \ldots)$ consists of the points (n, x_n). The first few points of the graph—$(1, 1)$, $(2, 5)$, $(3, 9)$, and $(4, 13)$—are shown in the figure below. The sequence continues with addition points for $n > 4$, but there are no points with $n < 1$.



□

Example 4.1.9 Create the graph for $u = (u_k)_{k=-1}^{\infty} = (8, 5, 2, -1, -4, \ldots)$.

Solution. The points in the graph use an index starting at $k = -1$. They include $(-1, 8)$, $(0, 5)$, $(1, 2)$, $(2, -1)$, and $(3, -4)$. The sequence continues to the right of these points.



□

4.1.3 Explicit Sequence Representations

We sometimes have an **explicit representation** for a sequence, where the value of the sequence is a dependent variable in terms of the index as the independent variable. The expression defining the dependent variable could be used with each of the different notations.

Example 4.1.10 Each of the following notations define the same sequence.

$$x : n \in \{1, \ldots, \infty\} \mapsto x_n = \frac{n}{n+1}$$

$$x = \left(x_n = \frac{n}{n+1} \right)_{n=1}^{\infty}$$

$$x = \left(\frac{n}{n+1}\right)_{n=1}^{\infty}$$

We could simplify even further and write

$$x_n = \frac{n}{n+1}, n = 1, \ldots, \infty,$$

as the subscript notation x_n itself implies we have a sequence.

Writing the sequence value as a dependent variable provides a compact way of representing the entire sequence. To find a particular value of the sequence, we substitute the value for the independent variable into the expression.

$$\begin{aligned} x_1 = x(1) &= \frac{1}{1+1} = \frac{1}{2}, \\ x_2 = x(2) &= \frac{2}{2+1} = \frac{2}{3}, \\ x_{10} = x(10) &= \frac{10}{10+1} = \frac{10}{11}. \end{aligned}$$

□

In addition to substitution using actual integer values, we can also use substitution of expressions that have integer values. This includes using other variables that have integer values. To do this, we substitute the expression that appears in the subscript for every occurrence of the index in the expression.

Example 4.1.11 For the sequence defined by

$$x_n = \frac{n}{n+1}, n = 1, \ldots, \infty,$$

find the expressions defined by x_k, x_{n+1}, x_{2n}, and x_{n^2}.

Solution. With this interpretation, we can even do composition to find the sequence value at an index defined by a formula. Substituting the variable k for the index n in the dependent variable's expression, we find

$$x_k = \frac{k}{k+1}.$$

In a similar way, we substitute the expressions $n+1$, $2n$, and n^2 in the formula where n originally appeared to obtain

$$\begin{aligned} x_{n+1} &= \frac{(n+1)}{(n+1)+1} = \frac{n+1}{n+2}, \\ x_{2n} = x(2n) &= \frac{2n}{2n+1}, \\ x_{n^2} = x(n^2) &= \frac{n^2}{n^2+1}. \end{aligned}$$

□

Some sequences have patterns where we can easily find an explicit formula by recognizing how the numbers defining the sequence values relate to the index.

Example 4.1.12 Find an explicit formula for the sequence

$$x = (\frac{1}{4}, \frac{1}{9}, \frac{1}{16}, \frac{1}{25}, \ldots),$$

and then find x_{12} and x_{2n}.

Solution. To find the explicit formula, we look for a pattern in the sequence and then try to find a relationship between the index and the pattern. Because the sequence domain was not specified, it is understood to be the natural numbers $\mathbb{N} = (1, \ldots, \infty)$. In this case, every sequence value is the reciprocal of a perfect square. If we look at this pattern with a table showing the index and the pattern, we find a relationship.

n	1	2	3	4
x_n	$\frac{1}{4} = \frac{1}{2^2}$	$\frac{1}{9} = \frac{1}{3^2}$	$\frac{1}{16} = \frac{1}{4^2}$	$\frac{1}{25} = \frac{1}{5^2}$

The pattern is that the number that is squared is always 1 greater than the index. So the explicit formula for this sequence is given by

$$x_n = \frac{1}{(n+1)^2}, \quad n \in \{1, 2, 3, 4, \ldots\}.$$

Using this explicit formula, we can find the desired values.

$$x_{12} = x(12) = \frac{1}{(12+1)^2} = \frac{1}{169}$$
$$x_{2n} = x(2n) = \frac{1}{(2n+1)^2}$$

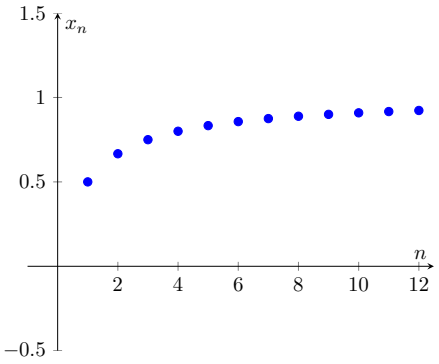
□

Knowing the explicit formula for a sequence, we can compute the values of the sequence to use in a graph.

Example 4.1.13 Graph the sequence $x_n = \dfrac{n}{n+1}$, defined for $n = 1, 2, 3, \ldots$.

Solution. This is the sequence discussed above. The plot will include the points

$$\{(n, x_n) : x_n = \frac{n}{n+1}, n = 1, 2, 3, \ldots\} = \{(1, \frac{1}{2}), (2, \frac{2}{3}), (3, \frac{3}{4}), (4, \frac{4}{5}), \ldots\}.$$



□

4.1.4 Summary

- Sequences are functions with domains that are intervals of integers. The independent variable (input) is called the **index**, and the dependent variable (output) is called the **value**. The value of a sequence x at index n is represented using subscripts for the index x_n.

- An **explicit representation** of a sequence x is when the function or map $n \mapsto x_n$ can be written with x_n as a dependent variable in terms of

the index n.

- Sequence evaluation with an explicit formula involves substitution of the index variable by whatever expression appears in the subscript position.
- The standard graph of a sequence x uses the points (n, x_n) where n is chosen from the domain set of the sequence.

4.1.5 Exercises

In the following group of exercises, a sequence is defined. Identify the variable representing the sequence, the variable representing the index, and the domain or interval of integers the values of the index come from.

1. $u : i \in \{-3, -2, \ldots, 5\} \mapsto u_i \in \mathbb{R}$

2. $v : n \in \{5, \ldots, \infty\} \mapsto v_n \in \mathbb{R}$

3. $z = (z_k)_{k=2}^{\infty}$

4. $M = (M_t)_{t=-\infty}^{10}$

In the following group of exercises, a sequence with a pattern is given. Identify the values of the requested terms from the sequence. Create a graph that includes the first ten values from the sequence.

5. $x = (x_n)_{n=1}^{\infty} = (2, 4, 6, 8, \ldots)$
Find x_3, x_5, and x_7.

6. $y = (y_k)_{k=0}^{\infty} = (12, 9, 6, 3, \ldots)$
Find y_1, y_4, and y_6.

7. $w = (w_i)_{i=-2}^{\infty} = (24, 12, 6, 3, \ldots)$
Find w_0, w_2, and w_4.

8. $P = (P_t)_{t=0}^{\infty} = (100, 110, 125, 145, 170, \ldots)$
Find P_1, P_4, and P_6.

Find an explicit formula for each of the following sequences by identifying patterns relating the index and the expressions shown for the values.

9. $x = (x_n)_{n=0}^{\infty} = (1, 4, 9, 16, 25, \ldots)$

10. $y = (y_n)_{n=1}^{\infty} = (\frac{1}{4}, \frac{2}{9}, \frac{3}{16}, \frac{4}{25}, \ldots)$

11. $z = (z_n)_{n=0}^{\infty} = (0, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots)$

In each of the following exercises, a sequence is defined explicitly. Evaluate the requested expressions.

12. $x_n = -3n + 20, \qquad n = 0, 1, 2, 3, \ldots$
Find x_0, x_1, and x_2.
Evaluate x_{k+2} and $x_k + 2$.

13. $y_k = \frac{2^{k-2}}{3^k}, \qquad k = 0, 1, 2, 3, \ldots$
Find y_0, y_1, and y_2.
Evaluate y_{n+1}, y_{k+1}, and $\dfrac{y_{k+1}}{y_k}$.

4.2 Increments of Sequences

4.2.1 Overview

In the introductory section of this chapter, we learned that the increments of a sequence, calculated using the backward difference, can be used to analyze the monotonicity and concavity of a sequence. Those examples all focused on sequences with given values.

In this section, we continue our study of increments by looking at sequences defined explicitly and recursively. For a sequence defined with an explicit formula, we will compute the increments using that formula and index substitution. For a sequence defined with a projection function, we will compute the increments as a function of the previous sequence value. Once a formula or function for the increments has been computed, we will solve inequalities to characterize the monotonicity and concavity of the sequence.

4.2.2 Increments of Explicit Sequences

When we know the explicit formula for a sequence x_n, we can find a corresponding formula for the increment of that sequence ∇x_n. Recall that an explicit formula gives us a function mapping the value of the index to the value of the sequence,

$$n \mapsto x_n.$$

We can think of n and x_n as state variables. We can also think of x_{n-1} as a state variable, one that represents the *previous* value of the sequence. The expanded state of the system becomes (n, x_n, x_{n-1}). We want to include yet another state variable, the increment, ∇x_n, which is defined by the backward difference

$$\nabla x_n = x_n - x_{n-1}.$$

We can find explicit formulas for these additional variables by making a substitution on the index. Suppose the map $n \mapsto x_n$ were a function, $S : n \mapsto x_n$. The symbol $S(n)$ would represent the explicit formula for x_n. Then $S(n-1)$ would represent the formula for x_{n-1}, calculated by substituting the expression $n-1$ everywhere the original variable n appeared in the formula. The process of substituting an expression in the place of the independent variable of a function is called **composition**.

Example 4.2.1 Consider the sequence defined explicitly,

$$x = (3n + 5)_{n=0}^{\infty}.$$

Find explicit formulas for x_{n-1} and ∇x_n.

Solution. The explicit formula for the sequence, $x_n = 3n + 5$, defines a function,

$$S(n) = 3n + 5.$$

The independent variable in the function is a placeholder for the input expression,

$$S(\square) = 3\square + 5.$$

We can find the formula for the previous term using a substitution $\square = n - 1$,

$$x_{n-1} = S(n-1) = 3(n-1) + 5.$$

Simplifying the expression to a sum, this gives

$$x_{n-1} = 3n + 2.$$

Formally, because x has its first index $n = 0$, there is no value x_{-1}. The formula for x_{n-1} is only valid for $n = 1, 2, \ldots$.

The increment ∇x_n is defined by the backward difference $\nabla x_n = x_n - x_{n-1}$. To calculate the backward difference, we substitute the explicit formulas in place of the state variables x_n and x_{n-1} and simplify:

$$\begin{aligned}\nabla x_n &= x_n - x_{n-1} \\ &= \big(3n + 5\big) - \big(3n + 2\big) \\ &= 3n + 5 - 3n - 2 \\ &= 3.\end{aligned}$$

Again, this only applies for $n = 1, 2, \ldots$. Because the increments were constant, we realize that x was an arithmetic sequence with $\beta = 3$.

To illustrate the connection between the formulas with which we are now working and the actual values of the sequence, consider the actual values of the sequence,

$$x = (5, 8, 11, 14, 17, \ldots).$$

Now, consider the table created using the explicit formulas above.

n	x_n	x_{n-1}
0	$3(0) + 5 = 5$	undefined
1	$3(1) + 5 = 8$	$3(1) + 2 = 5$
2	$3(2) + 5 = 11$	$3(2) + 2 = 8$
3	$3(3) + 5 = 14$	$3(3) + 2 = 11$

You should notice how the formula for x_{n-1} uses the *current* value of n to find the value of the *previous* value of the sequence. □

Example 4.2.2 Consider the sequence defined explicitly,

$$u = (n^2 + 2n)_{n=0}^{\infty}.$$

Find explicit formulas for u_{n-1} and ∇u_n.

Solution. The explicit formula for the sequence, $u_n = n^2 + 2n$, defines a function,

$$S(n) = n^2 + 2n.$$

The independent variable in the function is a placeholder for the input expression,

$$S(\square) = \square^2 + 2\square.$$

We can find the formula for the previous term using a substitution $\square = n - 1$,

$$u_{n-1} = S(n - 1) = (n - 1)^2 + 2(n - 1).$$

Expanding the square and then simplifying the expression to a sum, this gives

$$\begin{aligned}u_{n-1} &= (n - 1)(n - 1) + 2(n - 1) \\ &= n^2 - 2n + 1 + 2n - 2 \\ &= n^2 - 1\end{aligned}$$

The increment ∇u_n is defined by the backward difference $\nabla u_n = u_n - u_{n-1}$. To calculate the backward difference, we substitute the explicit formulas in place of the state variables u_n and u_{n-1} and simplify:

$$\nabla u_n = u_n - u_{n-1}$$

$$\begin{aligned} &= \left(n^2 + 2n\right) - \left(n^2 - 1\right) \\ &= n^2 + 2n - n^2 + 1 \\ &= 2n + 1. \end{aligned}$$

We can illustrate that the formulas using a table. Notice that the formula u_{n-1} calculate the previous value using the current index, and the formula for ∇u_n calculates the increment of the sequence using the index.

n	u_n	u_{n-1}	∇u_n
0	$(0)^2 + 2(0) = 0$	undefined	undefined
1	$(1)^2 + 2(1) = 3$	$(1)^2 - 1 = 0$	$2(1) + 1 = 3$
2	$(2)^2 + 2(2) = 8$	$(2)^2 - 1 = 3$	$2(2) + 1 = 5$
3	$(3)^2 + 2(3) = 15$	$(3)^2 - 1 = 8$	$2(3) + 1 = 7$

□

4.2.3 Increments of Recursive Sequences

When a sequence is defined recursively, we know that there is a projection function $f : x_{n-1} \mapsto x_n$. That is, the sequence value x_n can be found using the previous value x_{n-1} through the projection function,

$$x_n = f(x_{n-1}).$$

Instead of depending on the index, the increment is computed in terms of the previous value,

$$\nabla x_n = x_n - x_{n-1} = f(x_{n-1}) - x_{n-1}.$$

This suggests that we have another function, $g : x_{n-1} \mapsto \nabla x_n$, defined by

$$g(x) = f(x) - x,$$

which projects the increment instead of the new sequence value. We might call this function the **increment projection function**.

Example 4.2.3 A sequence is defined recursively by the recurrence relation

$$x_n = 1.25x_{n-1} - 10.$$

Find the formula for the increment in terms of x_{n-1}.

Solution. The recurrence relation is already in the form of a recursive equation with projection function $f(x) = 1.25x - 10$. The increment $\nabla x_n = x_n - x_{n-1}$ is computed by subtracting the x_{n-1} from the formula for x_n:

$$\nabla x_n = x_n - x_{n-1} = \left(1.25x_{n-1} - 10\right) - x_{n-1}.$$

Simplifying this formula gives

$$\nabla x_n = 0.25x_{n-1} - 10,$$

corresponding to an increment projection function $g(x) = f(x) - x = 0.25x - 10$.

We can illustrate the role of these formulas by creating a table of a sequence. Suppose the initial value is $x_0 = 20$. We can compute both x_n and ∇x_n in terms of the previously computed value x_{n-1}.

n	x_n	∇x_n
0	20	undefined
1	$1.25(20) - 10 = 15$	$0.25(20) - 10 = -5$
2	$1.25(15) - 10 = 8.75$	$0.25(15) - 10 = -6.25$
3	$1.25(8.75) - 10 = 0.9375$	$0.25(8.75) - 10 = -7.8125$

Suppose we had only used the recursive formula to find the sequence. We would have found

$$x = (x_n)_{n=0}^{\infty} = (20, 15, 8.75, 0.9375 \ldots).$$

Then if we found the increments directly, we would have subtracted consecutive terms and found

$$\nabla x = (\nabla x_n)_{n=1}^{\infty} = (-5, -6.25, -7.8125, \ldots),$$

in agreement with the calculations using the increment projection formula. □

Example 4.2.4 A sequence is defined recursively by a projection function

$$f(x) = 1.25x - 0.05x^2.$$

Find the formula for the increment as a function of the previous sequence value.

Solution. Knowing the sequence's projection function, the increment projection function is given by

$$\begin{aligned} g(x) &= f(x) - x \\ &= 1.25x - 0.05x^2 - x \\ &= 0.25x - 0.05x^2. \end{aligned}$$

This means that the increment is computed as $g : x_{n-1} \mapsto \nabla x_n$, or

$$\nabla x_n = 0.25x_{n-1} - 0.05x_{n-1}^2.$$

□

4.2.4 Analysis of Monotonicity and Concavity

When we have formulas to compute the increments, we can solve inequalities to determine under what conditions the increments are positive or negative. We can use the solutions of these inequalities to analyze where a sequence is increasing or decreasing. If we also compute the second backward difference, or the increments of the increments, then solving an additional inequality allows us to analyze the concavity of the sequence.

There are many ways to solve an inequality. One approach is to isolate the independent variable use balanced operations. Inequalities have a complication in that balanced multiplication (or division) by a negative number reverses the inequality. Another approach that works for continuous functions is to solve an equation in order to create intervals to test. Using the principle of continuity of formulas, which we will justify later in this text, we can check one point in as a representative for each interval. Because the first approach only works in some cases, we will emphasize practicing using the second approach which works more generally. We will learn later in the text how to deal with inequalities involving discontinuous functions.

Example 4.2.5 Determine the intervals of monotonicity and concavity for the sequence

$$x = (40n - n^2)_{n=0}^{\infty}.$$

Identify any local extremes.

Solution. The explicit formula $x_n = 40n - n^2$ allows us to compute formulas for the previous term and the increment. Notice the use of parentheses to em-

phasize the role of grouped terms, especially when there will be a subtraction.

$$\begin{aligned} x_{n-1} &= 40(n-1) - (n-1)^2 \\ &= 40(n-1) - (n-1)(n-1) \\ &= \big(40n - 40\big) - \big(n^2 - 2n + 1\big) \\ &= 40n - 40 - n^2 + 2n - 1 \\ &= 42n - n^2 - 41 \end{aligned}$$

$$\begin{aligned} \nabla x_n &= x_n - x_{n-1} \\ &= \big(40n - n^2\big) - \big(42n - n^2 - 41\big) \\ &= 40n - n^2 - 42n + n^2 + 41 \\ &= -2n + 41 \end{aligned}$$

We can verify that our work looks correct by starting a table and checking whether the explicit formulas match what the terms should be.

n	x_n	x_{n-1}	∇x_n
0	$40(0) - 0^2 = 0$	undefined	undefined
1	$40(1) - 1^2 = 39$	$42(1) - 1^2 - 41 = 0$	$-2(1) + 41 = 39$
2	$40(2) - 2^2 = 76$	$42(2) - 2^2 - 41 = 39$	$-2(2) + 41 = 37$

Checking thes few values in the table gives us confidence that we did the algebra correctly. The formula for the previous sequence value is matching what we expect, as is the formula for the increment.

Now that we have a formula for the increments, we want to find the intervals where the increments are positive or negative. This corresponds to solving inequalities $\nabla x_n > 0$ and $\nabla x_n < 0$. The increment is defined for index values $n = 1, 2, \ldots$.

The approach of solving an inequality by isolating the independent variable would go as follows. Start with the inequality in terms of the independent variable n, because we have an explicit definition for the sequence. To solve $\nabla x_n > 0$, we use balanced operations to create equivalent inequalities.

$$\begin{gathered} \nabla x_n > 0 \\ -2n + 41 > 0 \\ -2n > -41 \\ \frac{-2n}{-2} < \frac{-41}{-2} \\ n < 20\tfrac{1}{2} \end{gathered}$$

When we divided both sides by -2 (multiplied by $-\frac{1}{2}$), the equivalent relation showed a reversed inequality. The other inequality $\nabla x_n < 0$ follows the same steps, resulting in the equivalent inequality

$$\nabla x_n < 0 \quad \Leftrightarrow \quad n > 20\tfrac{1}{2}.$$

The alternate approach involves solving the equation $\nabla x_n = -2n + 41 = 0$. Solving the equation involves the same steps to give an equivalent equation

$$\nabla x_n = 0 \quad \Leftrightarrow \quad n = 20\frac{1}{2}.$$

We now consider the intervals of values for n on either side of this value. The intervals are $\{1, \ldots, 20\}$ and $\{21, \ldots, \infty\}$. The principle for solving the

inequality is to choose one value from each interval and use it to find the sign of ∇x_n. For example, we can use $n = 10$ and $n = 25$.

$$\nabla x_{10} = -2(10) + 41 = 21$$
$$\nabla x_{25} = -2(25) + 41 = -9$$

Both methods of solving the inequalities give the same intervals, which allow us to analyze the monotonicity of the sequence as shown in the table below.

Sign of ∇x_n	Monotonicity of x_n
Positive on $\{1, \ldots, 20\}$	Increasing on $\{0, \ldots, 20\}$
Negative on $\{21, \ldots, \infty\}$	Decreasing on $\{20, \ldots, \infty\}$

Because x is increasing on $\{0, \ldots, 20\}$ and then decreasing on $\{20, \ldots, \infty\}$, x must have a maximum value at $n = 20$. The value of the sequence at that index is

$$x_{20} = 40(20) - 20^2 = 800 - 400 = 400.$$

To find concavity, we need to compute the second backward difference. This is computed like other backward differences.

$$\begin{aligned}
\nabla^2 x_n &= \nabla x_n - \nabla x_{n-1} \\
&= \big(-2n + 41\big) - \big(-2(n-1) + 41\big) \\
&= \big(-2n + 41\big) - \big(-2n + 43\big) \\
&= -2n + 41 + 2n - 43 \\
&= -2
\end{aligned}$$

The second backward difference is always negative, for $n = 2, 3, \ldots$. Consequently, x is concave down on $\{0, \ldots, \infty\}$. □

One of the things you might notice is that completing analysis of a sequence is an involved process. You might be used to thinking that mathematics questions should have answers that take a limited amount of work. Complex questions might therefore seem overwhelming. Have confidence in your ability and develop a pattern of perseverance. Develop a pattern of big picture steps, breaking the overall problem into a series of manageable tasks.

Example 4.2.6 Determine the intervals of monotonicity and concavity for the sequence

$$z = (n^3 - 70n^2 + 1000n)_{n=-\infty}^{\infty}.$$

Identify any local extremes.

Solution. We review the big picture steps.

1. Compute the backward difference ∇z_n.
2. Solve the equation $\nabla z_n = 0$ to create test intervals.
3. Test the sign of ∇z_n in the intervals.
4. Interpret the monotonicity and extreme values of the sequence based on the sign analysis.
5. Compute the second backward difference $\nabla^2 z_n$.
6. Solve the equation $\nabla^2 z_n = 0$ to create test intervals.
7. Test the sign of $\nabla^2 z_n$ in the intervals.

8. Interpret the concavity of the sequence based on the sign analysis.

The explicit formula $z_n = n^3 - 70n^2 + 1000n$ is used to compute the formulas for the previous term and the increment.

$$\begin{aligned} z_{n-1} &= (n-1)^3 - 70(n-1)^2 + 1000(n-1) \\ &= (n-1)(n-1)(n-1) - 70(n-1)(n-1) + 1000(n-1) \\ &= (n-1)(n^2 - 2n + 1) - 70(n^2 - 2n + 1) + 1000n - 1000 \\ &= n^3 - 3n^2 + 3n - 1 - 70n^2 + 140n - 70 + 1000n - 1000 \\ &= n^3 - 73n^2 + 1143n - 1071 \end{aligned}$$

$$\begin{aligned} \nabla z_n &= z_n - z_{n-1} \\ &= \left(n^3 - 70n^2 + 1000n\right) - \left(n^3 - 73n^2 + 1143n - 1071\right) \\ &= n^3 - 70n^2 + 1000n - n^3 + 73n^2 - 1143n + 1071 \\ &= 3n^2 - 143n + 1071 \end{aligned}$$

Solving the equation $\nabla z_n = 0$ to identify our test intervals requires solving the quadratic equation

$$\nabla z_n = 3n^2 - 143n + 1071 = 0.$$

We use the quadratic formula:

$$\begin{aligned} n &= \frac{-(-143) \pm \sqrt{(-143)^2 - 4(3)(1071)}}{2(3)} \\ &= \frac{143 \pm \sqrt{7597}}{6}. \end{aligned}$$

To find the intervals, we need decimal approximations.

$$\begin{aligned} n_1 &= \frac{143 - \sqrt{7597}}{6} \approx 9.3065 \\ n_2 &= \frac{143 + \sqrt{7597}}{6} \approx 38.3601 \end{aligned}$$

The sequence is defined for an index interval $\{-\infty, \ldots, \infty\}$. These two breakpoints separate the interval into three test intervals:

$$\{-\infty, 9\}, \quad \{10, \ldots, 38\}, \quad \{39, \ldots, \infty\}.$$

We perform sign analysis by choosing a test value for the index n from each interval and identifying the sign of ∇z_n.

$$\begin{aligned} n = 0: &\quad \nabla z_0 = 3(0)^2 - 143(0) + 1071 = 1071 \\ n = 10: &\quad \nabla z_{10} = 3(10)^2 - 143(10) + 1071 = -59 \\ n = 40: &\quad \nabla z_{40} = 3(40)^2 - 143(40) + 1071 = 151 \end{aligned}$$

We can interpret these results:

1. Because $\nabla z_n > 0$ for all n in $\{-\infty, \ldots, 9\}$, we know z_n is *increasing* on the interval $\{-\infty, \ldots, 9\}$.

   Because $\nabla z_n < 0$ for all n in $\{10, \ldots, 38\}$, we know z_n is *decreasing* on the interval $\{9, \ldots, 38\}$.

   Because $\nabla z_n > 0$ for all n in $\{39, \ldots, \infty\}$, we know z_n is *increasing* on the interval $\{38, \ldots, \infty\}$.

The turning points correspond to local extreme values. The value z_9 is greater than values to its left and right and is a *local maximum*. The value z_{38} is less than values to its left and right and is a *local minimum*. Because z is decreasing on $\{-\infty, \ldots, 9\}$ and increasing on $\{38, \ldots, \infty\}$, we do not yet know if the sequence surpasses these values to determine global extreme values.

To analyze concavity, we repeat the process for the second backward difference.

$$\begin{aligned}\nabla z_{n-1} &= 3(n-1)^2 - 143(n-1) + 1071 \\ &= 3(n^2 - 2n + 1) - 143(n-1) + 1071 \\ &= 3n^2 - 6n + 3 - 143n + 143 + 1071 \\ &= 3n^2 - 149n + 1217\end{aligned}$$

$$\begin{aligned}\nabla^2 z_n &= \nabla z_n - \nabla z_{n-1} \\ &= \big(3n^2 - 143n + 1071\big) - \big(3n^2 - 149n + 1217\big) \\ &= 3n^2 - 143n + 1071 - 3n^2 + 149n - 1217 \\ &= 6n - 146\end{aligned}$$

Solving the equation $\nabla^2 z_n = 0$ gives

$$\begin{aligned}6n - 146 &= 0 \\ 6n &= 146 \\ n &= \frac{146}{6} = \frac{73}{3} \\ n &= 24\tfrac{1}{3}\end{aligned}$$

The intervals to test are separated by this value, $\{-\infty, \ldots, 24\}$ and $\{25, \ldots, \infty\}$. Test one point in each interval:

$$\begin{aligned}\nabla^2 z_0 &= 6(0) - 146 = -146, \\ \nabla^2 z_{25} &= 6(25) - 146 = 4.\end{aligned}$$

Now we can interpret our results.

- Because $\nabla^2 z_n < 0$ for all n in $\{-\infty, \ldots, 24\}$, we know z_n is *concave down* on the interval $\{-\infty, \ldots, 24\}$.

  Because $\nabla^2 z_n > 0$ for all n in $\{25, \ldots, \infty\}$, we know z_n is *concave up* on the interval $\{23, \ldots, \infty\}$.

□

4.2.5 Behavior of Recursive Sequences

When a sequence is defined recursively through a projection function, we found that we could create an increment projection function $g(x) = f(x) - x$. Because this does not directly give us any information about the index, we can not describe the interval of integers on which the sequence is increasing or decreasing. Instead we can describe which sequence values will lead to an increase or decrease in the next step.

Theorem 4.2.7 *Suppose a sequence u is defined recursively with $f : u_{n-1} \mapsto u_n$.*

- *If $f(x) > x$, or equivalently $f(x) - x > 0$, then $u_n = x$ implies u* increases *on $\{n, n+1\}$.*
- *If $f(x) < x$, or equivalently $f(x) - x < 0$, then $u_n = x$ implies u* decreases *on $\{n, n+1\}$.*
- *If $f(x) = x$, or equivalently $f(x) - x = 0$, then $u_n = x$ implies u is* constant*. In this case, we call x a **fixed point** of f and an **equilibrium** for u.*

Concavity requires comparing two increments, so we would need two projections into the future. Given u_n, we know $u_{n+1} = f(u_n)$ and $u_{n+2} = f(u_{n+1})$. Using composition of the function with itself, we discover

$$u_{n+2} = f\big(f(u_n)\big).$$

We can now compute the increments:

$$\begin{aligned} \nabla_{n+1} &= u_{n+1} - u_n \\ &= f(u_n) - u_n \\ \nabla_{n+2} &= u_{n+2} - u_{n+1} \\ &= f\big(f(u_n)\big) - f(u_n) \end{aligned}$$

If $u_n = x$, then the second backward difference is computed as

$$\begin{aligned} \nabla^2 u_{n+2} &= \nabla u_{n+2} - \nabla u_{n+1} \\ &= \Big(f\big(f(x)\big) - f(x)\Big) - \Big(f(x) - x\Big) \\ &= f\big(f(x)\big) - 2f(x) + x. \end{aligned}$$

Sign analysis on this formula allows us to answer questions about concavity involving consecutive increments.

Theorem 4.2.8 *Suppose a sequence u is defined recursively with $f : u_n \mapsto u_{n+1}$. Define the second-order increment projection function $h(x) = f\big(f(x)\big) - 2f(x) + x$.*

- *If $h(x) > 0$, then $u_n = x$ implies u is* concave up *on $\{n, n+1, n+2\}$.*
- *If $h(x) < 0$, then $u_n = x$ implies u is* concave down *on $\{n, n+1, n+2\}$.*
- *If $h(x) = 0$, then $u_n = x$ implies u is* linear *(constant increments) on $\{n, n+1, n+2\}$.*

Example 4.2.9 For a recursive sequence u defined by projection function $f(x) = 1.25x - 10$, describe the conditions for which the sequence is increasing, decreasing, concave up, or concave down.

Solution. The increment projection is defined by $g(x) = f(x) - x = 0.25x - 10$. We analyze the inequalities $g(x) > 0$ and $g(x) < 0$ by solving the equation $g(x) = 0$ and then doing sign analysis on resulting test intervals.

$$\begin{aligned} 0.25x - 10 &= 0 \\ 0.25x &= 10 \\ x &= 40 \end{aligned}$$

We now know that $x = 40$ is an equilibrium for the sequence. Our test intervals are $x < 40$ and $x > 40$.

$$g(30) = 0.25(30) - 10 = -2.5$$

$$g(50) = 0.25(50) - 10 = 2.5$$

Consequently, the sequence will decrease for an initial value $u_n < 40$ and increase for an initial value $u_n > 40$.

Analysis of concavity is more involved, requiring the calculation of the composition $f\big(f(x)\big)$. We emphasize the importance of thinking of this as substitution, with $f(\square) = 1.25\square - 10$. For an initial value $u_n = x$, the projection of the sequence value u_{n+1} is given by

$$u_{n+1} = f(x) = 1.25x - 10$$

Projecting a second step into the future for u_{n+2} is given by

$$\begin{aligned} u_{n+2} &= f\big(f(x)\big) = f(1.25x - 10) \\ &= 1.25(1.25x - 10) - 10 = 1.5625x - 12.5 - 10 \\ &= 1.5625x - 22.5 \end{aligned}$$

This gives the increment $\nabla u_{n+2} = u_{n+2} - u_{n+1}$ as

$$\begin{aligned} \nabla u_{n+2} &= \big(1.5625x - 22.5\big) - \big(1.25x - 10\big) \\ &= 0.3125x - 12.5. \end{aligned}$$

The second backward difference is therefore

$$\begin{aligned} \nabla^2 u_{n+2} &= \nabla u_{n+2} - \nabla u_{n+1} \\ &= \big(0.3125x - 12.5\big) - \big(0.25x - 10\big) \\ &= 0.0625x - 2.5. \end{aligned}$$

Solving the equation $\nabla^2 u_{n+2} = 0$ (Try it!) gives $x = 40$, giving us the same test intervals as our sign analysis for monotonicity.

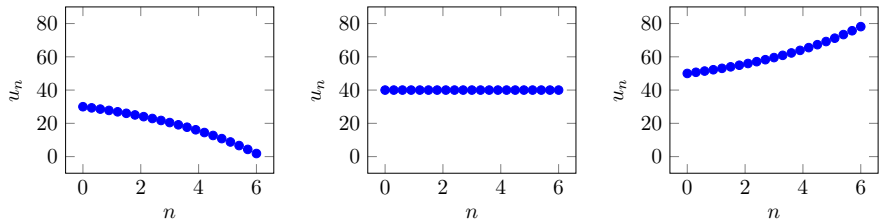
$$\begin{aligned} x = 0 \ (x < 40): &\quad \nabla^2 u_{n+2} = 0.0625(0) - 2.5 = -2.5 \\ x = 50 \ (x > 40): &\quad \nabla^2 u_{n+2} = 0.0625(50) - 2.5 = 0.625 \end{aligned}$$

Consequently, the sequence will be concave down for an initial value $u_n < 40$ and concave up for an initial value $u_n > 40$.

To visualize these results, consider the sequence with three different initial values.

$$\begin{aligned} u_0 = 30: &\quad u = (30, 27.5, 24.375, 20.46875, \ldots) \\ u_0 = 40: &\quad u = (40, 40, 40, 40, \ldots) \\ u_0 = 50: &\quad u = (50, 52.5, 55.625, 59.53125, \ldots) \end{aligned}$$

Graphs of these sequence are shown below. The first sequence is decreasing and concave down. The second sequence is constant (an equilibrium value). The third sequence is increasing and concave up.



(a) $u_0 = 30$ **(b)** $u_0 = 40$ **(c)** $u_0 = 50$

Figure 4.2.10 The sequence u defined by $u_{n+1} = 1.25u_n - 10$ and selected initial values.

□

4.2.6 Summary

- Explicit formulas for the values of a sequence x, $n \mapsto x_n$, allow us to compute an explicit formula for the increments $n \mapsto \nabla x_n$ using the backward difference
$$\nabla x_n = x_n - x_{n-1}$$
using *substitution*, or *composition*, with the expression $n-1$ in place of the index variable n. An explicit formula for the second backward difference $n \mapsto \nabla^2 x_n$ can be computed using substitution and the formula of the increments,
$$\nabla^2 x_n = \nabla x_n - \nabla x_{n-1}.$$

- Using an explicit formula $n \mapsto \nabla x_n$, we can use inequalities to perform sign analysis of the increments ∇x_n. Sign analysis provides intervals for the index n where $\nabla x_n > 0$, $\nabla x_n = 0$, and $\nabla x_n < 0$. We use these intervals to determine intervals for the index n where x_n is increasing, constant, or decreasing, respectively.

- Using an explicit formula $n \mapsto \nabla^2 x_n$, we can use inequalities to perform sign analysis of the increments $\nabla^2 x_n$. Sign analysis provides intervals for the index n where $\nabla^2 x_n > 0$, $\nabla^2 x_n = 0$, and $\nabla^2 x_n < 0$. We use these intervals to determine intervals for the index n where x_n is concave up, linear, or concave down, respectively.

- A general strategy for solving inequalities with continuous functions is to solve the corresponding equation. Solutions to the equation create the end-points of test intervals. We then choose one test point from each interval to determine the inequality and every other value in the interval will satisfy the same relation as the test point.

  In simple cases, an inequality can be solved more quickly by isolating the variable using balanced operations. Multiplication or division by a negative value reverses any inequalities. Multiplication by an expression is problematic if that expression might be negative—the inequality then reverses only for some values of the variable. In such cases, the general strategy is preferred.

- Using a recursive formula defined by a projection function $f : x_n \mapsto x_{n+1}$, we can create an increment projection function $g : x_n \mapsto \nabla x_{n+1}$, defined by
$$g(x) = f(x) - x.$$
Sign analysis on $g(x)$ determines intervals for *initial values* at which a sequence would increase or decrease to the *next* value. Any values where $g(x) = 0$ are called **fixed points** of the projection function f and correspond to **equilibrium values** of the recursive sequence.

  It is also possible to create a second-order increment projection function $h : x_n \mapsto \nabla^2 x_{n+2}$ defined by
$$h(x) = f\big(f(x)\big) - 2f(x) + x.$$
Sign analysis of $h(x)$ determines initial values where the first two *increments* are increasing, constant, or decreasing.

4.2.7 Exercises

Practice using composition (i.e., substitution) to find explicit formulas. Simplify to a form that is a sum of terms.

1. If $a_n = 3n - 5$, find a_{n-1} and a_{n+1}.

2. If $b_k = k^2 - 20k$, find b_{k-1} and b_{k+1}.

3. If $c_n = 2n^2 - 15n + 3$, find c_{n-1} and c_{n+1}.

For the each sequence, compute the explicit formula for the backward difference, perform sign analysis, and interpret the monotonicity of the sequence. Identify any local extreme values.

4. $x = (25 - 4k)_{k=0}^{\infty}$

5. $z = (j^2 - 40j + 10)_{k=0}^{\infty}$

6. $u = (40n - 3n^2)_{n=-5}^{\infty}$

7. $w = (k^3 - 500k)_{k=-\infty}^{\infty}$

For the each sequence, compute the explicit formula for the second backward difference, perform sign analysis, and interpret the concavity of the sequence. (These are the same sequences as in the previous exercise group.)

8. $x = (25 - 4k)_{k=0}^{\infty}$

9. $z = (j^2 - 40j + 10)_{k=0}^{\infty}$

10. $u = (40n - 3n^2)_{n=-5}^{\infty}$

11. $w = (k^3 - 500k)_{k=-\infty}^{\infty}$

For each recursively defined sequence, identify initial values that will result in an increase or a decrease or are equilibrium values.

12. $u_{n+1} = 50 - 3u_n$

13. $v_{k+1} = 1.1v_k - 30$

14. $w_{n+1} = 1.2w_n - 0.04w_n^2$

15. $z_{n+1} = 4z_n e^{-0.2z_n}$

16. $P_{n+1} = \frac{50P_n}{P_n+20}$, restricted to $P \geq 0$.

Applications.

17. You are about to receive some money (inheritance, lottery, etc.) and plan to invest it in an account that earns 5% annually, compounded quarterly. Your plan is to withdraw \$9000 each quarter (\$3000 per month). You want to analyze what will happen to your investment.

- Create a recursive definition for a sequence that represents the quarterly balance of your investment.
- Analyze the monotonicity and concavity of your sequence.
  - What size of an investment would result in an equilibrium?
  - What will happen to the investment if you receive less than the equilibrium amount?
  - What will happen to the investment if you receive more than the equilibrium amount?

18. A population of at risk birds has a constant per capita yearly death rate of 1 death per four individuals, $d = 0.25$. The per capita yearly birth rate is observed to be a decreasing function of the population size P, modeled by a linear function $b = 0.5 - 0.0002P$.

- Create a recursive definition for a sequence that represents the annual population size.
- Analyze the monotonicity of your sequence.
  - What is the equilibrium population size?
  - What will happen to the population if it begins below equilibrium?
  - What will happen to the population if it begins above equilibrium?
- Create a cobweb diagram for the sequence. How does the cobweb diagram relate to your analysis of monotonicity? How does the cobweb diagram relate to concavity

Suppose that the tail feathers of these birds are valuable so that poachers come and kill an additional 100 birds per year.

- Create a recursive definition for a new sequence that models the natural births and deaths as well as the illegal harvesting by poachers.
- Analyze the monotonicity of the modified sequence. What does the model predict for the consequence of poaching?

4.3 Accumulation Sequences

Overview. One of the most important mathematical ideas in calculus is that of an accumulation of change for physical quantities. As we have been learning about sequences, we have talked about how we can define sequences using explicit formulas and using recursive definitions. More recently, we have looked at how the increments of a sequence can help us understand the behavior of a sequence. For some sequences, we learned that patterns in the increments could be used to find additional terms in a sequence.

We are now ready to think about this more generally. Given any sequence of values, we wish to find that sequence for which the given sequence matches the increments. We call the sequence that we are finding the **accumulation sequence** of the given sequence.

In this section, we formally define and discuss the theory of accumulation sequences. Summation notation is introduced. We establish conditions that guarantee two sequences are equivalent. Then we illustrate applying these conditions to demonstrate that the explicit and recursive definitions for arithmetic and geometric sequences are equivalent.

4.3.1 Accumulation of Change

There are many examples of quantities where we track changes to the quantity rather than repeated measure the quantity itself. Consider a bank balance. We do not count our money every month. Instead, we add up all of our deposits and withdrawals and use them to adjust our record for the balance. Similarly, consider a population under study. It could be very costly to count all of the individuals every month. If instead we could track how many births and deaths occurred during the month, we could calculate a new population count by adding births and subtracting deaths.

Example 4.3.1 At the start of the year, you had \$1500 in an account. Suppose that the sequence

$$W = (W_m)_{m=1}^{12} = (240, 300, 270, 450, 250, 310, 360, 270, 320, 300, 350, 480)$$

represents the total of monthly withdrawals from the account, and the sequence

$$D = (D_m)_{m=1}^{12} = (280, 280, 280, 280, 280, 280, 280, 280, 280, 280, 280, 280)$$

represents the total of monthly deposits into the account. Find the sequence of monthly balances in the account.

Solution. Let B represent the monthly balance. Before any months pass, we have a balance of 1500 dollars. This gives an initial value $B_0 = 1500$. We wish to define the sequence $B = (B_m)_{m=0}^{12}$.

After one month, our account has had \$240 withdrawn and \$280 deposited. The balance after the end of the month is thus given by

$$B_1 = B_0 - W_1 + D_1 = 1500 - 240 + 280 = 1540.$$

Once we have the balance after one month, we can repeat this process for the other eleven months.

m (month)	B_m (balance in dollars)
0	1500
1	$1500 - 240 + 280 = 1540$
2	$1540 - 300 + 280 = 1520$
3	$1520 - 270 + 280 = 1530$
4	$1530 - 450 + 280 = 1360$
5	$1360 - 250 + 280 = 1390$
6	$1390 - 310 + 280 = 1360$
7	$1360 - 360 + 280 = 1280$
8	$1280 - 270 + 280 = 1290$
9	$1290 - 320 + 280 = 1250$
10	$1250 - 300 + 280 = 1230$
11	$1250 - 350 + 280 = 1160$
12	$1160 - 480 + 280 = 960$

□

When we create a sequence of values based on knowing the increments, we are creating what we call an **accumulation sequence**.

Definition 4.3.2 Given a sequence $x = (x_k)_{k=m}^{n}$, we say u is an **accumulation sequence** of x if $u = (u_k)_{k=m-1}^{n}$ with $\nabla u_k = x_k$. ◊

4.3.2 Equivalent Sequences

A given sequence of increments has infinitely many different accumulation sequences which differ in their initial value. However, for a given initial value and sequence of increments, the resulting accumulation sequence is unique. That is, any two sequences that have the same initial value and increments sequences that are equal for all values, then the sequences themselves are equal for all values.

Theorem 4.3.3 Uniqueness Conditions for Accumulation Sequences. *Given two sequences u and w. If $u_m = w_m$ and $\nabla u_k = \nabla w_k$ for all $k > m$, then $u_k = w_k$ for all $k \geq m$.*

Proof. In mathematics, to prove that every statement from a sequence of statements is true, we often use an approach called the **Principle of Mathematical Induction**. This requires demonstrating that the first statement in the sequence is true, and then showing that anytime one of the statements is true, the subsequent statement must also be true.

This theorem is perfectly suited to apply mathematical induction. The sequence of statements we wish to prove is

$$u_k = w_k, \quad k = m, m+1, \ldots.$$

The first statement in the sequence, $u_m = w_m$ is true by assumption—one condition is that the sequences u and w have the same initial values. The inductive step is to go from an arbitrary statement in the sequence of statements to the next. So suppose $u_k = w_k$ for some index k in $\{m, m+1, \ldots\}$. We know that $\nabla u_{k+1} = \nabla w_{k+1}$ by the assumption that the sequences have equal increments. We now use substitution twice:

$$\begin{aligned} u_{k+1} &= u_k + \nabla u_{k+1} \\ &= w_k + \nabla w_{k+1} \\ &= w_{k+1}. \end{aligned}$$

This shows that $u_{k+1} = w_{k+1}$ whenever $u_k = w_k$. By mathematical induction, the entire sequence of statements must be true. ■

Example 4.3.4 Consider the explicitly defined sequence $x = (3k+4)_{k=1}^{\infty}$ and the sequence $y = (y_n)_{n=1}^{\infty}$ defined recursively with an initial value $y_1 = 7$ and iteration $y_n = y_{n-1} + 3$ for $n = 2, 3, \ldots$. Show that x and y represent the same sequence.

Solution. To apply Theorem 4.3.3, we need to show that the sequences have the same initial value and the same increments. We just show the two initial values and verify they are the same.

$$x_1 = 3(1) + 4 = 7$$
$$y_1 = 7$$

We next compare the increments. Using the explicit formula for x, we find

$$\begin{aligned} \nabla x_k &= x_k - x_{k-1} \\ &= \big(3k+4\big) - \big(3(k-1)+4\big) \\ &= 3k + 4 - (3k - 3 + 4) \\ &= 3. \end{aligned}$$

Using the recursive formula for y, we find

$$\nabla y_k = y_k - y_{k-1} = 3.$$

The sequences x and y have the same initial value and the same increments. Therefore, they have all the same values: $x_k = y_k$ for all $k = 1, 2, \ldots$. □

Theorem Theorem 4.3.3 can be generalized from having two sequences with equal increments to two sequences sharing any recurrence relation involving the previous term. For example, a geometric sequence has a recurrence relation $x_n = \rho x_{n-1}$, so that the increment using the relation itself depends on the previous term, $\nabla x_n = (\rho - 1)x_{n-1}$.

Theorem 4.3.5 *Suppose u and w are two sequences with common initial values, $u_m = w_m$. If there is a sequence of projection functions f_k so that u and w satisfy the same relations,*

$$u_k = f_k(u_{k-1})$$

and

$$w_k = f_k(w_{k-1}),$$

then $u_k = w_k$ for all $k = m, m+1, \ldots$.

For a recursively defined sequence, the sequence of projection functions would all be the same function.

Example 4.3.6 Consider the explicitly defined sequence $x = (10 \cdot \frac{1}{2^k})_{k=1}^{\infty}$ and the sequence $y = (y_n)_{n=1}^{\infty}$ defined recursively with an initial value $y_1 = 5$ and iteration $y_n = \frac{1}{2} y_{n-1}$ for $n = 2, 3, \ldots$. Show that x and y represent the same sequence.

Solution. To apply Theorem 4.3.5, we need to show that the sequences have the same initial value and the satisfy the same recurrence relations. The initial values are:

$$x_1 = 10 \cdot \frac{1}{2^1} = 5,$$
$$y_1 = 5.$$

We next compare the recurrence relations. We know that y has projection

function $f : y_{n-1} \mapsto y_n = \frac{1}{2}y_{n-1}$. We need to show that x satisfies the same relation, $x_k = \frac{1}{2}x_{k-1}$. Using the explicit formula for x, we compute both sides of the recurrence equation and show they are equivalent.

$$\begin{aligned} x_k &= 10 \cdot \frac{1}{2^k} \\ \frac{1}{2}x_{k-1} &= \frac{1}{2} \cdot 10 \cdot \frac{1}{2^{(k-1)}} \\ &= 10 \cdot \frac{1}{2} \cdot \frac{1}{2^{(k-1)}} \\ &= 10 \cdot \frac{1}{2^{(k-1)+1}} \\ &= 10 \cdot \frac{1}{2^k} \end{aligned}$$

Comparing the formulas, we see that $x_k = \frac{1}{2}x_{k-1}$.

The sequences x and y have the same initial value and the same sequence of recurrence relations. Therefore, they have all the same values: $x_k = y_k$ for all $k = 1, 2, \ldots$. □

We end our discussion of showing two sequences are equivalent by establishing an explicit formula for sequences defined recursively by a *linear* projection function,

$$x_n = \alpha x_{n-1} + c,$$

with $\alpha \neq 1$. When $\alpha = 1$ we have an arithmetic sequence, which is a sequence we already know. When $c = 0$, we have a geometric sequence. The projection function $f : x_{n-1} \mapsto x_n$ is defined by the formula $f(x) = \alpha x + c$. The fixed point x^* is the solution to

$$\alpha x + c = x \qquad \Leftrightarrow \qquad x^* = \frac{c}{1-\alpha},$$

defined only for $\alpha \neq 1$.

The linear projection function can be rewritten in terms of the fixed point using slope α and point (x^*, x^*) as

$$f(x) = x^* + \alpha(x - x^*).$$

This means that the recurrence relation can be written

$$x_n = x^* + \alpha(x_{n-1} - x^*) \quad \Leftrightarrow \quad x_n - x^* = \alpha(x_{n-1} - x^*).$$

Consequently, $x_n - x^*$ is a geometric sequence with ratio α. This allows us to find an explicit formula for x_n.

Theorem 4.3.7 Explicit Formula for Linear Recursive Sequences. *Suppose x_n is defined recursively by the equation*

$$x_n = \alpha x_{n-1} + c$$

with $\alpha \neq 1$. Then x_n is defined explicitly by a shifted geometric sequence

$$x_n = x^* + (x_0 - x^*) \cdot \alpha^n = x^* + (x_1 - x^*) \cdot \alpha^{n-1},$$

where $x^* = \dfrac{c}{1-\alpha}$ is the equilibrium of the sequence.

4.3.3 Summation Notation

In mathematics, the idea of adding terms from a sequence appears so frequently that a special notation, called summation notation or sigma notation for the Greek letter sigma Σ, was created to represent the sum.

Definition 4.3.8 Summation Notation. Given any sequence x and integers $m \leq n$, the sum of terms x_k with index k satisfying $m \leq k \leq n$ is written

$$\sum_{k=m}^{n} x_k = x_m + x_{m+1} + \cdots + x_n.$$

The starting index m is called the **lower limit** of the sum while the ending index n is called the **upper limit**. $\Diamond$

The sequence of terms being added is often given as an explicit function of the index. In that case, the explicit formula is used in place of the sequence name in the summation.

Example 4.3.9 Evaluate the following sums.

1. $\displaystyle\sum_{k=3}^{7}[2k+3]$
2. $\displaystyle\sum_{k=1}^{4}\frac{1}{k^2+k}$

Solution.

1. The sum $\displaystyle\sum_{k=3}^{7}[2k+3]$ involves the increment sequence $a_k = 2k+3$ and is the sum of terms with index from 3 to 7:

   $$a_3 = 2(3)+3 = 9,\ a_4 = 2(4)+3 = 11,\ a_5 = 13,\ a_6 = 15,\ a_7 = 17.$$

   Consequently, we can find the sum

   $$\sum_{k=3}^{7}[2k+3] = 9+11+13+15+17 = 65.$$

2. The sum $\displaystyle\sum_{k=1}^{4}\frac{1}{k^2+k}$ involves an increment sequence $b_k = \dfrac{1}{k^2+k}$. The index values involved go from 1 to 4 so that we find

   $$\begin{aligned}\sum_{k=1}^{4}\frac{1}{k^2+k} &= \frac{1}{1^1+1} + \frac{1}{2^2+2} + \frac{1}{3^2+3} + \frac{1}{4^2+4}\\ &= \frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \frac{1}{20}\\ &= \frac{30}{60} + \frac{10}{60} + \frac{5}{60} + \frac{3}{60}\\ &= \frac{48}{60} = \frac{4}{5}.\end{aligned}$$

$\square$

An accumulation sequence is closely related to summation. The accumula-

tion sequence is a new sequence formed by starting with an initial value and then adding one increment at a time. Suppose $x = (x_k)_{k=1}^{\infty}$ and u is the corresponding accumulation sequence with initial value u_0. We can write each term of u as the initial value plus a partial sum of the increments.

$$\begin{aligned} u_1 &= u_0 + x_1 = u_0 + \sum_{k=1}^{1} x_k \\ u_2 &= u_0 + x_1 + x_2 = u_0 + \sum_{k=1}^{2} x_k \\ u_3 &= u_0 + x_1 + x_2 + x_3 = u_0 + \sum_{k=1}^{3} x_k \\ &\vdots \end{aligned}$$

In general, we have

$$u_n = u_0 + \sum_{k=1}^{n} x_k.$$

Notice how the index for u appears as the upper limit of the summation and that the index of summation is a different variable. The index of summation can be any other unused variable, so that we might have instead written

$$u_n = u_0 + \sum_{i=1}^{n} x_i.$$

Also, notice that for consistency, we require

$$\sum_{k=1}^{0} x_k = 0,$$

regardless of the sequence x to indicate that no terms have been added in the summation. In general, we have the following representation.

Theorem 4.3.10 *If $x = (x_k)_{k=m}^{n}$ and u is the accumulation sequence with initial value u_{m-1}, then we can write*

$$u_k = u_{m-1} + \sum_{i=m}^{k} x_i,$$

for $i = m, \ldots, n$.

Example 4.3.11 Write the accumulation sequence $z = (z_n)_{n=0}^{\infty}$ with initial value $z_0 = 4$ and an increment sequence $a = (3, 5, 7, 9, 11, 13, \ldots)$ as a summation with an explicit formula for the increments.

Solution. The sequence z has initial value 4 which corresponds to index 0,

$$z_0 = 4.$$

For index values $n > 0$, the sequence is computed with an accumulation of values from the sequence a.

$$z_1 = 4 + \sum_{k=1}^{1} a_k = 4 + 3 = 7$$

$$z_2 = 4 + \sum_{k=1}^{2} a_k = 4 + 3 + 5 = 12$$
$$z_3 = 4 + \sum_{k=1}^{3} a_k = 4 + 3 + 5 + 7 = 19$$
$$z_4 = 4 + \sum_{k=1}^{4} a_k = 4 + 3 + 5 + 7 + 9 = 28$$

We need an explicit formula for the sequence $k \mapsto a_k$. We recognize that a is an arithmetic sequence with $a_1 = 3$ and constant increment $\nabla a_k = 2$. By Theorem 13.2.8, we know $a_k = a_1 + 2(k-1) = 2k + 1$. Using this explicit formula in the summation, we find

$$z_n = 4 + \sum_{k=1}^{n} (2k + 1).$$

□

Example 4.3.12 Show that $\sum_{k=1}^{n} (2k - 1) = n^2$ for $n = 0, 1, \ldots$.

Solution. There are two distinct sequences appearing in the equation— the sequence defined by the accumulation,

$$u_n = \sum_{k=1}^{n} (2k - 1),$$

and the sequence defined by an explicit formula. As u_n includes only a summation, we must have a zero initial value, $u_0 = 0$.

Because we know that $x_k = 2k - 1$ is the increment sequence for u, we only need to show that w has the same initial value and increment sequence. The initial value of w is

$$w_0 = 0^2 = 0,$$

in agreement with that of $u_0 = 0$. The increment is computed using the backward difference.

$$\begin{aligned} \nabla w_n &= w_n - w_{n-1} \\ &= n^2 - (n-1)^2 \\ &= n^2 - (n^2 - 2n + 1) \\ &= n^2 - n^2 + 2n - 1 \\ &= 2n - 1 \end{aligned}$$

The explicit formula for the increment of w is the same as that for u. Consequently, we know that $u_n = w_n$ for all $n = 0, 1, 2, \ldots$. □

4.3.4 Summary

- An accumulation sequence is a sequence generated from an initial value and a given sequence of increments.

- If x is the sequence of increments and u is the accumulation sequence,

then u satisfies the recurrence relation

$$u_n = u_{n-1} + x_n.$$

- If two sequences share the same initial value and the same increments, then the sequences are identical (Theorem 4.3.3). More generally, if two sequences share the same initial value and sequence of recurrence relations involving the previous term, then the sequences are identical (Theorem 4.3.5).
- Summation notation (sigma notation) provides a method to communicate the sequence of increments as well as the range of index values. The index variable is sometimes called the dummy variable because any other variable could be used in its place.
- Every accumulation sequence u can be represented as the initial value added to the summation of its increments x_k with the index variable appearing as the upper limit,

$$u_n = u_m + \sum_{k=m+1}^{n} x_k.$$

4.3.5 Exercises

Find the first six terms of the indicated accumulation sequence for the given increment sequence. Clearly indicate the relevant index values.

1. Find u with increments defined by $x = (x_k)_{k=2}^{\infty} = (-4, -2, 0, 2, 4, 6, \ldots)$ and initial value 21.

2. Find w with increments defined by $y = (y_i)_{i=0}^{\infty} = (1, -1, 1, 1, -1, -1, 1, 1, 1, -1, \ldots)$ and initial value 0.

In the following problems, show that explicit definition and recursive definition define the same sequence. If not, explain why.

3. $x_n = 3n - 5$ for $n = -2, -1, 0, \ldots$ defines the same sequence as $x_n = x_{n-1} + 3$ with $x_{-2} = -11$.

4. $x_n = 4 + \frac{3^{(n+1)}}{4^n}$ for $n = 0, 1, 2, \ldots$ defines the same sequence as $x_n = \frac{3}{4}x_{n-1} + 1$ with $x_0 = 7$.

5. $x_n = 2^{n+1} - 1$ for $n = 0, 1, 2, \ldots$ defines the same sequence as $x_n = x_{n-1} + 2^n$ with $x_0 = 1$.

6. $x_n = n^2 - n$ for $n = 0, 1, 2, \ldots$ defines the same sequence as $x_n = x_{n-1} + n$ with $x_0 = 0$.

7. $x_n = \frac{1}{2}(3^n - 1)$ for $n = 0, 1, 2, \ldots$ defines the same sequence as $x_{n+1} = x_n + 3^n$ with $x_0 = 0$. Note: The recursive formula uses a forward recurrence, so either compare forward differences or rewrite the recursive equation as a backward recurrence.

Determine the intervals of monotonicity and concavity for each sequence defined by the given increments.

8. $\nabla x_n = 4n - 70$ for $n = 1, 2, \ldots$ with $x_0 = 20$.

9. $\nabla x_n = 50 - 3n$ for $n = 1, 2, \ldots$ with $x_0 = -10$.

10. $\nabla x_n = n^2 - 30n$ for $n = 1, 2, \ldots$ with $x_0 = 0$.

11. $\nabla x_n = -100 + 75n - n^2$ for $n = 1, 2, \ldots$ with $x_0 = 0$.

For each of the following summations, write down the sum of individual terms. Then compute the value of the sum. For example, $\sum_{k=2}^{5} 2k$ would be $2(2) + 2(3) + 2(4) + 2(5) = 4 + 6 + 8 + 10 = 28$.

12. $\displaystyle\sum_{k=12}^{15} 3k$

13. $\displaystyle\sum_{k=-2}^{2} 2^k$

14. $\displaystyle\sum_{k=2}^{5} \frac{2k+1}{5k}$

Rewrite the following sums in summation notation. Find an appropriate formula for the increment sequence and identify the correct lower and upper limits of the sum.

15. $15 + 20 + 25 + 30 + \cdots + 90$

16. $21 + 25 + 29 + 33 + \cdots + 61$

17. $\frac{1}{4} + \frac{2}{9} + \frac{3}{16} + \cdots + \frac{12}{169}$

18. The sum of all four digit odd numbers.

19. The sum of all two-digit squares, $16 + 25 + 36 + 49 + 64 + 81$.

20. The sum of all three-digit odd squares.

Show that the summation formulas below are valid for $n = 0, 1, 2, \ldots$ by showing that two sequences are equal to one another.

21. $\displaystyle\sum_{k=1}^{n}(2k) = n^2 + n.$

22. $\displaystyle\sum_{k=1}^{n}(4k - 3) = n(2n - 1).$

23. $\displaystyle\sum_{k=1}^{n}(6k^2 - 2k) = 2n^2(n + 1).$

4.4 Summation Formulas

4.4.1 Overview

In the previous section, we learned that accumulation sequences could be written using summation notation. Consequently, summations can always be interpreted in the context of a sequence. We have seen some examples where we could show that an accumulation sequence representing a summation was equivalent to a sequence defined explicitly. Unfortunately, that process is only useful if we can somehow discover the explicit formula to compare. We seek for computational methods that will allow us to find the explicit values for summations.

This section studies the properties of summation and their application. We will learn that any summation can be interpreted as a net change in an accumulation sequence. We will also learn about algebraic properties of summation, particularly a property known as linearity. Once we know summation formulas for elementary building blocks, these properties will allow us to combine them for more complicated formulas.

4.4.2 Summation of as Net Accumulated Change

In the previous section, we learned that every accumulation sequence can be written using summation notation. The reverse is true. For every summation, we can define a corresponding accumulation sequence. Suppose we are interested in a summation

$$S = \sum_{k=m}^{n} x_k,$$

where $m > 0$ and $n \geq m$ and $x = (x_k)$ is the sequence whose terms are being added. Let u be *any* accumulation sequence with increments x and initial value u_0. Then we know we can write

$$u_k = u_0 + \sum_{i=1}^{k} x_i.$$

We want to find the relation between the summation S and the accumulation sequence.

First, we observe that the equation defining the accumulation sequence can be rewritten with the summation isolated:

$$\sum_{i=1}^{k} x_i = u_k - u_0.$$

The summation is equal to the difference between the initial value and the final value of the accumulation sequence. This observation can be generalized to other index ranges, but to explain it we first need the splitting property of summation.

Theorem 4.4.1 Summation Splitting Property. *For any summation* $\sum_{i=m}^{n} x_i$ *and intermediate index* k *with* $m < k < n$*, we can split the sum at* k *as*

$$\sum_{i=m}^{n} x_i = \sum_{i=m}^{k} x_i + \sum_{i=k+1}^{n} x_i.$$

Proof. The property is simply a generalization of the associative properties of addition. The basic idea is to group terms,

$$\begin{aligned}\sum_{i=m}^{n} x_i &= x_m + x_{m+1} + \cdots + x_k + x_{k+1} + \cdots + x_n \\ &= \big(x_m + x_{m+1} + \cdots + x_k\big) + \big(x_{k+1} + \cdots + x_n\big) \\ &= \sum_{i=m}^{k} x_i + \sum_{i=k+1}^{n} x_i\end{aligned}$$

■

This splitting property allows us to rewrite a summation as a change in an associated accumulation sequence.

Theorem 4.4.2 Summation as Net Accumulated Change. *Given any sequence of terms x_k and an accumulation sequence u with $\nabla u_k = x_k$,*

$$\sum_{k=m}^{n} x_k = u_n - u_{m-1}.$$

Proof. The accumulation sequence can be written

$$u_n = u_0 + \sum_{k=1}^{n} x_k.$$

By the Summation Splitting Property, if we split the sum at index $k = m-1$, we have

$$u_n = u_0 + \sum_{k=1}^{m-1} x_k + \sum_{k=m}^{n} x_k.$$

However, once we recognize

$$u_{m-1} = u_0 + \sum_{k=1}^{m-1} x_k,$$

we have

$$u_n = u_{m-1} + \sum_{k=m}^{n} x_k.$$

Solving for the summation gives the stated conclusion,

$$\sum_{k=m}^{n} x_k = u_n - u_{m-1}.$$

■

In this theorem, notice that which accumulation sequence is used does not matter. The initial value is irrelevant. In addition, notice that the accumulated change represented by the sum is equal to the change from *one before* the lower limit to the upper limit. This extra index step corresponds to increments matching backward differences. Finally, notice that the difference in accumulations can also be written

$$\sum_{k=m}^{n} x_k = \sum_{k=1}^{n} x_k - \sum_{k=1}^{m-1} x_k.$$

Example 4.4.3 In a previous example, we showed that the accumulation of odd integers was the sequence of squares. Use this to compute the sum of all odd three digit numbers.

Solution. The question is asking us to compute

$$101 + 103 + 105 + \cdots + 997 + 999.$$

In order to apply summation properties and accumulation sequences, we need the explicit formula for the sequence of terms as well as the lower limit and upper limit of the sum.

The sequence of odd integers $x = (1, 3, 5, \ldots)$ has an explicit formula $x_k = 2k - 1$, $k = 1, 2, 3, \ldots$. It was for this sequence that we had 4.3.12

$$u_n = \sum_{k=1}^{n} 2k - 1 = n^2.$$

We want to write the sum of odd three digit numbers in terms of the sequence of increments. Then we will be able to use the explicit formula of the accumulation sequence to compute the sum.

To find the limits of summation, we need to find the value of the index k such that $x_k = 101$ (lower limit) and $x_k = 999$ (upper limit). For the lower limit, we have

$$\begin{aligned} 2k - 1 &= 101 \\ 2k &= 102 \\ k &= 51, \end{aligned}$$

and for the upper limit we have

$$\begin{aligned} 2k - 1 &= 999 \\ 2k &= 1000 \\ k &= 500. \end{aligned}$$

Consequently, the sum of interest is

$$S = \sum_{k=51}^{500} 2k - 1 = 101 + 103 + 105 + \cdots + 997 + 999.$$

We are finally ready to apply the Summation as Net Accumulated Change. The summation is equal to the change in the accumulation sequence from 50 (index prior to first increment) to 500 (index of last increment),

$$\begin{aligned} S = \sum_{k=51}^{500} 2k - 1 &= u_{500} - u_{50} \\ &= 500^2 - 50^2 \\ &= 250000 - 2500 \\ &= 247500. \end{aligned}$$

□

4.4.3 Algebraic Properties of Summation

Now that we know that we can write a summation as the change of an accumulation sequence, we have a tool to compute summations once we are able to identify the accumulation. However, it can be tedious to find the accumulation sequence for every problem. We benefit from properties of summation that allow us to use elementary building blocks to compute the summation for a variety of different problems. These properties of summation correspond to the basic properties of addition.

Suppose we have a sequence x and a constant α. We can create a new sequence αx, called a **constant multiple**, by multiplying every term of x by the same constant α. Using the constant multiple as an increment sequence, every term will have a common factor of α. This leads to a property of summation called the constant multiple rule—constant multiples factor out of summation.

Theorem 4.4.4 Constant Multiple Rule of Summation. *Let x be a sequence and α a constant. Then for any lower and upper limits,*

$$\sum_{k=m}^{n} \alpha x_k = \alpha \sum_{k=m}^{n} x_k.$$

The next property considers a sequence that is itself formed by adding two sequences together. Suppose we have two sequences u and w and we form a new sequence $u + w$ with values that are the sum of the corresponding values, $(u + w)_n = u_n + w_n$. Because addition is both commutative and associative, any sum of a finite number of terms can be regrouped in any convenient way. A summation of terms $u + w$ can therefore be grouped in a way that we add only the terms from u and then add only the terms from v and then add the results. This leads to a property of summation called the sum rule.

Theorem 4.4.5 Sum Rule of Summation. *Let u and w be any two sequences defined for the range $k = m, \ldots, n$. Then*

$$\sum_{k=m}^{n} [u_k + w_k] = \sum_{k=m}^{n} u_k + \sum_{k=m}^{n} w_k.$$

Using the rules together creates a new rule called linearity involving two sequences x and y. The idea for this rule is that an individual term in the increment sequence is the sum of a constant multiple of each, $\alpha x + \beta y$. Such a sum is called a **linear combination** of x and y with coefficients α and β. This name results from the general equation of a line being of the form $ax + by = c$. Linearity applies the sum rule and the constant multiple as if in a single step.

Theorem 4.4.6 Linearity of Summation. *Let x and y be any two sequences with common domain and let α and β be any two constants. Then for any lower and upper limits,*

$$\sum_{k=m}^{n} [\alpha x_k + \beta y_k] = \alpha \sum_{k=m}^{n} x_k + \beta \sum_{k=m}^{n} y_k.$$

Using $\alpha = 1$ and $\beta = -1$, the linear combination becomes a difference, $\alpha x + \beta y = x - y$. So the difference rule is a special case of linearity.

Theorem 4.4.7 Difference Rule of Summation. *Let x and y be any two sequences with common domain. Then for any lower and upper limits,*

$$\sum_{k=m}^{n} [x_k - y_k] = \sum_{k=m}^{n} x_k - \sum_{k=m}^{n} y_k.$$

There are no corresponding rules for multiplication or division. This is really no different than emphasizing the importance of multiplying all terms using the distributive property, such as occurs with the FOIL method for multiplying binomials. For example, $\sum_{k=1}^{3}[k] = 1 + 2 + 3 = 6$. The product of the sum gives one result:

$$\sum_{k=1}^{3}[k] \cdot \sum_{k=1}^{3}[k] = (1 + 2 + 3) \cdot (1 + 2 + 3) = 6 \cdot 6 = 36.$$

But the sum of the products gives a different result:

$$\sum_{k=1}^{3}[k \cdot k] = (1^2 + 2^2 + 3^2) = 1 + 4 + 9 = 14.$$

In general,

$$\sum_{k=m}^{n}[x_k \cdot y_k] \neq \sum_{k=m}^{n} x_k \cdot \sum_{k=m}^{n} y_k.$$

4.4.4 Elementary Summation Formulas

There are some elementary increment sequences for which we can find an explicit formula for the accumulation sequence. We will state the results and prove them using the uniqueness criteria for accumulation sequences 4.3.3. The simplest accumulation sequence, and that used in each of the elementary summation formulas, use an initial value $s_0 = 0$. Thus, where we normally would have $s_n - s_0$ as the accumulated change, we only have s_n.

Theorem 4.4.8 Sum of Constant Sequence.

$$\sum_{k=1}^{n} c = cn$$

Proof. The accumulation sequence of interest is

$$u_n = \sum_{k=1}^{n} c.$$

The increment sequence x is a sequence of constants, $c_k = c$. The proposed explicit sequence is

$$w_n = cn.$$

The initial value of u is $u_0 = 0$ which matches the initial value of the explicit sequence $w_0 = c(0) = 0$. To show that $w = u$, we need to show that w has the same increments.

$$(\nabla w)_n = w_n - w_{n-1} = cn - c(n-1) = cn - cn + c = c$$

Since $(\nabla w)_n = c$ is the same increment as x_k, u and w are the same sequence. ■

Theorem 4.4.9 Sum of Natural Numbers.

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

Proof. The accumulation sequence of interest is

$$u_n = \sum_{k=1}^{n} k.$$

The increment sequence x is defined by $x_k = k$. The proposed explicit sequence is

$$w_n = \frac{n(n+1)}{2}.$$

The initial values agree:

$$u_0 = 0,$$
$$w_0 = \frac{0(1)}{2} = 0.$$

The increment for w is given by:

$$\begin{aligned}(\nabla w)_n = w_n - w_{n-1} &= \frac{n(n+1)}{2} - \frac{(n-1)n}{2} \\ &= \frac{n}{2}\big((n+1) - (n-1)\big) = \frac{n}{2} \cdot 2 = n\end{aligned}$$

Since $(\nabla w)_n = n = x_n$, u and w have the same increments and same initial value. By Theorem 4.3.3, u and w are equivalent. ■

Theorem 4.4.10 Sum of Squares.

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

Proof. The accumulation sequence of interest is

$$u_n = \sum_{k=1}^{n} k^2$$

so that increment sequence x is defined by $x_k = k^2$. The proposed explicit sequence is

$$w_n = \frac{n(n+1)(2n+1)}{6}.$$

The initial values agree:

$$u_0 = 0,$$
$$w_0 = \frac{0(1)(1)}{6} = 0.$$

The increment for w is given by:

$$\begin{aligned}(\nabla w)_n &= w_n - w_{n-1} \\ &= \frac{n(n+1)(2n+1)}{6} - \frac{(n-1)n\big(2(n-1)+1\big)}{6} \\ &= \frac{n}{6}\big((n+1)(2n+1) - (n-1)(2n-1)\big) \\ &= \frac{n}{6}\big((2n^2+3n+1) - (2n^2-3n+1)\big) \\ &= \frac{n}{6}(6n) = n^2\end{aligned}$$

Since $(\nabla w)_n = n^2 = x_n$, u and w have the same increments and same initial value. By Theorem 4.3.3, u and w are equivalent. ■

Theorem 4.4.11 Sum of Cubes.

$$\sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4}$$

Proof. The accumulation sequence of interest is

$$u_n = \sum_{k=1}^{n} k^3$$

so that increment sequence x is defined by $x_k = k^3$. The proposed explicit sequence is

$$w_n = \frac{n^2(n+1)^2}{4}.$$

The initial values agree:

$$u_0 = 0,$$
$$w_0 = \frac{0(1)(1)}{6} = 0.$$

The increment for w is given by:

$$\begin{aligned}(\nabla w)_n &= w_n - w_{n-1} \\ &= \frac{n^2(n+1)^2}{4} - \frac{(n-1)^2 n^2}{4} \\ &= \frac{n^2}{4}\big((n+1)^2 - (n-1)^2\big) \\ &= \frac{n^2}{4}\big((n^2+2n+1) - (n^2-2n+1)\big) \\ &= \frac{n^2}{4}(4n) = n^3\end{aligned}$$

Since $(\nabla w)_n = n^3 = x_n$, u and w have the same increments and same initial value. By Theorem 4.3.3, u and w are equivalent. ■

Theorem 4.4.12 Sum of a Geometric Sequence.

$$\sum_{k=0}^{n} b^k = \frac{b^{n+1} - 1}{b - 1}$$

Proof. The accumulation sequence of interest is

$$u_n = \sum_{k=0}^{n} b^k$$

so that increment sequence x is defined by $x_k = b^k$. The proposed explicit sequence is

$$w_n = \frac{b^{n+1} - 1}{b - 1}.$$

Because the summation lower index is 0, the sequence u has a non-zero initial value $u_0 = b^0 = 1$. The initial value of w is given by

$$w_0 = \frac{b^1 - 1}{b - 1} = 1,$$

which matches the initial value of u. The increment for w is given by:

$$\begin{aligned}(\nabla w)_n &= w_n - w_{n-1} = \frac{b^{n+1}-1}{b-1} - \frac{b^n-1}{b-1} \\ &= \frac{1}{b-1}\big((b^{n+1}-1)-(b^n-1)\big) \\ &= \frac{b^{n+1}-b^n}{b-1} = \frac{b^n(b-1)}{b-1} = b^n\end{aligned}$$

Since $(\nabla w)_n = b^n = x_n$, u and w have the same increments and same initial value. By Theorem 4.3.3, u and w are equivalent. ■

Our first examples consider sums involving just the elementary terms.

Example 4.4.13 Find the sum of the integers $1, 2, \ldots, 100$.

Solution. Start by recognizing this as the accumulation of the sequence $x = (k : k = 1, 2, 3, \ldots)$ over a range $1 \le k \le 100$. This allows us to rewrite our problem as a summation:

$$\sum_{k=1}^{100} k.$$

Theorem Theorem 4.4.9 applies directly with $n = 100$, so we know

$$\sum_{k=1}^{100} k = \frac{100(101)}{2} = 5050.$$

□

Example 4.4.14 Find the sum of the integers $100, 101, \ldots, 200$.

Solution. This example uses the same basic sequence (the integers) but instead of starting at $k = 1$, we are summing the sequence $x = (k : k = 1, 2, 3, \ldots)$ over an index range $100 \le k \le 200$,

$$100 + 101 + \cdots + 200 = \sum_{k=100}^{200} k.$$

Using the Summation as Net Accumulated Change theorem, we can write the summation as a difference

$$\sum_{k=100}^{200} k = \sum_{k=1}^{200} k - \sum_{k=1}^{99}.$$

The two summations are accumulations from Theorem 4.4.9:

$$\sum_{k=1}^{200} k = \frac{200(201)}{2} = 20100,$$

$$\sum_{k=1}^{99} k = \frac{99(100)}{2} = 4950.$$

Consequently,

$$\sum_{k=100}^{200} k = 20100 - 4950 = 15150.$$

□

4.4.5 Summations of Linear Combinations

The elementary summation formulas allow us to compute sums involving only the elementary terms. Combining these formulas using the properties of summation, namely using the constant multiple rule and the sum rule, we can compute sums of any linear combination of the elementary terms.

Example 4.4.15 Find $\sum_{k=1}^{20}(500 + 60k - 2k^2)$.

Solution. The increments in the sum consist of a constant (500), a constant multiple of the index ($60k$), and a constant multiple of the square of the index ($-2k^2$). The linearity property of summation 4.4.6 allows us to compute the sum using the elementary formulas. Although linearity allows the two steps to be done at once, the following illustrates the steps (sum and constant multiple rules) in order:

$$\begin{aligned}\sum_{k=1}^{20}[500 + 60k - 2k^2] &= \sum_{k=1}^{20}[500] + \sum_{k=1}^{20}[60k] + \sum_{k=1}^{20}[-2k^2] \\ &= \sum_{k=1}^{20}[500] + 60\sum_{k=1}^{20}[k] - 2\sum_{k=1}^{20}[k^2].\end{aligned}$$

The brackets emphasize that the increments of a summation are given by a particular value or formula. Each of these summations involve elementary increment sequences for which we have explicit formulas.

$$\begin{aligned}\sum_{k=1}^{20}[500] &= 500(20) = 10000, \\ \sum_{k=1}^{20}[k] &= \frac{20(21)}{2} = 210, \\ \sum_{k=1}^{20}[k^2] &= \frac{20(21)(41)}{6} = 2870.\end{aligned}$$

Consequently,

$$\sum_{k=1}^{20}[500 + 60k - 2k^2] = 10000 + 60(210) - 2(2870) = 16860.$$

□

The same strategy still applies if the constant multiple coefficients are written using parameters or even using variables other than the dummy index variable of summation. In particular, when the upper limit of the summation is a variable, the formula for the sequence might also involve that variable as well as the index variable. Because this will be encountered frequently, an example is provided below.

Example 4.4.16 Find a formula for $\sum_{k=1}^{n}\left(\frac{3k}{n^2} - \frac{k^2}{n^3}\right)$ that involves only n.

Solution. The sequence of increments is $x_k = \frac{3k}{n^2} - \frac{k^2}{n^3}$. We recognize this as a linear combination of the more elementary sequences k and k^2 if we rewrite

the sequence

$$x_k = \frac{3}{n^2} \cdot k + \frac{-1}{n^3} \cdot k^2.$$

Because the coefficients of this linear combination only involve n and not the dummy variable of the summation k, we can rewrite the summation as a corresponding linear combination and then apply the elementary summation formulas to find our desired formula.

$$\begin{aligned}\sum_{k=1}^{n}\left(\frac{3k}{n^2} - \frac{k^2}{n^3}\right) &= \frac{3}{n^2}\sum_{k=1}^{n}[k] - \frac{1}{n^3}\sum_{k=1}^{n}[k^2] \\ &= \frac{3}{n^2} \cdot \frac{n(n+1)}{2} - \frac{1}{n^3} \cdot \frac{n(n+1)(2n+1)}{6} \\ &= \frac{3n(n+1)}{2n^2} - \frac{n(n+1)(2n+1)}{6n^3}.\end{aligned}$$

□

There are no convenient summation rules for products or quotients, with one exception. If the product can be rewritten as a sum using the distributive property of multiplication, then we can sometimes use linearity after this simplification in terms of elementary formulas. If the increments are not linear combinations of elementary terms, then we have no methods for simplifying the calculation.

Example 4.4.17 Find a formula for $\sum_{k=1}^{n}(2 - \frac{3k}{n})(1 + \frac{2k}{n})$ that only involves n.

Solution. Use the distributive property (aka FOIL) to rewrite the product as a sum which can be identified as a linear combination of a constant term, k, and k^2:

$$\begin{aligned}(2 - \frac{3k}{n})(1 + \frac{2k}{n}) &= 2 + \frac{4k}{n} - \frac{3k}{n} - \frac{12k^2}{n^2} \\ &= 2 + \frac{1}{n} \cdot k - \frac{12}{n^2}k^2.\end{aligned}$$

The linearity property of summation 4.4.6 allows us to compute the sum as the same linear combination of the elementary accumulations:

$$\begin{aligned}\sum_{k=1}^{n}(2 - \frac{3k}{n})(1 + \frac{2k}{n}) &= \sum_{k=1}^{n} 2 + \frac{1}{n} \cdot k - \frac{12}{n^2}k^2 \\ &= \sum_{k=1}^{n} 2 + \frac{1}{n}\sum_{k=1}^{n} \cdot k - \frac{12}{n^2}\sum_{k=1}^{n} k^2 \\ &= 2n + \frac{1}{n} \cdot \frac{n(n+1)}{2} - \frac{12}{n^2}\frac{n(n+1)(2n+1)}{6}\end{aligned}$$

To simplify the answer, we need to cancel common factors and then rewrite the expression with a common denominator.

$$\begin{aligned}\sum_{k=1}^{n}(2 - \frac{3k}{n})(1 + \frac{2k}{n}) &= 2n + \frac{n+1}{2} - \frac{2(n+1)(2n+1)}{n} \\ &= \frac{(2n)(2n) + n(n+1) - 4(n+1)(2n+1)}{2n} \\ &= \frac{4n^2 + n^2 + n - 8n^2 - 12n - 4}{2n}\end{aligned}$$

$$= \frac{-3n^2 - 11n - 4}{2n}$$

□

4.4.6 Summary

- Summation of terms is equivalent to an accumulation of those terms as increments.
- The Summation Splitting Property allows us to split a summation over an index range into the sum of two summations over adjacent ranges.
- Every summation can be computed as the accumulated change of the terms as increments. The Summation as Net Accumulated Change theorem states that if we know the accumulation sequence u with increments x, then
$$\sum_{k=m}^{n} x_k = u_n - u_{m-1}.$$
- The linearity properties of summation (the constant multiple rule and the sum rule) allow us to break summations involving sums into simpler summations over the same index range.
  - Constant Multiple Rule of Summation: $\sum_{k=m}^{n} \alpha x_k = \alpha \sum_{k=m}^{n} x_k$
  - Sum Rule of Summation: $\sum_{k=m}^{n} [u_k + w_k] = \sum_{k=m}^{n} u_k + \sum_{k=m}^{n} w_k$
  - Linearity of Summation: $\sum_{k=m}^{n} [\alpha u_k + \beta w_k] = \alpha \sum_{k=m}^{n} u_k + \beta \sum_{k=m}^{n} w_k$
- Elementary accumulation formulas:
  - Sum of Constant Sequence: $\sum_{k=1}^{n} c = cn$
  - Sum of Natural Numbers: $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$
  - Sum of Squares: $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$
  - Sum of Cubes: $\sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4}$
  - Sum of a Geometric Sequence: $\sum_{k=0}^{n} b^k = \frac{b^{n+1} - 1}{b - 1}$

4.4.7 Exercises

The following collection of problems practice applying the properties of summation. Given the following information about the sequence x and y, compute the desired summations.

$$x_0 = 8 \qquad \sum_{k=1}^{10} x_k = 42 \qquad \sum_{k=0}^{20} x_k = 30$$

$$y_{19} = 4 \qquad y_{20} = -8 \qquad \sum_{k=0}^{18} y_k = -4 \qquad \sum_{k=11}^{20} y_k = 5$$

1. $\displaystyle\sum_{k=0}^{10} 4x_k$

2. $\displaystyle\sum_{k=0}^{20} x_k + y_k$

3. $\displaystyle\sum_{k=11}^{20} 3x_k - 2y_k$

Compute the following sums using the summation properties and the elementary summation formulas.

4. $\displaystyle\sum_{k=1}^{20} 3k$

5. $\displaystyle\sum_{k=1}^{30} 4k - 100$

6. $\displaystyle\sum_{k=12}^{20} k^2$

7. $\displaystyle\sum_{k=1}^{n} (1+3k)(2-5k)$

8. $\displaystyle\sum_{k=0}^{6} 3^k$

9. $3 + 3^3 + 3^5 + 3^7 + \cdot + 3^{19}$

Hint: Rewrite as a summation that can use the geometric sum.

10. The sum of three-digit multiples of 5

11. The sum of three-digit perfect squares

12. $\displaystyle\sum_{k=1}^{n} \left(\frac{4}{n} - \frac{5k}{n^2}\right)$

13. $\displaystyle\sum_{k=1}^{n} \frac{3k^2}{n^3}$

14. $\displaystyle\sum_{k=1}^{n} \left(2 + \frac{3k}{n}\right)^2 \cdot \frac{1}{n}$

4.5 Limits of Sequences

4.5.1 Overview

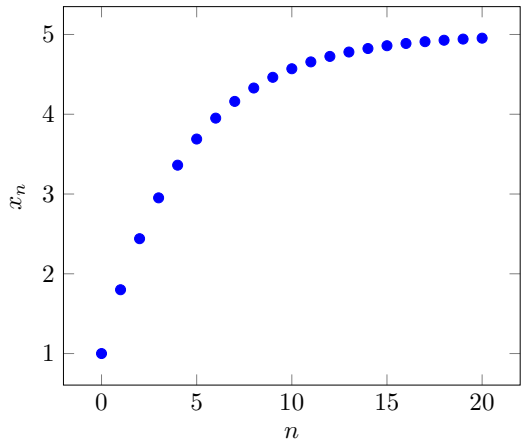
The limit of a sequence describes its end behavior. Some sequences converge to a particular value, meaning that the values of the sequence keep getting closer and closer to that value. Other sequences grow without bound. Still other sequences alternate between values or behave chaotically. Limits allow us to establish some mathematical language that characterizes some of these behaviors.

The objectives for this section are as follows. We will focus on what it means for a sequence to have a limit. Much of our intuition will focus on graphs and tables. As we do this, we will learn about limits of sequences defined recursively in terms of fixed points of the projection function. We will then discuss limit arithmetic involving infinity and how this can be used to find limits of some sequences defined explicitly.

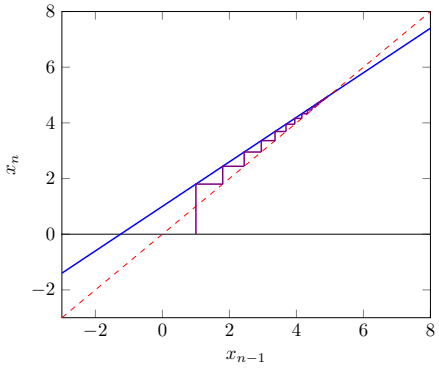
4.5.2 What Is A Limit?

Consider the sequence defined recursively by $x_n = 0.8x_{n-1}+1$ and initial value $x_0 = 1$. The next ten values (to the nearest ten-thousandth) are shown in a table and the first twenty values are illustrated in a graph.

n	x_n
0	1
1	1.8
2	2.44
3	2.952
4	3.3616
5	3.68928
6	3.951424
7	4.161139
8	4.328911
9	4.463129
10	4.570503



The graph illustrates that the sequence x is increasing and concave down. It appears that the values of the sequence might be leveling off at some value. A cobweb diagram, shown below, suggests that this sequence has values that are converging to a fixed point of the projection function $f(x) = 0.8x + 1$, where $f(x) = x$.



The fixed point corresponds to an equilibrium of the recursively defined sequence. We find the equilibrium value by solving the fixed point equation.

$$\begin{aligned} f(x) &= x \\ 0.8x + 1 &= x \\ 1 &= 0.2x \\ x &= 5 \end{aligned}$$

Now, let us compare the decimal approximations to the sequence with higher index values with the equilibrium.

n	x_n
20	4.953883140
25	4.984888427
30	4.995048240
35	4.998377407
40	4.999468309
45	4.999825775
50	4.999942910
55	4.999981293
60	4.999993870

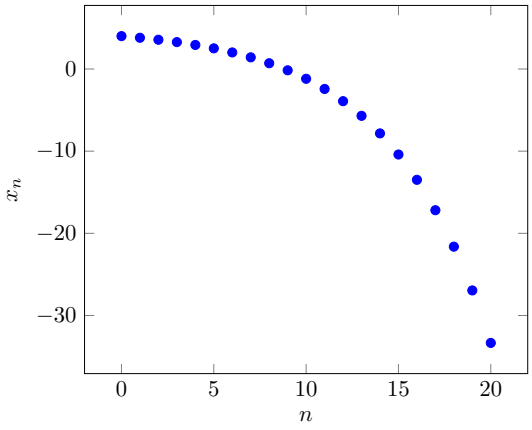
Notice that the values for the sequence have decimal approximations that are converging to the equilibrium value $x = 5$. The greater the index value, the closer the sequence value is to equilibrium. Consequently, we say that our sequence has a limit of 5 and write $x_n \to 5$ or

$$\lim_{n\to\infty} x_n = 5.$$

Next consider another example—a recursive sequence x defined by a recurrence relation $x_n = 1.2x_n - 1$ and initial value $x_0 = 4$. The projection function $f(x) = 1.2x - 1$ has the same fixed point $x = 5$ because

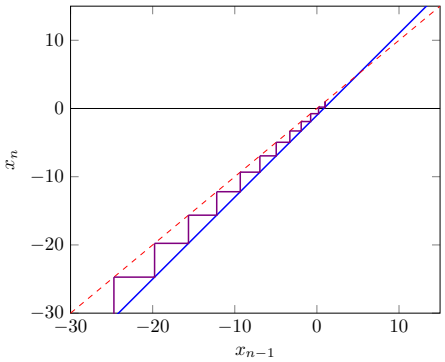
$$f(5) = 1.2(5) - 1 = 5.$$

n	x_n
0	4
1	3.8
2	3.56
3	3.272
4	2.9264
5	2.51168
6	2.014016
7	1.416819
8	0.700183
9	-0.1597804
10	-1.191736



However, this time the sequence is decreasing and concave down, moving away from the equilibrium value. Instead, the value of the sequence is becoming more and more negative. The cobweb diagram illustrates that this will continue forever. For a sequence like this, we say $x_n \to -\infty$ or

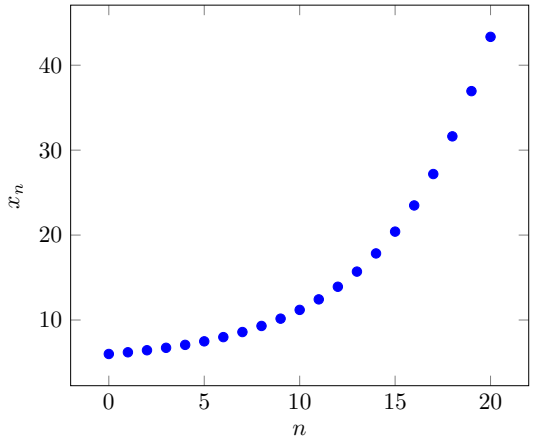
$$\lim_{n\to\infty} x_n = -\infty.$$



For the same recursive definition $x_n = 1.2x_{n-1} - 1$ with an initial value above the equilibrium, $x_0 = 6$, the sequence is increasing and concave up, shown below. The values become more and more positive. For a sequence like this, we say $x_n \to \infty$ or

$$\lim_{n\to\infty} x_n = \infty.$$

n	x_n
0	6
1	6.2
2	6.44
3	6.728
4	7.0736
5	7.48832
6	7.985984
7	8.583181
8	9.299817
9	10.15978
10	11.19174



We have seen that a recursive sequence sometimes converges to a fixed point and sometimes it diverges away from a fixed point. We might wonder if a recursive sequence can have a limit that is *not* a fixed point of the projection function. The answer is no, so long as the projection function is continuous.

Theorem 4.5.1 Recursive Limits as Fixed Points. *If a sequence is defined recursively with a continuous projection function f, $x_n = f(x_{n-1})$, and x_n has a limit, then the limit must be a fixed point of f.*

Proof. We have to wait for a definition of continuity before this can be proved. ■

Note some things that this theorem does not guarantee. First, just because a function has a fixed point does not mean that it will be a limit. (The sequence might not have a limit.) Second, the projection function needs to be continuous. If a function is not continuous, then it is possible to have a limit that is not a fixed point. Fortunately, functions defined by simple algebraic formulas will be continuous everywhere they are defined. For any continuous projection functions, the only limits *will* be fixed points. Finally, if the sequence is not defined recursively, then we will need other methods to find the limits.

In practice, we first observe that a sequence defined recursively has a limit (perhaps through a graph or a table). If we find all of the fixed points for the projection function, then we can determine which of those is the appropriate limit. The limit may depend on the initial value of the sequence, so we compare approximate values from the table with the approximate (decimal) values of

the fixed points.

We consider some additional examples. As you attempt the examples or exercises, take advantage of technology to generate tables and graphs. Refer back to Section 13.3 for guidance as needed.

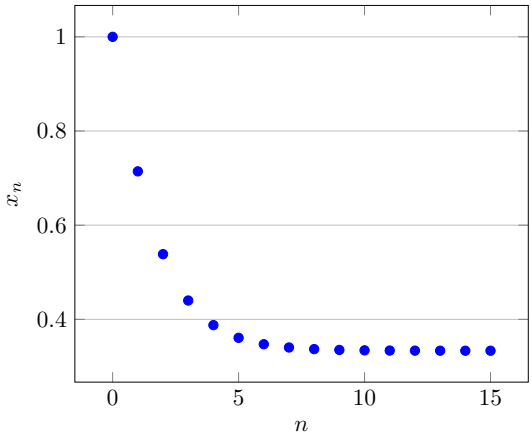
Example 4.5.2 Use a table and a graph of the sequence defined explicitly as

$$x_n = \frac{3 + 2^n}{1 + 3 \cdot 2^n}$$

to estimate the limit of x_n.

Solution. The explicit definition of the sequence allows us to create a table. Because we are looking for the decimal approximation to converge, we need to show quite a few decimal places. Once the table is generated, we can create a plot.

n	x_n
0	1.0
1	0.714285714286
2	0.538461538462
3	0.44
4	0.387755102041
5	0.360824742268
6	0.347150259067
7	0.34025974026
8	0.336801040312
9	0.335068314899
10	0.334201106411
11	0.33376729048
12	0.333550329563
13	0.333441835863
14	0.333387585702
15	0.333360459793



The plot for the sequence shows that the sequence appears to be leveling off. Looking at the table, we see that more and more of the digits are converging to a 3. If this pattern is real and continues, the limiting value would be the repeating decimal $0.3333\ldots$ which is the rational number $\frac{1}{3}$. We would say that this sequence has a limit $x_n \to \frac{1}{3}$:

$$\lim_{n\to\infty} x_n = \frac{1}{3}.$$

□

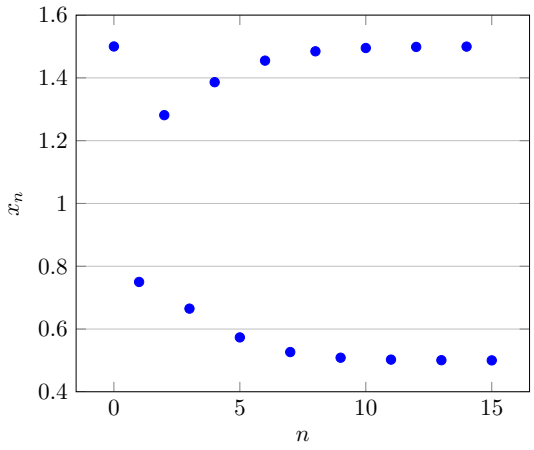
Example 4.5.3 Use a table and a graph of the sequence defined explicitly as

$$u_n = 1 + \frac{1}{2} \cdot (-1 + \frac{1}{2^n})^n,$$

to estimate the limit of u_n.

Solution. We generate a table of sequence values and plot the results.

n	u_n
0	1.5
1	0.75
2	1.28125
3	0.6650390625
4	1.38623809814
5	0.573392406106
6	1.45491835196
7	0.526711160622
8	1.48458696224
9	0.508720709959
10	1.49513858939
11	0.502678999959
12	1.4985371216
13	0.500792876146
14	1.49957292337
15	0.500228832948



Notice that for this sequence, the graph appears to level off to two different values. However, this is because the sequence is actually approaching two alternating values. This is an example of approaching a repeating pattern rather than a value, most easily seen in the table. As we look further down the table, the odd index values correspond to a sequence value that is getting closer to 0.5 while the even index values correspond to a sequence value that is getting closer to 1.5.

This sequence does not have a limit because the sequence is not approaching a single value. We say $\lim_{n\to\infty} x_n$ does not exist. □

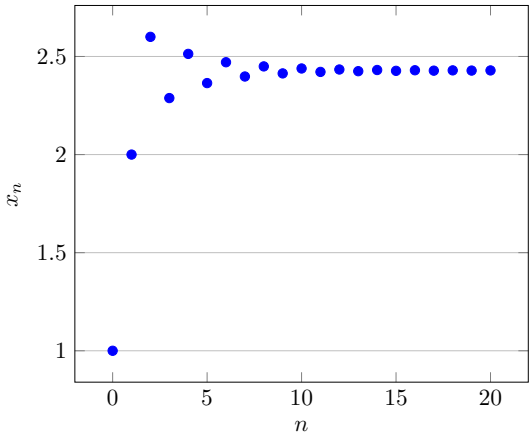
Example 4.5.4 Use a table and a graph of the sequence defined recursively by

$$z_{n+1} = 2.7z_n - 0.7z_n^2$$

and an initial value $z_0 = 1$ to estimate the limit of z_n.

Solution. We can use the recursive definition and a computer to generate approximate values and then graph the sequence.

n	z_n
0	1.00000000000000
1	2.00000000000000
2	2.60000000000000
3	2.28800000000000
4	2.51313920000000
5	2.36436779299635
6	2.47062849869924
7	2.39789332147854
8	2.44938730115809
9	2.41369700737469
10	2.43882864952499
11	2.42131772649675
12	2.43361218868805
13	2.42502511000601
14	2.43104504810447
15	2.42683561174279
16	2.42978439120944
17	2.42772132482997
18	2.42916599531699
19	2.42815498439281
20	2.42886281809844



Graphically, you should see that the plot shows the sequence values appears to level off to a single value. However, the graph also shows that the values are alternately above and below that limit. We look in the table to determine the limiting value, but it is not obvious from the table what the limiting value should be. Because the values are alternately above and then below whatever the limit should be, we can conclude that the limit must be between the last two values listed.

The limit is approximately 2.428 but we do not know the next decimal place without computing more values in the sequence. With additional computation (not shown), we find $z_{39} = 2.42857109636261$ and $z_{40} = 2.42857166111753$ so that our approximation for the limit (a value between z_{39} and z_{40}) can be estimated as close to 2.428571,

$$\lim_{n\to\infty} z_n \approx 2.428571.$$

To find a better approximation using data would require more computation.

Because the sequence is defined recursively, the limit will be a fixed point of the projection function. If we solve the fixed point equation, we can find the exact value of the limit.

$$\begin{aligned} 2.7x - 0.7x^2 &= x \\ 1.7x - 0.7x^2 &= 0 \\ x(1.7 - 0.7x) &= 0 \end{aligned}$$

The factored equation indicates there are two fixed points: $x = 0$ and $x = \frac{17}{7}$. The decimal approximation for $x = \frac{17}{7}$ is 2.4285714286, which is precisely where our sequence is converging,

$$\lim_{n\to\infty} z_n = \frac{17}{7}.$$

□

4.5.3 The Definition of a Limit

With these examples, we can introduce the mathematical definition for the limit of a sequence. The idea of a limit $x_n \to L$ is that the value of the sequence x_n approximates the value of L closer and closer as n increases. Recall that we measure the quality of approximation using the error of approximation, $|x_n - L|$. The sequence successfully approximates the limit L if the error eventually become smaller than any desired accuracy of approximation.

Definition 4.5.5 Limit of a Sequence. For a real number L, a sequence x has a limit L if for any desired accuracy of approximation $\epsilon > 0$, the error of approximation is eventually $|x_n - L| < \epsilon$ and we write

$$\lim_{n\to\infty} x_n = L \quad \text{or} \quad x_n \to L.$$

This means that for every $\epsilon > 0$, there is a threshold index N so that $|x_n - L| < \epsilon$ whenever $n > N$. $\Diamond$

In most cases, verifying the definition directly will be too challenging. We will usually learn to apply rules that guarantee that this definition will be satisfied rather than directly show that the approximation rule is satisfied. However, for sequences defined by sufficiently simple explicit formulas, we can determine how far down the table we would need to go to reach a desired accuracy. In these cases, we use strategies for solving inequalities to find the interval of integers where the error of approximation is sufficiently small.

Example 4.5.6 Consider the sequence $x_n = \dfrac{n}{2n+1}$, for $n = 0, 1, 2, \ldots$. Find numerical evidence that $\lim_{n\to\infty} x_n = \dfrac{1}{2}$. Then determine the index threshold N so that $|x_n - \frac{1}{2}| < 0.001$ for index values $n > N$.

Solution. Using the following simple Sage script, we can generate a plot and table quickly. This script includes some modifications so that only a portion of the table is printed. We also have modified the format code to include more decimal values.

```
# What range of index values to plot
lastIndex = 100
# Which values in the calculation to show?
# Some at the beginning
showFirst = 10
# Some at the end
showLast = 10

# Calculate the table for plotting
data = []
for n in range(lastIndex+1):
    xn = n/(2*n+1)
    data.append( [n,xn] )
graph = list_plot(data, frame=True, axes_labels=["n", "x_n"])
show(graph)

# Display beginning of the table
for i in range(showFirst):
    # data is a list of points, so data[i] = [n,xn]
    [n,xn] = data[i]
    print("%d\t%,8f" % (n,xn))
# Print a break-line to show there are missing terms
print("...")
# Display end of the table
for i in range(lastIndex-showLast,lastIndex+1):
    # data is a list of points, so data[i] = [n,xn]
    [n,xn] = data[i]
    print("%d\t%.8f" % (n,xn))
```

```
0       0.00000000
1       0.33333333
2       0.40000000
3       0.42857143
4       0.44444444
5       0.45454545
6       0.46153846
7       0.46666667
8       0.47058824
9       0.47368421
...
90      0.49723757
91      0.49726776
92      0.49729730
93      0.49732620
94      0.49735450
95      0.49738220
96      0.49740933
97      0.49743590
98      0.49746193
99      0.49748744
100     0.49751244
```

It appears that the sequence is leveling out to a value near 0.5. However, at index $n = 100$, the value is at approximately $x_{100} \approx 0.49751244$ (8 decimal places). If the limit really is $x_n \to 0.5$, then the error of approximation for $n = 100$ is

$$|x_{100} - 0.5| \approx 0.00248756.$$

We need to ensure that the error continues to go down. At index $n = 1000$ (by

modifying the script and running again), we have $x_{1000} \approx 0.49975012$ with an error of approximation

$$|x_{1000} - 0.5| \approx 0.00024988.$$

The explicit formula for the sequence uses the function $f(n) = \dfrac{n}{2n+1}$ for integer values of n. We can think of the mapping $n \overset{f}{\mapsto} x$ for arbitrary values of n and not just integers. The inverse $f^{-1} : x \mapsto n$ can inform us of when the sequence passes different values. We solve for n as the dependent variable.

$$\begin{gathered} x = \frac{n}{2n+1} \\ (2n+1)x = n \\ 2nx + x = n \\ 2nx - n = -x \\ (2x-1)n = -x \\ n = \frac{-x}{2x-1} = \frac{x}{1-2x} \end{gathered}$$

This function, $f^{-1}(x) = \dfrac{x}{1-2x}$, gives the relation $x \mapsto n$.

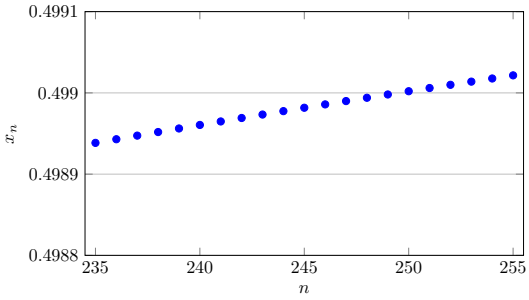
From our graph, we have seen that (n, x_n) is increasing and concave down. The sequence will always be below the limit. Given *any* desired error of approximation $\epsilon > 0$, we can use our inverse function to see exactly when the sequence rises above $x_n > \frac{1}{2} - \epsilon$. For example, with $\epsilon = 0.001$, we want to find when $x_n > 0.5 - 0.001 = 0.499$. We find

$$f^{-1}(0.499) = \frac{0.499}{1 - 2(0.499)} = 249.5.$$

The sequence, which requires integer index values, will therefore be above $x_n > 0.499$ for $n \geq 250$. In the definition of a limit, this corresponds to an index threshold $N = 249$ for accuracy $\epsilon = 0.001$.

We can verify this numerically by running the script to show a table and graph for index values surrounding $n = 250$. The table and graph show that the sequence value is below 0.499 for $n \leq 249$ and above 0.499 for $n \geq 250$, agreeing with our calculation.

n	x_n
235	0.49893843
236	0.49894292
237	0.49894737
238	0.49895178
239	0.49895616
240	0.49896050
241	0.49896480
242	0.49896907
243	0.49897331
244	0.49897751
245	0.49898167
246	0.49898580
247	0.49898990
248	0.49899396
249	0.49899800
250	0.49900200
251	0.49900596
252	0.49900990
253	0.49901381
254	0.49901768
255	0.49902153



The inverse function, $f^{-1}(x) = \dfrac{x}{1-2x}$, is undefined for $x = \frac{1}{2}$. For values $x > \frac{1}{2}$, we will have negative values for n, $f^{-1}(x) < 0$. The sequence therefore will never reach or surpass the value $x = \frac{1}{2}$. But we will be able to identify when the function rises above every value below $x = \frac{1}{2}$. This is how we *know* that

$$\lim_{n\to\infty} x_n = \frac{1}{2}.$$

□

4.5.4 Summary

- A sequence has a limit, $x_n \to L$, if the values of the sequence get closer and closer to the value of L. The graph of the sequence (n, x_n) approaches a horizontal line $x = L$.

- If $x_n \to L$, then a table of values for the sequence x should have decimal approximations that converge to the decimal approximation of L.

- Sequences defined recursively using a continuous projection function f can only have limits that are fixed points of f. (See Theorem 4.5.1)

- The formal definition of a sequence limit $x_n \to L$, written

  $$\lim_{n\to\infty} x_n = L,$$

  is that for any desired threshold of approximation error $\epsilon > 0$, there is an index threshold N so that $|x_n - L| < \epsilon$ whenever $n > N$. (See Definition 4.5.5)

4.5.5 Exercises

Use a computer-generated table to approximate the limit of the following sequences to at least four decimal places. If the limit does not exists, state why.

1. $x_n = \sqrt{n} \cdot (\sqrt{3n+1} - \sqrt{3n})$

2. $x_n = n \cdot (\sqrt{3n+1} - \sqrt{3n})$

3. $x_n = \left(1 + \dfrac{1}{3n}\right)^n$

4. $x_n = n^2 \cdot \left(\dfrac{1}{4n+5} - \dfrac{1}{4n}\right)$

5. $x_{n+1} = \dfrac{x_n}{5} + \dfrac{2}{x_n}$ with $x_0 = 2$.

6. $x_{n+1} = \dfrac{x_n}{5} - \dfrac{2}{x_n}$ with $x_0 = 2$.

7. $x_{n+1} = 5x_n e^{-x_n}$ with $x_0 = 0.5$.

8. $x_{n+1} = 8x_n e^{-x_n/3}$ with $x_0 = 0.5$.

For each of the following functions, find all fixed points. Then, using a sequence defined recursively $x_n = f(x_{n-1})$ and the given initial value, test if the sequence has a limit and give its exact value.

9. $f(x) = 1.3x + 12$; $x_0 = 1$

10. $f(x) = 0.8x + 6$; $x_0 = 5$

11. $f(x) = 1.2x - 0.04x^2$; $x_0 = 3$

12. $f(x) = \dfrac{4x}{1+x^2}$; $x_0 = 3$

For each of the following sequences, explore the definition of a limit by finding the index threshold associated with the approximation error threshold.

13. The sequence $x_n = \frac{1}{2n+1}$ has a limit $x_n \to 0$. For $\epsilon = 0.01$, find N so that $|x_n - 0| < \epsilon$ for $n > N$.

14. The sequence $x_n = \frac{n}{2n+1}$ has a limit $x_n \to \frac{1}{2}$. For $\epsilon = 0.01$, find N so that $|x_n - \frac{1}{2}| < \epsilon$ for $n > N$.

15. The sequence $x_n = \left(-\frac{2}{3}\right)^n$ has a limit $x_n \to 0$. For $\epsilon = 0.01$, find N so that $|x_n - 0| < \epsilon$ for $n > N$.

4.6 Calculating Sequence Limits

4.6.1 Overview

In the previous section, we learned about limits of sequences. Unfortunately, using a table of values to find a limit only allows us to estimate its value. A finite table alone can never clearly show whether a perceived pattern will continue or change after the values shown. It would be helpful to have some rules for finding limits based on the formula rather than numerical patterns.

This section establishes the rules of limits of sequences. Many limits can be calculated by identifying terms that are unbounded in the limit. We learn about how infinity behaves in the context of limit arithmetic. We also learn about indeterminate limit forms.

4.6.2 Infinite Limits

In order to compute limits of sequences, we begin with sequences that grow without bound, which is written $x_n \to \infty$ when the sequence grows in a positive direction or $x_n \to -\infty$ when the sequence grows in a negative direction. Arithmetic sequences with increments $\beta \neq 0$ (recall Theorem 13.2.8) must either steadily increase (positive increments $\beta > 0$) or steadily decrease (negative increments $\beta < 0$). The special case that $\beta = 0$ is somewhat boring, as this corresponds to a constant sequence so that the limit is just the constant value.

Theorem 4.6.1 Limit of an Arithmetic Sequence. *An arithmetic sequence with explicit formula $x_n = a+c\cdot n$ (for constants a and c) has unbounded growth when $c \neq 0$. The corresponding limit statements are*

$$\lim_{n\to\infty}(a+cn) = +\infty \quad (c>0)$$
$$\lim_{n\to\infty}(a+cn) = -\infty \quad (c<0)$$
$$\lim_{n\to\infty}(a) = a \quad (c=0)$$

Geometric sequences are a little more complicated, depending on the ratio ρ (recall Theorem 13.2.10) and the initial value. Repeated multiplication by a number whose magnitude is larger than 1 makes the resulting magnitude increase without bound. Repeated multiplication by a number whose magnitude is smaller than 1 makes the resulting magnitude converge to 0. If the ratio ρ is negative, then the sign of the sequence values will alternate between positive and negative. This is summarized by another theorem.

Theorem 4.6.2 Limit of a Geometric Sequence. *A geometric sequence with explicit formula $x_n = a\cdot\rho^n$ and ratio ρ is unbounded when $|\rho| > 1$, meaning that $|x_n| \to \infty$.*

- *If $\rho \leq -1$, x_n alternates sign and the limit does not exist.*
- *If $\rho > 1$, then the limit depends on the sign of a:*

$$\lim_{n\to\infty} a\cdot\rho^n = +\infty, \quad (a>0),$$
$$\lim_{n\to\infty} a\cdot\rho^n = -\infty, \quad (a<0).$$

- *If $|\rho| < 1$ (i.e., $-1 < \rho < 1$), then $\lim_{n\to\infty} a\cdot\rho^n = 0$.*

Example 4.6.3 Find the appropriate limits of the following sequences.

1. $\lim_{n\to\infty} 3 - 4n$
2. $\lim_{n\to\infty} 100 + 0.02n$
3. $\lim_{n\to\infty} -3 \cdot 1.05^n$
4. $\lim_{n\to\infty} 100 \cdot (-0.75)^n$
5. $\lim_{n\to\infty} 5 \cdot (-1.5)^n$

Solution.

1. The sequence $x_n = 3 - 4n$ is recognized as the explicit formula of an arithmetic sequence with increment $c = -4$. Since this is a negative increment, the sequence decreases without bound. So we write
$$\lim_{n\to\infty} 3 - 4n = -\infty.$$
2. The sequence $x_n = 100 + 0.02n$ is arithmetic with increment $c = 0.02$. Since the increment is positive, the sequence increases without bound and we write
$$\lim_{n\to\infty} 100 + 0.02n = +\infty.$$
3. The sequence $x_n = -3 \cdot 1.05^n$ is a geometric sequence with ratio $\rho = 1.05$. Because $\rho > 1$, the sequence grows without bound. Furthermore, because the terms are all negative, we have a limit
$$\lim_{n\to\infty} -3 \cdot 1.05^n = -\infty.$$
4. The sequence $x_n = 100 \cdot (-0.75)^n$ is a geometric sequence with ratio $\rho = -0.75$. Because the ratio is negative, the signs of the terms alternate between positive and negative. However, since $|\rho| = 0.75 < 1$, the magnitude of the terms converges to zero so that
$$\lim_{n\to\infty} 100 \cdot (-0.75)^n = 0.$$
5. The sequence $x_n = 5 \cdot (-1.5)^n$ has a negative ratio $\rho = -1.5$. Since $|\rho| > 1$, the terms have alternating signs but grow in magnitude. Consequently, $\lim_{n\to\infty} 5 \cdot (-1.5)^n$ does not exist.

□

4.6.3 Arithmetic of Infinity

Once we know how to identify when sequences have unbounded terms, we can use that information to find limits of related sequences. We can think of this as the arithmetic of infinity. Infinities can add and multiply but should never be subtracted or divided from one another. The signs of arithmetic involving infinity behave like for numbers, such as having a negative times a positive be negative.

The most important principle to remember is that infinities should never cancel one another. Cases where the formula looks like infinities *might* cancel are called indeterminate. This includes trying to cancel infinity by multiplying

by zero. An indeterminate limit form means that the value of the limit *can not be determined* without further analysis that resolves the apparent cancellation.

Theorem 4.6.4 Arithmetic Rules for Infinity. *Suppose unbounded sequences are combined using arithmetic operations. Then the following arithmetic relating to limits will be valid, where c will represent a positive number.*

- *Adding a number to infinity has no effect:*

$$+\infty \pm c = +\infty$$
$$-\infty \pm c = -\infty$$

- *Multiplying infinity by a non-zero number is still infinite, but changes sign if multiply by a negative number:*

$$c \cdot \pm\infty = \pm\infty$$
$$-c \cdot \pm\infty = \mp\infty$$

- *Adding infinities of the same sign are infinite. Don't cancel opposite infinities.*

$$+\infty + +\infty = +\infty$$
$$-\infty + -\infty = -\infty$$
$$+\infty + -\infty = \textit{indeterminate}$$

- *Multiplying infinities are infinite, and negative if opposite signs.*

$$+\infty \cdot +\infty = +\infty$$
$$-\infty \cdot -\infty = +\infty$$
$$+\infty \cdot -\infty = -\infty$$

- *The reciprocal of infinity is zero, but they can't cancel.*

$$\frac{1}{\pm\infty} = 0$$
$$0 \cdot \pm\infty = \textit{indeterminate}$$
$$\frac{\pm\infty}{\pm\infty} = \textit{indeterminate}$$
$$\frac{0}{0} = \textit{indeterminate}$$

The previous theorem was stated somewhat imprecisely in order to convey the idea of arithmetic of infinities without being bogged down by formal notation relating to limits. Each statement really is about a limit.

As an example, the arithmetic on infinity $+\infty + +\infty = +\infty$ would more carefully be stated as follows. Suppose that there are two sequences x_n and y_n such that $x_n \to +\infty$ and $y_n \to +\infty$. The sequence defined by $u_n = x_n + y_n$ has limit

$$\lim_{n\to\infty} x_n + y_n = +\infty.$$

The shorthand notation of performing arithmetic with infinity allows this to be simplified as writing

$$\lim_{n\to\infty} x_n + y_n = +\infty + +\infty = +\infty.$$

The intermediate step $+\infty + +\infty$ is not truly arithmetic, but points out that $x_n \to +\infty$ and $y_n \to +\infty$, and since those sequences were added, the final limit is also $+\infty$. We are substituting limits of individual terms into the formula defining the expression. As long as the arithmetic involves no cancellation of infinities, it will result in a correct statement.

To deal with indeterminate forms, we usually need to try to rewrite the formula in a new way so that the cancellation is avoided. The most common approach for rewriting is to factor out a **dominant term**. When there are infinities trying to cancel, we identify which of the terms should dominate and we factor that expression from both terms and simplify.

Example 4.6.5 Determine the following limits, if possible.

1. $\lim_{n\to\infty} 3 + \dfrac{5}{2^n}$
2. $\lim_{n\to\infty} \dfrac{n^2 - 3n}{5n - 1}$
3. $\lim_{n\to\infty} \dfrac{1 + 2^n}{3 + 5^n}$

Solution.

1. The sequence $x_n = 3 + \frac{5}{2^n}$ is the sum of terms 3 and $\frac{5}{2^n}$. The constant sequence has a limit $3 \to 3$ (since it never changes). The geometric sequence $\frac{5}{2^n} = 5 \cdot (\frac{1}{2})^n$ has a ratio $|\rho| < 1$ so that $\frac{5}{2^n} \to 0$. The form of the limit (using the terms) is

$$\lim_{n\to\infty} 3 + \frac{5}{2^n} = 3 + 0,$$

and since this does not involve any cancelation of infinities, will give the correct limit,

$$\lim_{n\to\infty} 3 + \frac{5}{2^n} = 3 + 0 = 3.$$

2. The sequence $u_n = \dfrac{n^2 - 3n}{5n - 1}$ is a quotient of terms $n^2 - 3n$ and $5n - 1$. To find the limit, we explore the terms individually first.

Because $n^2 = n \cdot n$, we know $n^2 \to +\infty \cdot +\infty = +\infty$. Similarly, the arithmetic sequence $3n \to +\infty$. However, the difference $n^2 - 3n$ would have a limit of the form $+\infty - \infty$, which is a cancellation of infinities. As written, $n^2 - 3n$ is an indeterminate form.

Our strategy will be to rewrite this as a product, and the best practice is to factor out (divide out) the greatest power of n (dominant term),

$$n^2 - 3n = n^2(\frac{n^2}{n^2} - \frac{3n}{n^2}) = n^2(1 - \frac{3}{n}).$$

From this, we find

$$\frac{3}{n} \to \frac{3}{+\infty} = 0 \quad \Rightarrow \quad 1 - \frac{3}{n} \to 1 - 0 = 1.$$

Since we already know $n^2 \to +\infty$, we have the limit of the numerator

$$\lim_{n\to\infty} n^2 - 3n = \lim_{n\to\infty} n^2(1 - \frac{3}{n}) = +\infty \cdot 1 = +\infty.$$

The term in the denominator $5n-1$ is an arithmetic sequence (linear function) with increment (slope) $c=5$. We know

$$\lim_{n\to\infty} 5n-1 = +\infty.$$

Unfortunately, that means our limit form as a quotient is itself an indeterminate form,

$$\lim_{n\to\infty} \frac{n^2-3n}{5n-1} = \frac{+\infty}{+\infty}.$$

We can not cancel infinities, so we must rewrite our formula.

For this example, I worked out the numerator separately to make a point about that term itself being an indeterminate form. In practice, our strategy will be to apply that factoring principle to the entire formula at one step. This is illustrated below.

The problem can be solved up if we just factor out from the numerator and denominator the dominant term (greatest power) and simplify as needed.

$$\begin{aligned}\lim_{n\to\infty} \frac{n^2-3n}{5n-1} &= \lim_{n\to\infty} \frac{n^2(1-\frac{3}{n})}{n(5-\frac{1}{n})}\\ &= \lim_{n\to\infty} \frac{n(1-\frac{3}{n})}{5-\frac{1}{n}}\\ &= \frac{+\infty\cdot(1-\frac{3}{\infty})}{5-\frac{1}{\infty}}\\ &= \frac{+\infty\cdot 1}{5} = +\infty.\end{aligned}$$

3. The sequence $w_n = \dfrac{1+2^n}{3+5^n}$, by quick inspection, involves the geometric sequences 2^n and 5^n, both of which grow exponentially so that $w_n \to \frac{+\infty}{+\infty}$. This indeterminate form involves canceling infinities, so we must rewrite the formula. Following the method of the previous example, we factor out the dominant term, in this case the geometrically growing powers.

$$\begin{aligned}\lim_{n\to\infty} \frac{1+2^n}{3+5^n} &= \lim_{n\to\infty} \frac{2^n(\frac{1}{2^n}+\frac{2^n}{2^n})}{5^n(\frac{3}{5^n}+\frac{5^n}{5^n})}\\ &= \lim_{n\to\infty} \frac{2^n(\frac{1}{2^n}+1)}{5^n(\frac{3}{5^n}+1)}\end{aligned}$$

This is still indeterminate form $\frac{+\infty}{+\infty}$, so we rewrite $\dfrac{2^n}{5^n} = (\frac{2}{5})^n$, which is a geometric sequence with ratio $\rho = \frac{2}{5}$ satisfying $|\rho| < 1$.

$$\begin{aligned}\lim_{n\to\infty} \frac{1+2^n}{3+5^n} &= \lim_{n\to\infty} \frac{2^n(\frac{1}{2^n}+1)}{5^n(\frac{3}{5^n}+1)}\\ &= \lim_{n\to\infty} \frac{(\frac{2}{5})^n(\frac{1}{2^n}+1)}{\frac{3}{5^n}+1}\\ &= \frac{0\cdot(0+1)}{0+1} = \frac{0}{1} = 0\end{aligned}$$

□

Knowing how to find limits of sequences with explicit formulas, we can also find limits for recursive sequences whose explicit formulas are known.

Example 4.6.6 Find the limit of a recursive sequence u defined by

$$u_n = 0.8u_{n-1} + 20$$

and initial value $u_0 = 50$.

Solution. A sequence that has a linear projection function, involving both multiplication by a ratio and the addition of an increment, has a shifted geometric sequence as its explicit formula. The equilibrium value is found by solving the fixed point equation.

$$\begin{aligned} 0.8x + 20 &= x \\ 20 &= 0.2x \\ 100 &= x \end{aligned}$$

Thus, the equilibrium value is $u^* = 100$.

The explicit formula is a geometric sequence with ratio $\alpha = 0.8$ shifted by the equilibrium,

$$\begin{aligned} u_n &= u^* + (u_0 - u^*)(0.8)^n \\ &= 100 + (50 - 100)(0.8)^n \\ &= 100 + -50(0.8)^n \end{aligned}$$

Be careful not to violate the order of operations by adding the 100 and -50 or multiplying the -50 and 0.8,

Using this explicit formula, we can find the limit of the sequence. The geometric sequence has a limit 0 because the ratio has magnitude smaller than 1.

$$\lim_{n\to\infty} u_n = 100 + (-50)(0) = 100$$

That is, the sequence converges to the equilibrium value. □

Example 4.6.7 Find the limit of the sequence defined by the recursive equation

$$x_{n+1} = 1.05x_n - 20$$

and initial value $x_0 = 300$.

Solution. Find the fixed point by solving the equation $x = 1.05x - 20$.

$$\begin{aligned} x &= 1.05x - 20 \\ -0.05x &= -20 \\ x &= 400 \end{aligned}$$

Using the fixed point $x^* = 400$ and the growth factor $\alpha = 1.05$, we can write down the explicit formula,

$$\begin{aligned} x_n &= x^* + (x_0 - x^*)\alpha^n \\ &= 400 + (300 - 400)1.05^n \\ &= 400 - 100 \cdot 1.05^n \end{aligned}$$

The geometric term with ratio $\alpha = 1.05$ grows without bound. The limit of the sequence can be found:

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} 400 - 100 \cdot 1.05^n$$

$$\begin{aligned} &= 400 - 100 \cdot +\infty \\ &= 400 - \infty \\ &= -\infty. \end{aligned}$$

The sequence will decrease without bound. □

4.6.4 Limit Applications to Modeling

With these tools, we can analyze sequences associated with physically meaningful models. Limits tell us about the long-term behavior. As time progresses, the sequence will progressively get closer and closer to the limit.

For example, the concentration of a drug in a patient taking repeated doses can be modeled by a sequence. A limit of this sequence can tell us something about what will happen to that concentration if the dosing continues for an extended amount of time.

Example 4.6.8 Suppose a patient begins taking 500 mg of a drug every four hours. However, the body metabolizes 60% of the drug in the body every four hours. Find a formula for the amount of drug in the body immediately after each dose and then determine the limiting value.

Solution. The patient's body starts with no drug. Immediately after the first dose, there are 500 mg. Four hours later, 60% has been removed and then another dose is added in. If we let D_n be the sequence of drug mass in the body as a function of the number of doses n, then this is modeled recursively by the equation

$$D_{n+1} = D_n - 0.6D_n + 500,$$

with initial value $D_1 = 500$.

This model has a linear projection function $f(x) = 0.4x + 500$ and corresponding fixed point

$$0.4x + 500 = x \quad \Leftrightarrow \quad x = \frac{500}{0.6} = \frac{2500}{3}.$$

The explicit formula, using Theorem 4.3.7, is given by

$$D_n = \frac{2500}{3} + (500 - \frac{2500}{3}) \cdot 0.4^{n-1} = \frac{2500}{3} - \frac{1000}{3} \cdot 0.4^{n-1}.$$

Because the slope of the projection function $\alpha = 0.4$ has magnitude less than 1, the limiting value is the fixed point $x^* = \frac{2500}{3} \approx 833.33$. Thus, if the patient continues to take the drug, the amount in the body immediately after each dose will be approximately 833.33 mg. (Immediately before the dose, it must have been approximately 333.33 mg.) □

In the coming chapters, we will learn about the definite integral. The calculation of an integral is as a limit of a summation. The following example illustrates how such calculations are performed.

Example 4.6.9 Find $\lim_{n\to\infty} \sum_{k=1}^{n} \left(1 + \frac{3k}{n}\right) \cdot \frac{3}{n}$.

Solution. There are two major steps needed for this problem: (i) find an explicit formula for the summation that depends only on n and (ii) compute the limit of that explicit formula.

To compute the explicit formula for the summation, we need to remember that the variable n is a constant with respect to the summation index k. We will rewrite the formula of the sequence in summation to be a sum so that we

can use the linearity property.

$$\begin{aligned}\sum_{k=1}^{n}\left(1+\frac{3k}{n}\right)\cdot\frac{3}{n} &= \sum_{k=1}^{n}\frac{3}{n}+\frac{9k}{n^2}\\ &= \frac{3}{n}\cdot\sum_{k=1}^{n}1+\frac{9}{n^2}\cdot\sum_{k=1}^{n}k\\ &= \frac{3}{n}\cdot n+\frac{9}{n^2}\cdot\frac{n(n+1)}{2}\\ &= 3+\frac{9(n+1)}{2n}\\ &= 3+\frac{9n+9}{2n}\end{aligned}$$

We see that the summation is itself a sequence involving an index n. We find the limit of that sequence. The fraction $\frac{9n+9}{2n}$ will be indeterminate form $\frac{\infty}{\infty}$, so we will factor out the dominant term of n from top and bottom.

$$\begin{aligned}\lim_{n\to\infty}\sum_{k=1}^{n}\left(1+\frac{3k}{n}\right)\cdot\frac{3}{n} &= \lim_{n\to\infty}3+\frac{9n+9}{2n}\\ &= \lim_{n\to\infty}3+\frac{n(9+\frac{9}{n})}{2n}\\ &= \lim_{n\to\infty}3+\frac{9}{2}+\frac{9}{2n}\\ &= 3+\frac{9}{2}+\frac{9}{\infty}\\ &= 3+\frac{9}{2}+0\\ &= \frac{15}{2}=7\frac{1}{2}\end{aligned}$$

□

4.6.5 Summary

- Arithmetic sequences with non-zero increments are unbounded. Geometric sequences are unbounded when the ratio has magnitude greater than 1 and converge to zero when the ratio has magnitude less than 1.
- Sequence limits obey the standard rules of arithmetic, including for infinite limits, with the exception that any formula that would cancel infinity are indeterminate. (See Theorem 4.6.4) Indeterminate means the limit can not be determined from simple arithmetic but requires rewriting in another form.
  - $\infty-\infty$: rewrite by factoring out dominant term
  - $\frac{\infty}{\infty}$: factor out dominant term in numerator and denominator, look to simplify
  - $\frac{0}{0}$: try to factor and simplify
  - $0\cdot\infty$: use negative powers to rewrite as fraction, then treat as $\frac{0}{0}$ or $\frac{\infty}{\infty}$

  Look for terms in the limit that vanish: $\frac{1}{\infty}=0$.

4.6.6 Exercises

Use limit arithmetic to find the exact value for each limit. If the limit does not exist, explain why.

1. $\lim_{n\to\infty} 3 - 4n$

2. $\lim_{n\to\infty} -5 + \frac{n}{200}$

3. $\lim_{n\to\infty} 3 \cdot (\frac{2}{5})^n$

4. $\lim_{n\to\infty} -3 \cdot (-0.8)^n$

5. $\lim_{n\to\infty} 4 - 5 \cdot (1.05)^n$

6. $\lim_{n\to\infty} 5 - (-1.5)^n$

7. $\lim_{n\to\infty} \frac{5 + 3n}{5 - n^2}$

8. $\lim_{n\to\infty} \frac{5 + 3n^2}{5 - n^2}$

9. $\lim_{n\to\infty} \frac{5 + 3n^3}{5 - n^2}$

10. $\lim_{n\to\infty} \frac{2^n - 3^n}{7 + 2 \cdot 3^n}$

11. $\lim_{n\to\infty} \frac{3 \cdot 2^n - 5 \cdot 3^n}{1 + 2 \cdot 5^n}$

Find an explicit formula for each sequence. Then determine the limit of the sequence.

12. $x_t = 1.05x_{t-1} - 10$ with $x_0 = 500$.

13. $x_{k+1} = 0.8x_k + 12$ with $x_1 = 5$.

14. $y_n = -1.5y_{n-1} + 5$ with $y_0 = 2$.

Use the properties and elementary formulas for summation to find a formula for the summation in order to compute the limit.

15. $\lim_{n\to\infty} \sum_{k=1}^{n} \frac{5k}{n^2}$

16. $\lim_{n\to\infty} \sum_{k=1}^{n} \frac{k^2}{n^3}$

17. $\lim_{n\to\infty} \sum_{k=1}^{n} \left(2 + \frac{3k}{n}\right) \cdot \frac{3}{n}$

18. $\lim_{n\to\infty} \sum_{k=1}^{n} \left(1 + \frac{2k}{n}\right)^2 \cdot \frac{2}{n}$

Applications of sequences.

19. When a patient starts taking a drug, there is no drug in the blood. The prescription is to take 250 mg every six hours. Then, during the six hours between doses, the body metabolizes 20% of whatever drug is in the body.

Find a recursive formula using a linear projection function for the sequence of drug levels with an index counting the total number of doses. Clearly state the initial value. Find an explicit formula for the sequence and determine the limit. Is there a steady state amount of drug in the body?

20. A pond has been polluted by a stream. The pond holds 12,000 gallons of water and the stream replaces 900 gallons per day. The stream has been carrying the chemical pollutant at a concentration of $0.4 \frac{\text{g}}{\text{gal}}$ and now the pond has that same pollution level. With stream cleanup, the chemical pollutant in the stream has just been reduced to a concentration of $0.1 \frac{\text{g}}{\text{gal}}$.

Find a recursive formula using a linear projection function for the sequence of pollutant levels (total mass) in the pond with an index counting the number of days since the cleanup occurred. Clearly state the initial value. Find an explicit formula for the sequence and determine the limit. On what day will the total amount of pollutant in the pond be reduced to half of the pollution level prior to cleanup? Is there a steady state amount of pollutant in the pond?

Chapter 5

Limits and Differentiability

5.1 An Overview of Calculus

The previous chapter 13 studied sequences. There are three major concepts in calculus that we can use sequences to motivate. These are limits, derivatives, and integrals.

In this chapter, we will focus on the idea of the definite integral as a generalization of accumulating increments of change. Thinking of sequences in terms of their increments of change is simpler because the domain consists only of integers which are equally spaced. More general functions are defined with domains consisting of intervals of the real numbers. (Some functions can be defined on even more complex sets, which gives rise to even more advanced mathematics.) Consequently, we can not think only in terms of increments of change but in terms of a **rate of change**.

5.1.1 Derivatives and Integrals

For sequences, we learned to think of complementary ideas of accumulation sequences and increments. With a sequence x, we had a forward difference

$$\Delta x_n = x_{n+1} - x_n$$

and a backward difference

$$\nabla x_n = x_n - x_{n-1}.$$

These differences measure the change in the sequence x for consecutive values of the index, which plays the role of the independent variable.

For functions defined on intervals, there is no meaning to *consecutive* values of the independent variable. Near a point of interest $x = c$, there are infinitely many other values close to c. Consequently, when measuring the change of a function Δf, we must also specify the change in the independent variable Δx. Consider two values for the independent variable, say $x = a$ and $x = b$, and we define $\Delta x = b - a$ and $\Delta f = f(b) - f(a)$.

Different increments Δx will usually result in different function increments Δf. However, for many functions, the ratio $\Delta f / \Delta x$, called the **average rate of change**, has a limit as $\Delta x \to 0$. This limiting rate of change is called the **instantaneous rate of change** and in calculus is named the **derivative**.

Definition 5.1.1 Instantaneous Rate of Change. Given a function f that relates variables $x \overset{f}{\mapsto} y$, the **instantaneous rate of change** of y with respect to x is the **derivative** $\dfrac{dy}{dx}$ defined by

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x},$$

if the limit exists. Consequently, for sufficiently small increments Δx, we have

$$\Delta f \approx \frac{dy}{dx} \cdot \Delta x.$$

$\Diamond$

The following example illustrates the role of the instantaneous rate of change to relate the increments of the independent variable with the increments of the dependent variable.

Example 5.1.2 A ball dropped from a tower has a height h (measured in feet) modeled as a function of time t (measured in seconds) given by

$$t \overset{f}{\mapsto} h = 40 - 16t^2.$$

At $t = 1$, the instantaneous rate of change is $\dfrac{dh}{dt} = -32$.

This rate of change is illustrated in the dynamic figure below. Thinking of $t_0 = 1$ as one value of the independent variable, you can adjust the second value t_1 to establish the increment $\Delta t = t_1 - t_0$. The function automatically computes $f(1)$ and $f(t_1)$ and shows $\Delta h = f(t_1) - f(1)$. The ratio $\Delta h/\Delta t$ will be close to $-32\,\frac{\text{ft}}{\text{s}}$ for small values of Δt.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 5.1.3

We can recover this instantaneous rate of change using limits, as shown in the solution below.

Solution. We know that $\Delta t = t_1 - 1$ and $\Delta h = f(t_1) - f(1)$. Using the formula, this gives

$$\Delta h = (40 - 16t_1^2) - (40 - 16(1)^2) = -16(t_1^2 - 1).$$

The average rate of change is defined by the quotient

$$\frac{\Delta h}{\Delta t} = \frac{-16(t_1^2 - 1}{t_1 - 1},$$

which has a value for all $t_1 \neq 1$.

The instantaneous rate of change is the limit of the average rate of change as $\Delta t \to 0$, which in this case requires $t_1 \to 1$. Even though the quotient is not defined at $t_1 = 1$, we can simplify the formula used on the sides to a formula that is defined.

$$\begin{aligned}
\frac{dh}{dt} = \lim_{\Delta t \to 0} \frac{\Delta h}{\Delta t} &= \lim_{t_1 \to 1} \frac{-16(t_1^2 - 1)}{t_1 - 1} \\
&= \lim_{t_1 \to 1} \frac{-16(t_1 + 1)(t_1 - 1)}{t_1 - 1} \\
&= \lim_{t_1 \to 1} -16(t_1 + 1) \\
&= -16(1 + 1) = -32
\end{aligned}$$

The key step in this limit calculation was changing the limit expression from one in which the formula is not continuous to a new formula. When the formula is continuous, we can just evaluate it at the point of interest. The variable t_1 used in the limit could have been chosen to be any convenient name. □

At this point, our emphasis is understanding that the rate of change or derivative measures the limiting ratio for increments of change in the value of the function to corresponding increments of change in the independent variable. Not every function has a derivative. We will study the calculation of the derivative in more depth in later chapters.

Computing a derivative for a given function is analogous to computing the increments of a sequence. The complementary calculation for sequences is to compute the accumulation sequence for given increments. That is, if $x = (x_n)_{n=1}^{\infty}$ is a sequence of increments, then the accumulation sequence u

with increments $\nabla u_n = u_n - u_{n-1} = x_n$ and initial value u_0 was written

$$u_n = u_0 + \sum_{k=1}^{n} x_k.$$

The calculus analogue is to be given a function that represents a rate of change and use it to find a new function, the **accumulation function**, that has that rate of change as its derivative. Suppose $f(x)$ is the rate of change or derivative of a quantity Q with respect to x,

$$\frac{dQ}{dx} = f(x).$$

We are then interested in finding Q as a function of x if we know an initial value $Q(x_0) = Q_0$. The analogue of summation of increments is the **definite integral**, and we will write

$$Q = Q_0 + \int_{x_0}^{x} f(z)\, dz.$$

The rest of this chapter is focused on bringing meaning to the idea of the definite integral. We study definite integrals before derivatives because we have just studied summation and sequences. The calculations involved in developing the ideas of definite integrals apply these concepts. Ultimately, the Fundamental Theorem of Calculus will provide a connection between the definite integral and the derivative, showing that our two ideas of rate of change represent the same thing.

5.1.2 A Technological Aside

Computational tools play an important role in the real-world application of mathematics. It is increasingly common to have a tool perform actual computations with the user responsible to formulate the appropriate problem.

For example, you may have heard of the website WolframAlpha. This site acts like a search engine for mathematical content, and you can enter queries like "`factor x^2+3x`". The ability extends to calculus tools as well. We might have asked for our earlier example "`derivative of 40-16t^2 at t=1`".

Disadvantages of a site like WolframAlpha is that you are limited to a single query at a time and it can sometimes be hard to state precisely what you want. More powerful tools are available, including advanced programmable calculators and commercial software tools like Wolfram's Mathematica and MapleSoft's Maple programs.

A free, but similarly powerful tool is SageMath. A calculation in SageMath uses a **script** based on the Python programming language. Comments in the scripts follow the # symbol and are ignored by the computer but are useful to understand what is happening.

Example 5.1.4 To factor the formula $x^2 + 3x$, we would use the following script.

```
# Tell Sage that x is a variable
var("x")
# Ask Sage to factor. Include the multiplication *
factor(x^2+3*x)
```

```
(x+3)*x
```

□

Example 5.1.5 To find the derivative of $40 - 16t^2$ at $t = 1$, we would use the following script.

```
# Tell Sage that t is an independent variable
var("t")
# Define h as function of t
# -- Notice how every operation must be typed
h(t) = 40-16*t^2
show(h(t))
# The derivative is also a function
# but let Sage figure it out using the derivative operation.
Dh(t) = derivative(h(t), t)
show(Dh(t))
# Find the value of the derivative at t=1
Dh(1)
```

```
-16*t^2+40
-32*t
-32
```

□

Example 5.1.6 A container of water has a volume V. Suppose that the volume has an instantaneous rate of change with respect to time t given by

$$\frac{dV}{dt} = -40 + 3t.$$

When $\frac{dV}{dt}$ is negative, the volume is decreasing; when $\frac{dV}{dt}$ is positive, the volume is increasing. The expression defines exactly how fast the water is entering or leaving the container. Find the volume of water as a function of time if $V = 500$ when $t = 1$.

The following SageMath script will start by defining the formula for the rate of change. It then uses a definite integral to create the variable for the volume,

$$V(t) = 500 + \int_1^t -40 + 3z \, dz.$$

```
# Define the independent variable.
var("t")
# Define dV as a function for rate
DV(t) = -40+3*t
show(DV(t))
# Define the V using integral, but need dummy variable
var("z")
V(t) = 500 + integrate(DV(z), [z, 1, t])
show(V(t))
```

```
3*t-40
3/2*t^2-40*t+1077/2
```

The integration variable z was needed in the integral for the same reason that a summation in sequence accumulations requires a dummy index variable. The formula `DV(z)` represents the formula for the rate of change evaluated at this integration variable instead of t, $-40+3z$. This could have been computed in WolframAlpha with the query `integrate -40+3z with respect to z from 1 to t`. □

5.1.3 Summary

- Calculus is developed using ideas similar to those for sequences—limits, increments, and accumulation— to limits of functions, derivatives, and integrals.
- The **derivative** $\dfrac{dQ}{dx}$ measures the **instantaneous rate of change** of a quantity Q with respect to the independent variable x, represented by a limit,
$$\frac{dQ}{dx} = \lim_{\Delta x \to 0} \frac{\Delta Q}{\Delta x}.$$
Consequently, increments of change in Q, ΔQ, will be approximately proportional to the increment in x,
$$\Delta Q \approx \frac{dQ}{dx} \cdot \Delta x,$$
for sufficiently small Δx.
- Given a function f' for the rate of change of a quantity Q, $x \overset{f'}{\mapsto} \frac{dQ}{dx}$, and an initial value Q_0 when $x = x_0$, the **accumulation function** will be that function with derivative $\frac{dQ}{dx} = f'(x)$, represented by the integral
$$Q = Q_0 + \int_{x_0}^{x} f'(z)\, dz.$$
- Computational tools, such as WolframAlpha and SageMath, are available to perform these calculations, leaving us the responsibility of formulating problems and interpreting the results.

5.1.4 Exercises

Use appropriate tables to approximate the following function limits. For a two-sided limit, be sure that your work verifies that both sides approximate the same value

1. $\displaystyle\lim_{x \to 3^-} \frac{2^x - 8}{x - 3}$

2. $\displaystyle\lim_{x \to 3^+} \frac{2^x - 8}{x - 3}$

3. $\displaystyle\lim_{x \to 2} \frac{x^2 - 4}{2^x - 4}$

4. $\displaystyle\lim_{x \to 1} \frac{x^2 - 1}{|x - 1|}$

Find the instantaneous rate of change for the relationship described in each problem using the limit of the average rate of change between the given point and a second variable point. Compare the instantaneous rate to the average rate for the specified increments.

5. An object tossed into the air has a height that changes in time. Let h measure the height from the ground in feet and let t measure the time since the object was tossed in seconds. Then h has a model
$$t \mapsto h = 4 + 30t - 16t^2.$$

Find $\dfrac{dh}{dt}$ at $t = 1$ and compare this to the average rate $\dfrac{\Delta h}{\Delta t}$ with $\Delta t = 0.1$.

6. The material cost for producing an aluminum box the shape of a cube is a function of the size of the cube. Let C be the cost in dollars and let s measure the length of each side of the box in centimeters. Then C has a model

$$s \mapsto C = 0.03s^2.$$

Find $\dfrac{dC}{ds}$ at $s = 10$ and compare this to the average rate $\dfrac{\Delta C}{\Delta s}$ with $\Delta s = -0.2$.

7. For a circle of radius r, the area A satisfies a relation

$$r \mapsto A = \pi r^2.$$

Find $\dfrac{dA}{dr}$ at $r = 2$ and compare this to the average rate $\dfrac{\Delta A}{\Delta r}$ with $\Delta r = 0.05$.

For each problem, write down the formula involving an integral for the quantity whose derivative and initial value are given. Use technology to find the algebraic formula of the quantity.

8. Given $\dfrac{dy}{dx} = 4$ and $y = 5$ when $x = 2$. Find y as a function of x.

9. Given $\dfrac{dy}{dx} = 2 + 3x$ and $y = 4$ when $x = 1$. Find y as a function of x.

10. Given $\dfrac{dQ}{dt} = t^3$ and $Q = 2$ when $t = 1$. Find Q as a function of t.

11. Given $\dfrac{dP}{dt} = 500e^{0.2t}$ and $P = 4000$ when $t = 0$. Find P as a function of t.

5.2 Functions Defined on Intervals

Overview. In this section, we consider how functions are defined on different sets. We learn about the domain of the function and how to find the domain given a formula. Finding the natural domain of a function involves solving inequalities using sign analysis. Sometimes, we need to restrict a function to use a rule only on a particular set, called the explicit domain. Other times, we need the function to use different rules on different sets, creating a piecewise function. With piecewise functions, a function might not be continuous. We learn about limit notation as a way of evaluating what a rule to the left or right of a point would have given at a point. Continuity requires that the function is defined and that the left- and right-limits both agree with the actual function value.

5.2.1 Functions and Sets

We earlier learned that sequences are functions. A sequence x defined a map $n \mapsto x_n$ from the value of the index to the value in the sequence list at that index position. An explicit definition of the sequence might even use a formula, say $x_n = 2n + 5$. An interactive figure below illustrates this mapping. As you move the value on the n-axis, an arrow shows the corresponding value on the x_n-axis. However, because n must be an integer, the sequence value is not defined for any other values on the axis.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 5.2.1 The sequence $x_n = 2n + 5$ for $n = 0, 1, 2, \ldots$ as a map.

We also learned that if we had an equation involving two variables, say x and y, and could solve that equation for y as a dependent variable being equal to an expression in x, then the map $x \mapsto y$ also defined a function. For example, we might have $y = 2x + 5$.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 5.2.2 The function $y = 2x + 5$ as a map.

The equations $x_n = 2n + 5$ and $y = 2x + 5$ involve the same operations or rule for going from the independent variable to the dependent variable. Nevertheless, we think of these as fundamentally different functions. The sequence only allows n to have integer values, but the dependent variable y would allow x to have any real value. In order to distinguish functions at this level, we must extend the definition of a function to include the domain and codomain.

Definition 5.2.3 Function. A function f is a rule or relation from a given set D (the domain) to another set D' (the codomain) such that *every* value $a \in D$ is related (mapped) to a *unique* value $b \in D'$. We write $f : D \to D'$. $\Diamond$

Our notation for a function involving sets uses a different arrow $f : D \to D'$ than the mapping arrow we used earlier for independent variable to dependent variable, $f : x \mapsto y$. Think of the set arrow $(\to)$ as specifying sets and the mapping arrow $(\mapsto)$ as specifying variables. When a function is described in terms of both sets and variables, we can use both to completely characterize the function.

Example 5.2.4 The sequence $x = (2n + 5)_{n=0}^{\infty}$ is a function from a domain $D = \mathbb{N}_0 = \{0, \ldots, \infty\}$ to a codomain of the real numbers $\mathbb{R}$. The notation that

gives indicates all of this would be

$$x : \mathbb{N}_0 \to \mathbb{R}; n \mapsto x_n = 2n + 5.$$

□

Sets relating to functions used in calculus, including domains, are usually expressed as a union of intervals. An interval represents all real numbers from a connected segment of the real number line. The left end-point of the segment is listed first and the right end-point is listed second, using infinity if the segment continues indefinitely. A square bracket is used when the end-point is included (closed) and a round parenthesis is used when the end-point is not included (open). For a review of interval notation, see Subsection A.1.2.

Example 5.2.5 The function $f(x) = 2x + 5$ is a function corresponding to $y = 2x + 5$. To indicate that the value of x is allowed to be any real number, we could write

$$f : \mathbb{R} \to \mathbb{R}; x \mapsto 2x + 5.$$

The set of all real numbers can also be represented as an interval, $(-\infty, \infty)$, so the function could also have been written

$$f : (-\infty, \infty) \to (-\infty, \infty); x \mapsto 2x + 5.$$

Instead of using mapping notation for the formula, we could also have written

$$f : (-\infty, \infty) \to (-\infty, \infty); f(x) = 2x + 5.$$

□

When we want to restrict the domain to a set smaller than the natural domain, we can use mapping notation as described above, or we can use a conditional statement on the formula. A conditional statement provides the condition for when the equation or rule should be applied.

Example 5.2.6 The function f defined by

$$f : [0, 1] \to \mathbb{R}; x \mapsto 2x + 5$$

has a domain $D = [0, 1]$. The interval corresponds to values x that satisfy $0 \leq x \leq 1$. Consequently, we could also write the function using a conditional statement as

$$f(x) = 2x + 5, \quad 0 \leq x \leq 1.$$

The graphical representations of f as a map and as a graph are shown in the interactive figures below. Note how the value of the dependent variable is undefined for values of the independent variable outside of the restricted domain.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 5.2.7 The restricted function $f(x) = 2x + 5$ for $0 \leq x \leq 1$ as a map.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 5.2.8 The graph of the restricted function $y = f(x) = 2x + 5$ for $0 \leq x \leq 1$ in the (x, y) plane.

□

Restricting a domain is necessary to define inverse functions when a function is not one-to-one. For example, we have earlier noted that $y = x^2$ is not one-

to-one because when solving for x, we get two solutions $x = \pm\sqrt{y}$. The next example explores this in more depth.

Example 5.2.9 We intuitively, but incorrectly, think of $f(x) = x^2$ and $g(x) = \sqrt{x}$ as inverse functions. The composition $g \circ f(x) = \sqrt{x^2}$ is not the identity because $\sqrt{x^2} = |x|$. This is illustrated in the figure below. For $x < 0$, $g \circ f(x) \neq x$; but for $x \geq 0$, $g \circ f(x) = x$.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 5.2.10 The chain $x \mapsto u = x^2 \mapsto y = \sqrt{u}$.

The function restricted to this domain, $f : [0, \infty) \to \mathbb{R}; x \mapsto x^2$, can be written using a standard equation with a constraint as

$$f(x) = x^2, \quad x \geq 0.$$

The restricted function is the inverse of the square root. □

5.2.2 Finding the Domain and Range

When a function is defined by a formula, as in $f(x) = 2x + 5$, if the domain is not specified, then the largest domain consistent with the formula is assumed. We call this the **natural domain** of the function. Related to the domain is a set known as the **range**, which is the set of all output values. The range is always a subset of the codomain.

Definition 5.2.11 For a function f defined by a formula, such as $y = f(x)$, the **natural domain** is the set of all real numbers for which the formula is defined. ♢

Definition 5.2.12 For a function $f : D \to D'$, the **range** is the set of all values y for which there exists a state (x, y). That is, there exists $x \in D$ so that $f(x) = y$. ♢

We find the **natural domain** by identifying which operations might not be defined for all values and then solve either equations or inequalities that will identify where the function is defined. Our elementary operations and functions use the following constraints to find the domain.

- Division is undefined if the denominator equals zero.
- Even roots (e.g., square roots) and irrational powers are undefined if the inner expression is negative.
- Logarithms are undefined if the inner expression is non-positive (zero or negative).

Example 5.2.13 Determine the domain of $f(x) = \dfrac{2x + 3}{x^2 - 4}$.

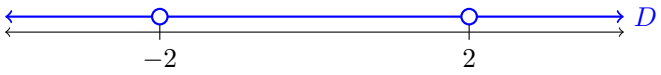
Solution. Because $f(x)$ is defined as a quotient, the domain will be the set of all values except where $x^2 - 4 = 0$. We solve this equation by factoring, since a product can only equal zero if one of the factors equals zero.

$$\begin{gathered} x^2 - 4 = 0 \\ (x + 2)(x - 2) = 0 \\ x + 2 = 0 \quad \text{or} \quad x - 2 = 0 \\ x = -2 \quad \text{or} \quad x = 2 \end{gathered}$$

This means $f(x)$ is defined for all inputs except $x = -2$ or $x = 2$.

To describe the domain using intervals, we think of the real number line and remove $x = \pm 2$. A graphical representation of the set using a number line is shown below. Intervals are read from the line left-to-right. It starts at $-\infty$ and continues until -2, then goes from -2 to 2, and finally goes from 2 until $+\infty$. We write

$$D = (-\infty, -2) \cup (-2, 2) \cup (2, +\infty).$$



□

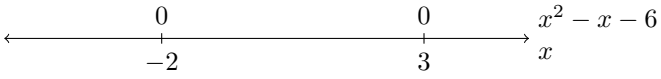
Sometimes finding the domain of a function involves solving an inequality (such as for a square root or a logarithm). To solve the inequality, we perform sign analysis. We identify end points of intervals where the expression of interest *might* change sign by solving equations. These end points only occur where the expression equals zero or where the expression itself is undefined (a discontinuity). We test the sign of the expression in each of the resulting intervals by using test points.

Example 5.2.14 Find the domain of the function $g(x) = \log_4(x^2 - x - 6)$.

Solution. The logarithm in $g(x)$ will only have a real value when the input expression is positive, $x^2 - x - 6 > 0$. Our task becomes determining the signs of the expression $x^2 - x - 6$. First, we find possible sign-changing points. The expression is always defined (no discontinuities) so we just solve for zeros $x^2 - x - 6 = 0$.

$$\begin{gathered} x^2 - x - 6 = 0 \\ (x-3)(x+2) = 0 \\ x - 3 = 0 \quad \text{or} \quad x + 2 = 0 \\ x = 3 \quad \text{or} \quad x = -2 \end{gathered}$$

If we mark these points on a number line, we can easily identify the intervals to test for signs. It is helpful to use the same number line to record the resulting signs, so we can label x-values below the line and the resulting sign or value of the expression above the line.



The number line shows we need to test the intervals $(-\infty, -2)$, $(-2, 3)$, and $(3, \infty)$. Choosing one value from each interval, we can evaluate the expression at that point and identify the sign.

$$\begin{aligned} x = -3 \quad &\Rightarrow \quad x^2 - x - 6 = (-3)^2 - (-3) - 6 = 6 \\ x = 0 \quad &\Rightarrow \quad x^2 - x - 6 = 0^2 - 0 - 6 = -6 \\ x = 4 \quad &\Rightarrow \quad x^2 - x - 6 = 4^2 - 4 - 6 = 6 \end{aligned}$$

We can now update the number line we started by recording either $+$ or $-$ above each interval that we tested.

We were finding the domain of $g(x) = \log_4(x^2 - x - 6)$, which requires $x^2 - x - 6 > 0$. Based on our summary, we need to find all values which result in the expression having a positive sign. So our solution is the set D formed

from the union of intervals $(-\infty, -2)$ and $(3, \infty)$,

$$D = (-\infty, -2) \cup (3, \infty).$$

A visualization of the domain on the number line might also help solidify the connections between the sign analysis number line and the domain set.

□

Example 5.2.15 Find the domain of the function $h(x) = \sqrt{\dfrac{4x}{x^2 - 9}}$.

Solution. A square root (any even root) requires that the input expression is non-negative. Our domain is to solve the inequality

$$D = \{x : \frac{4x}{x^2 - 9} \geq 0\}.$$

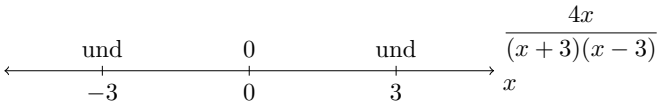
To use sign analyis, we need to know the zeros and discontinuities and then test each resulting interval. Discontinuities occur when we try to divide by zero.

$$\begin{gathered} x^2 - 9 = 0 \\ (x+3)(x-3) = 0 \\ x + 3 = 0 \quad \text{or} \quad x - 3 = 0 \\ x = -3 \quad \text{or} \quad x = 3 \end{gathered}$$

Zeros for a quotient require that the numerator equals zero.

$$\begin{gathered} 4x = 0 \\ x = 0 \end{gathered}$$

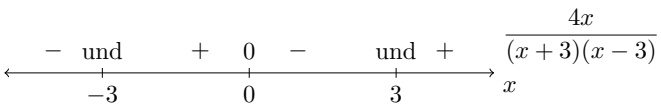
Our sign analysis number line will have three points.



Checking one point in each resulting interval gives us the sign. Because we only need to know the sign, it is simpler to think of factors of positive or negative values.

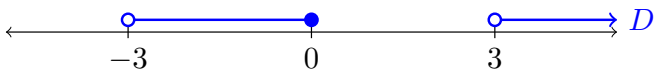
$$\begin{aligned} x = -4 \quad &\Rightarrow \quad \frac{4x}{(x+3)(x-3)} = \frac{4(-4)}{(-4+3)(-4-3)} = \frac{(-)}{(-)(-)} \\ x = -1 \quad &\Rightarrow \quad \frac{4x}{(x+3)(x-3)} = \frac{4(-1)}{(-1+3)(-1-3)} = \frac{(-)}{(+)(-)} \\ x = 1 \quad &\Rightarrow \quad \frac{4x}{(x+3)(x-3)} = \frac{4(1)}{(1+3)(1-3)} = \frac{(+)}{(+)(-)} \\ x = 4 \quad &\Rightarrow \quad \frac{4x}{(x+3)(x-3)} = \frac{4(4)}{(4+3)(4-3)} = \frac{(+)}{(+)(+)} \end{aligned}$$

The signs can be summarized on the number line.



We interpret our analysis to find the domain of $h(x)$. The domain must include intervals where the inner expression is positive, $(-3, 0)$ and $(3, \infty)$, along with points where the expression equals zero, $x = 0$. The set is visualized below. We do not include the points where the expression was undefined, $x = \pm 3$. The domain is the set

$$D = (-3, 0] \cup (3, \infty).$$



□

When a function is one-to-one, it has an inverse. Because the input and output values for a function and its inverse exactly switch roles, we can find the range of a function by finding the domain of its inverse.

Theorem 5.2.16 *Suppose $f : D \to D'$ is one to one, and let R be the range of f. Then the domain of f^{-1} is R and the range of f^{-1} is D.*

Example 5.2.17 Find the range of $f(x) = \dfrac{3x}{x+4}$.

Solution. We see if f is one-to-one by finding the inverse. Start with the equation $y = \dfrac{3x}{x+4}$ and solve for x. First, clear the fraction by multiplying both sides by $x + 4$. Then collect terms involving x.

$$\begin{aligned} y &= \frac{3x}{x+4} \\ y(x+4) &= 3x \\ xy + 4y &= 3x \\ xy - 3x &= -4y \end{aligned}$$

Now factor out the common factor of x and finish solving.

$$\begin{aligned} x(y-3) &= -4y \\ x &= \frac{-4y}{y-3} \end{aligned}$$

The inverse function is $f^{-1}(y) = \dfrac{-4y}{y-3}$.

The domain of f^{-1} is the set of all numbers y so $y - 3 \neq 0$. That is, $y \neq 3$. In interval notation, this is $(-\infty, 3) \cup (3, \infty)$. Because the domain of the inverse function is the range of the original function, we know that the range of f is $(-\infty, 3) \cup (3, \infty)$. □

When a function is not one-to-one, we will need to know how to find extreme values to find the range of a function. That will have to wait until we know about derivatives.

5.2.3 Piecewise Defined Functions

When different rules or formulas are used for different conditions, we have a **piecewise-defined function**. The standard notation for piecewise function is to create a list of rules using an equation with conditional statements on the domain for each given rule. To satisfy the unique-output property of a function, the conditional statements should ensure that each value in the domain only gets one output value.

Example 5.2.18 Describe the piecewise function

$$f(x) = \begin{cases} 3, & x < 0, \\ x^2, & 0 < x \leq 2, \\ 4 - x, & x > 2. \end{cases}$$

Include a graph.

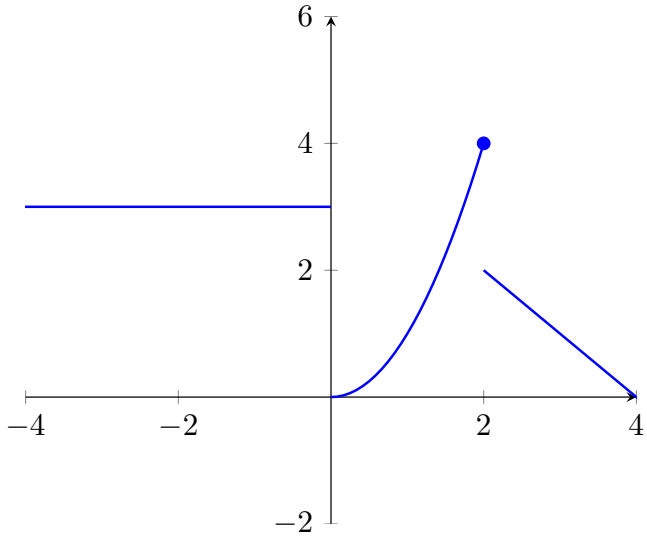
Solution. The function $f : x \mapsto y$ is listed with three different rules. For inputs $x \in (-\infty, 0)$ (i.e., $x < 0$), we use the rule $y = 3$; for inputs $x \in (0, 2]$, we use the rule $y = x^2$; and for $x \in (2, \infty)$, we use the rule $y = 4 - x$. Notice that for $x = 0$, there is no rule provided. The domain of f is the union of the component domains, so

$$D = (-\infty, 0) \cup (0, \infty).$$

Notice how we would evaluate the function at different points. Looking at the input, we determine which rule applies and then use only that rule.

$$\begin{aligned} f(-2) &= 3 \\ f(1) &= 1^2 = 1 \\ f(2) &= 2^2 = 4 \\ f(2.01) &= 4 - 2.01 = 1.99 \end{aligned}$$

The graph of the relation $y = f(x)$ is created by pasting the graphs $y = 3$, $y = x^2$, and $y = 4 - x$ into a single graph, but including only that portion of the graphs that is relevant for the constrained domains of those rules. For points at the edge of a domain interval, we use filled circles only when the point is explicitly included.



□

In the previous example, the graph of the function had breaks. Those breaks occurred at the edges of the constrained domains of the rule. We call such a break a **discontinuity**. When a function is connected, we say it is **continuous**. For a piecewise function to be continuous at a point, the rule used to the left and right of a point need to give the same value as the rule at the point.

We need a notation that says to use the different rules around a point. Function evaluation notation $f(x)$ finds the value using the rule at the point. We use a new notation, called **limit notation**, to apply the rules coming from the left or from the right to predict the value at a point.

Definition 5.2.19 Intuitive Meaning of Limit Notation. For a piecewise function using otherwise continuous expressions around a point $x = c$,

$$f(x) = \begin{cases} f_{\text{left}}(x), & x < c, \\ f_{\text{at}}(x), & x = c, \\ f_{\text{right}}(x), & x > c, \end{cases}$$

the left- and right-limits of $f(x)$ at c are the values of the expressions $f_{\text{left}}(c)$ and $f_{\text{right}}(c)$ and are written using limit notation,

$$\lim_{x \to c^-} f(x) = f_{\text{left}}(c),$$
$$\lim_{x \to c^+} f(x) = f_{\text{right}}(c).$$

$\Diamond$

Example 5.2.20 For the piecewise function

$$f(x) = \begin{cases} 3, & x < 0, \\ x^2, & 0 < x \leq 2, \\ 4 - x, & x > 2, \end{cases}$$

evaluate the limits at $x = 0$ and at $x = 2$.

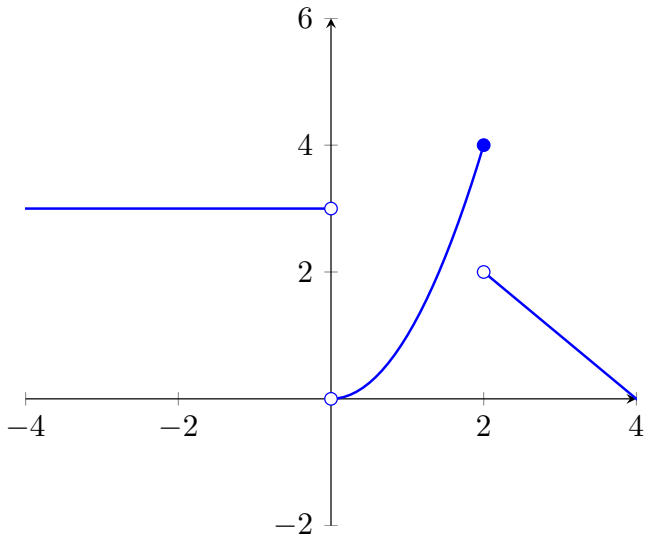
Solution. Around $x = 0$, the function $f(x)$ uses $f(x) = 3$ to the left of $x = 0$ and $f(x) = x^2$ immediately to the right of $x = 0$. Using limit notation, we write

$$\lim_{x \to 0^-} f(x) = 3,$$
$$\lim_{x \to 0^+} f(x) = 0^2 = 0.$$

Around $x = 2$, the function $f(x)$ uses $f(x) = x^2$ to the left and $f(x) = 4 - x$ to the right. For limits, we then have

$$\lim_{x \to 2^-} f(x) = 2^2 = 4,$$
$$\lim_{x \to 2^+} f(x) = 4 - 2 = 2.$$

In the earlier example using this same function, we included a filled circle at the point $(2, 4)$ for the value $f(2) = 2^2 = 4$. When a limit is different from the value of the function (or the function value doesn't exist), we can include an empty circle to show the limit of either the left or the right branch. The limits at $x = 0$ leads to two empty points at $(0, 3)$ (left-limit) and at $(0, 0)$ (right-limit). The limits at $x = 2$ leads to one empty point at $(2, 2)$, since the left-limit matches the value of the function at $(2, 4)$. This improved graph is shown below.



□

At a point away from break points of piecewise functions, the same rule is applied on the left and on the right. Consequently, we can compute left- and right-limits for formulas that are not defined piecewise as well.

Example 5.2.21 Find $\lim_{x\to 1^-}\left[x^2-2x\right]$ and $\lim_{x\to 1^+}\left[x^2-2x\right]$.

Solution. We can think of the expression x^2-2x as a function. The same rule is applied everywhere, so this is equivalent to a piecewise function

$$f(x)=\begin{cases} x^2-2x, & x<1,\\ x^2-2x, & x=1,\\ x^2-2x, & x>1. \end{cases}.$$

(You wouldn't normally write this down—just think it.) Consequently, we have

$$\lim_{x\to 1^-}\left[x^2-2x\right]=1^2-2(1)=-1,$$
$$\lim_{x\to 1^+}\left[x^2-2x\right]=1^2-2(1)=-1.$$

□

Continuity captures the idea of connectedness. The rule for a function to either side of a point should perfectly match up with the rule for the point itself. We express this with limits.

Definition 5.2.22 Continuity at a Point. The statement "the function f is continuous at a point $x=c$" means that the left-limit and right-limit at $x=c$ are equal to the value $f(c)$,

$$\lim_{x\to c^-} f(x)=f(c),$$
$$\lim_{x\to c^+} f(x)=f(c).$$

◊

Note 5.2.23 Our definition for continuity is, at the moment, a bit circular because our intuitive definition of limits (Definition 5.2.19) indicated that we needed "otherwise continuous expressions". We will need a definition for limits that captures the same idea but does not require continuous expressions. We

will then show that every simple algebraic expression is continuous using that new definition.

Example 5.2.24 For the function

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

what values of a and b are needed to make f continuous at $x = 2$?

Solution. The parameters a and b for these formulas set the y-intercepts of the lines, allowing us to slide the lines up or down. We are looking for values that make these lines intersect at the point $(2, 1)$. We use limit notation to create the equations we need to solve.

To make the rule $f(x) = 2x + a$ (to the left of $x = 2$) reach the correct point, we use the left-limit.

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

So that the left branch intersects at the correct point, we need $4 + a = 1$ with $a = -3$.

To make the rule $f(x) = -3x + b$ (to the right of $x = 2$) reach the correct point, we use the right-limit.

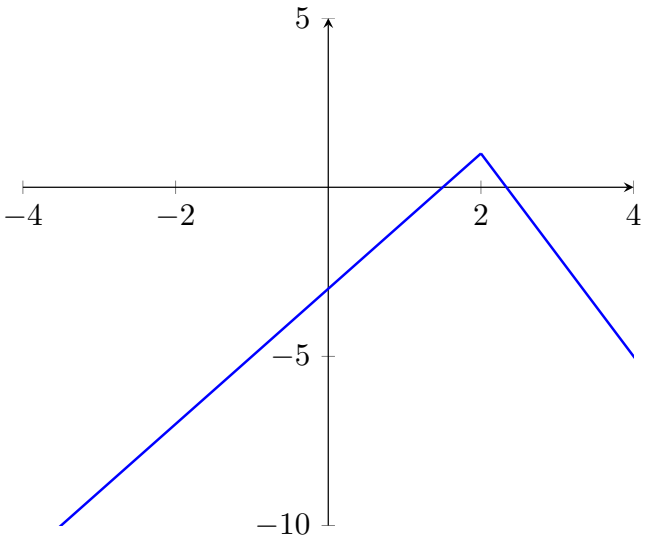
$$\begin{aligned} \lim_{x \to 2^+} f(x) &= \lim_{x \to 2^+} [-3x + b] \\ &= -3(2) + b = -6 + b \end{aligned}$$

We need $-6 + b = 1$ so that $b = 7$.

The function

$$f(x) = \begin{cases} 2x - 3, & x < 2, \\ 1, & x = 2, \\ -3x + 7, & x > 2, \end{cases}$$

is continuous at $x = 2$. A graph of this function is shown below.



□

5.2.4 Summary

- A complete definition of a function must specify the domain, namely the set of all possible inputs to the function. Functions using the same rule on different domains are different functions.
- Sets are often defined as the union of intervals. An open interval (a, b) describes a set that is a solution to $a < x < b$. A closed intervals $[a, b]$ includes the endpoints, $a \leq x \leq b$.
- The natural domain of a function is found by determining the set of inputs for which the function output is defined.
  - A quotient $\frac{u}{w}$ is defined for non-zero denominator $w \neq 0$.
  - An even root (e.g., square root) $\sqrt{u}$ is defined for a non-negative input $u \geq 0$.
  - A logarithm $\log_b u$ is defined for a positive input $u > 0$.
- Inequalities related to zero can be solved by sign analysis: (1) create intervals separated by zeros and discontinuities of the expression, (2) test the sign of the relevant expression on the resulting intervals, and (3) interpret the results.
- An explicit domain for a function can be specified using mapping notation,
$$f : D \to \mathbb{R}; x \mapsto f(x),$$
or using a constraint.
- The range of a one-to-one function is the domain of the inverse function.
- A piecewise function uses different rules for different sets of the domain. The boundaries of these sets are the break-points of the function.
- Continuity of functions is introduced as a concept to guarantee that piecewise functions are connected. At each break-point, we need to verify using limits that the formula to the left and the formula to the right both match the value at the break-point.
- Limit notation indicates that we use a function rule to the left or right of a point and find the value if that rule were extended continuously to the point of interest.
$$\lim_{x \to c^-} f(x) = \text{value from rule on left}$$
$$\lim_{x \to c^+} f(x) = \text{value from rule on right}$$

5.2.5 Exercises

1. Write the function
$$f : [0, 3) \to \mathbb{R}; x \mapsto 2x - 3$$
as an equation with a restriction.

2. Write the function
$$f(x) = 4x + 1, \quad -2 < x \leq 1$$
using mapping notation.

For each of the functions, find the natural domain.

3. $f(x) = \dfrac{3x}{x^2+1}$

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

5. $f(x) = \dfrac{x+2}{x^2-4x-21}$

6. $f(x) = \log_3(x+5)$

7. $f(x) = \log_{10}\left(\dfrac{x+2}{x-1}\right)$

8. $f(x) = \sqrt{x^2-2x-15}$

9. $f(x) = \sqrt[3]{\dfrac{x^2-2x-3}{x^2+2x}}$

10. $f(x) = \sqrt[4]{\dfrac{x^3-8x}{x^2-1}}$

11. Find the range of $f(x) = \dfrac{3}{x+1} - 2$.

12. Find the range of $f(x) = \dfrac{x+3}{x-2} + 4$.

For each function, find the indicated values.

13. $f(x) = \begin{cases} x^2-3x, & x<1, \\ 2, & x=1, \\ 3x-2, & x>1. \end{cases}$

(a) $f(\frac{1}{2})$

(b) $f(1)$

(c) $f(2)$

(d) $\lim\limits_{x\to 1^-} f(x)$

(e) $\lim\limits_{x\to 1^+} f(x)$

(f) $\lim\limits_{x\to 2^-} f(x)$

(g) $\lim\limits_{x\to 2^+} f(x)$

Is f continuous at $x=1$? Is f continuous at $x=2$?

14. $g(x) = \begin{cases} -3x+1, & x<0, \\ 2^x, & 0<x<3, \\ 2x+3, & x\geq 3. \end{cases}$

(a) $g(-2)$

(b) $g(0)$

(c) $g(3)$

(d) $\lim\limits_{x\to 0^-} g(x)$

(e) $\lim\limits_{x\to 0^+} g(x)$

(f) $\lim_{x \to 3^-} g(x)$

(g) $\lim_{x \to 3^+} g(x)$

Is g continuous at $x = 0$? Is g continuous at $x = 3$?

15. Find the value of a so that

$$f(x) = \begin{cases} 2x + 5, & x \leq 3, \\ ax - 4, & x > 3, \end{cases}$$

is continuous at $x = 3$.

16. Find the values of a and b so that

$$f(x) = \begin{cases} 5 - 2x, & x \leq -1, \\ ax + b, & -1 < x < 2, \\ 2x - 5, & x \geq 2, \end{cases}$$

is continuous at $x = -1$ and at $x = 2$.

5.3 Limits of Functions

We have previously studied limits of sequences. In the last section, we considered the continuity of piecewise functions as it depended on whether the function rules to the left and to the right of a point agreed with the value at the point. We used limit notation to describe the values coming from the left and from the right.

In this section, we seek to harmonize these two views of limits. We will introduce the idea that the limit of a function describes the limit of a sequence of output values for a converging sequence of input values. The behavior of limits of sequences justify our rules to calculate limits of functions. We will also discuss horizontal and vertical asymptotes of functions in the context of limits.

5.3.1 Limits

For sequences, we introduced the idea of limits as the value the sequence was approaching further and further in that sequence. We saw that the decimal approximations of the sequence values would eventually converge to the decimal approximation of the limiting value. The mathematical definition of the limit was stated in terms of the possibility of eventually waiting long enough in the sequence that the sequence values would approximate the limit value within *any* desired accuracy of approximation.

The only limits of interest in sequences were when the index went to infinity. For functions, in order to understand continuity 5.2.19, we have found that we also need to think about limits as the independent variable approaches a value from either the left or the right. Sequences can give us a way to think about this possibility.

Definition 5.3.1 Limits of Function. For a function f defined on intervals to the left and right of c, we say

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

to mean that for *every* independent sequence x such that $x_n \neq c$ and $x_n \to c$, the dependent sequence $y = (f(x_n))_{n=n_0}^{\infty}$ must have the limit L,

$$\lim_{n \to \infty} f(x_n).$$

One-sided limits add constraints to the independent sequences, with $x \to c^+$ requiring $x_n > c$ and $x \to c^-$ requiring $x_n < c$. ◊

Function limits are properties of the function itself and do not depend on the sequences chosen. If different independent sequences that converge to c result in different limits for the dependent sequence, then the function does not have a limit. The following example illustrates how this new definition relates our earlier concept of continuity of piecewise functions with the sequence definition of function limits. We create a table of sequence values, one column corresponding to the independent variable (input) and another column corresponding to the dependent variable (output). The input sequence is chosen to converge to the value c, and we examine what happens to the sequence of the dependent variable.

Example 5.3.2 A function is defined piecewise as

$$f(x) = \begin{cases} x+2, & x < 3, \\ 4x - x^2, & x > 3. \end{cases}$$

Find $\lim_{x\to 3^-} f(x)$, $\lim_{x\to 3^+} f(x)$, and $\lim_{x\to 3} f(x)$ using sequences for approximation.

Solution. The left-sided limit needs to consider an independent sequence $x_n < 3$ with $x_n \to 3$. The following partial table illustrates an example with $x_n = 3 - 10^{-n}$.

n	x_n	$f(x_n)$
1	2.9	$f(2.9) = 2.9 + 2 = 4.9$
2	2.99	$f(2.99) = 2.99 + 2 = 4.99$
3	2.999	$f(2.999) = 2.999 + 2 = 4.999$

If we compare the values of the dependent sequence $f(x_n)$ with the value of the formula $x+2$ evaluated at $x = 3$, which is $x + 2 = 5$, we can see that the dependent sequence is approaching that limit $f(x_n) \to 5$. We therefore write

$$\lim_{x\to 3^-} f(x) = 5.$$

In a similar way, a right-sided limit requires $x_n > 3$ with $x_n \to 3$, such as the sequence $x_n = 3 + 10^{-n}$.

n	x_n	$f(x_n)$
1	3.1	$f(3.1) = 4(3.1) - (3.1)^2 = 2.79$
2	3.01	$f(3.01) = 4(3.01) - (3.01)^2 = 2.9799$
3	3.001	$f(3.001) = 4(3.001) - (3.001)^2 = 2.997999$

If we compare the values of the dependent sequence $f(x_n)$ with the value of the formula $4x - x^2$ evaluated at $x = 3$, which is $4x - x^2 = 3$, we can see that the dependent sequence is approaching that limit $f(x_n) \to 3$. We therefore write

$$\lim_{x\to 3^+} f(x) = 3.$$

The two-sided limit requires only $x_n \to 3$. The sequence values might be either above or below 3. Above, we found that when the independent variable values are on the left $x_n < 3$, we had $f(x_n) \to 5$. But when $x_n > 3$, we had $f(x_n) \to 3$. Because different sequences with $x_n \to 3$ result in different limits for $f(x_n)$,

$$\lim_{x\to 3} f(x) \text{ does not exist.}$$

□

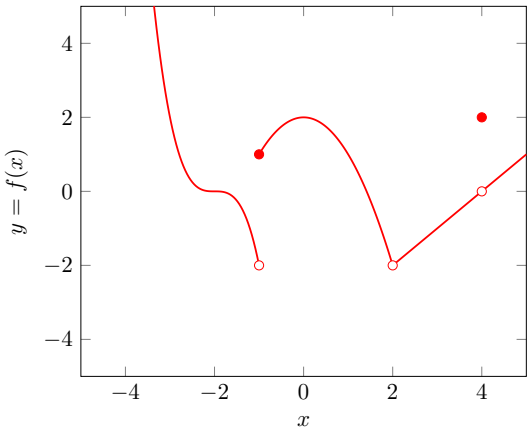
While the previous example attempted to connect our simpler understanding of limits with the limits of sequences, the given solution really only illustrated the first few terms from two out of infinitely many possible independent sequences. Using tables of sequence values might suggest possible values of the limits, but we need a more definitive reason that the limit agrees with simple evaluation of the formula.

When a function is visualized as a graph, a limit can be determined by looking at the branches of the graph immediately to the left or right of the point of interest. A limit of $f(x)$ with $x \to c^-$ means to look at the branch of the function with $x < c$ and identify what point that branch would lead to as $x \to c$. Similarly, a limit of $f(x)$ with $x \to c^+$ means to look at the branch of the function with $x > c$ and identify what point that branch would lead to as

$x \to c$.

Example 5.3.3 Consider the function f whose graph is shown below. Find the following limits based on the graph, assuming the coordinates of shown points are integers.

1. $\lim_{x \to -1} f(x)$
2. $\lim_{x \to 0} f(x)$
3. $\lim_{x \to 2} f(x)$
4. $\lim_{x \to 4} f(x)$



Solution.

1. $\lim_{x \to -1} f(x)$

   We consider a sequence for x on the x-axis that converges to -1. It is best to consider a sequence on the left and another on the right. For $x_n < -1$, our function will be using the cubic portion of the graph. As $x_n \to -1$ (from the left), the function will move closer and closer to the open point at $(-1, -2)$. The y-value of this point is the corresponding limit of the dependent sequence:

   $$\lim_{x \to -1^-} f(x) = -2.$$

   For $x_n > -1$, we will be somewhere to the right. As $x_n \to -1$ (from the right), we will eventually be on the portion of the function corresponding to the concave down parabola. The sequence will move us closer and closer to the filled-in point at $(-1, 1)$:

   $$\lim_{x \to -1^+} f(x) = 1.$$

   Because the left- and right-limits have different values, the two-sided limit *does not exist*.

2. $\lim_{x \to 0} f(x)$

   We consider a sequence for x on the x-axis that converges to 0. Regardless of whether the sequence is to the left or the right of $x = 0$, the value x_n will eventually use the function defined by the concave down parabola.

The point on the graph of the function will converge to the vertex of this parabola at $(0, 2)$. This is our limit.

$$\lim_{x \to 0} f(x) = 2.$$

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

   For a sequence $x_n \to 2$ with $x_n < 2$, the point on the graph will eventually be on the parabola and approaching the open point at $(2, -2)$. For values with $x_n > 2$, the point on the graph will eventually be on the line and also approaching the point $(2, -2)$. Because the sequence always results in approaching the point $(2, -2)$, we have a limit

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

   Notice that a limit does not depend on whether the point is included in the function or not. All that matters is whether the sequence of points converges to that point.

4. $\lim_{x \to 4} f(x)$

   For a sequence $x_n \to 4$ and $x_n \neq 4$, eventually the function will be on the line to the left or right of $x = 4$. Either way, the corresponding point on the graph will be converging to $(4, 0)$:

   $$\lim_{x \to 4} f(x) = 0.$$

   The value of the function $f(4) = 2$ has no effect on the limit.

□

When we have a formula for a function, we already know that a table can be helpful but will not guarantee the value of the limit. We might think we can graph the function, but even our graphs of functions are ultimately based on a table of values. We need some methods to evaluate limits based on the formulas alone.

Our earlier motivation for limits involving piecewise functions suggest that we can find a limit by evaluating a function at the point of interest. In most cases, this is true. This is a consequence of the algebraic structure of expressions and the fact that limits behave very nicely with algebraic operations. The rest of this section explains why limits usually behave so well.

5.3.2 Limit Rules for Combining Sequences

We start by formalizing some rules about how sequence limits relate to the arithmetic of sequences. These rules are stated as theorems. We begin each theorem with one or more sequence that is given with a particular limit. We then define a new sequence using arithmetic involving those sequences. The conclusion of each theorem describes the limit of the new sequence. To apply a theorem, we must verify that the hypotheses are satisfied before we can use the conclusion.

For a converging sequence $x = (x_n)$ with $x_n \to L$, we can think of the sequence values as *approximating* L. The absolute error of approximation $|x_n - L|$ must vanish as $n \to \infty$. In other words, for any margin of error $\epsilon > 0$, there must be some index N so that $|x_n - L| < \epsilon$ once $n > N$. The proofs of these theorems rely on showing how the error of approximation for the arithmetic combination of the sequences can be related to the errors of

approximations of the given sequences to their limits in a way to show that the approximation error will eventually vanish.

The first three rules involve elementary operations involving constants on a single sequence. They correspond to the operations used to construct expressions involving another expression and a constant, as discussed in (((Unresolved xref, reference "subsubsection-elementary-arithmetic-operations"; check spelling or use "provisional" attribute))) .

Theorem 5.3.4 Sequence Limit of a Constant Sum (SL:CS). *Given a sequence $u = (u_n)$ with $u_n \to L$ and any constant k, the transformed sequence $w_n = u_n + k$ has limit*

$$\lim_{n\to\infty} u_n + k = L + k.$$

Proof. The error of approximation for w_n from its proposed limit $L + k$ can be rewritten

$$|w_n - (L + k)| = |u_n + k - L - k| = |u_n - L|.$$

This is the same as the error of approximation for u_n from its limit L. As soon as $|u_n - L| < \epsilon$, we also have $|w_n - (L + k)| < \epsilon$. Because $|u_n - L| \to 0$, this proves

$$\lim_{n\to\infty} u_n + k = L + k.$$

■

Theorem 5.3.5 Sequence Limit of a Constant Multiple (SL:CM). *Given a sequence $u = (u_n)$ with $u_n \to L$ and any constant k, the transformed sequence $w_n = k \cdot u_n$ has limit*

$$\lim_{n\to\infty} k\,u_n = k\,L.$$

Proof. The error of approximation for w_n from its proposed limit kL can be rewritten

$$|w_n - kL| = |ku_n - kL| = |k(u_n - L)| = |k| \cdot |u_n - L|.$$

If $k = 0$, then $w_n = 0$ for all n and $w_n \to 0$ must be true. If $k \neq 0$, then the error of approximation for w from its proposed limit is exactly $|k|$ times the error of approximation for u from its given limit. As soon as $|u_n - L| < \frac{1}{|k|}\epsilon$, we must have $|w_n - kL| < \epsilon$. Because $|u_n - L| \to 0$, this proves

$$\lim_{n\to\infty} k\,u_n = k\,L.$$

■

Theorem 5.3.6 Sequence Limit of a Reciprocal (SL:MInv). *Given a sequence $u = (u_n)$ with $u_n \to L \neq 0$, the transformed sequence of multiplicative inverses $w_n = \frac{1}{u_n}$ has limit*

$$\lim_{n\to\infty} \frac{1}{u_n} = \frac{1}{L}.$$

Proof. Division is not defined when the denominator equals zero. Because $u_n \to L$ and $L \neq 0$, we know that $|u_n - L| < \frac{1}{2}|L|$ eventually. When $L > 0$, this means that $\frac{1}{2}L < u_n < \frac{3}{2}L$. If $L < 0$, then $\frac{3}{2}L < u_n < \frac{1}{2}L$. Either way, u_n is kept away from 0 and $w_n = \frac{1}{u_n}$ is guaranteed to be defined. (Before this point, we might have had $u_n = 0$ so that w_n is not defined.

The error of approximation for w_n from its proposed limit $\frac{1}{L}$ can be rewritten using a common denominator

$$|w_n - \frac{1}{L}| = |\frac{1}{u_n} - \frac{1}{L}| = |\frac{L - u_n}{u_n L}| = |u_n - L| \cdot \frac{1}{|u_n||L|}.$$

Because $|u_n| > \frac{1}{2}|L|$, we know that $\dfrac{1}{|u_n|} < \dfrac{2}{|L|}$. Thus,

$$|w_n - \frac{1}{L}| < |u_n - L| \cdot \frac{2}{|L|^2}.$$

The error of approximation for w_n is always smaller than $\frac{2}{|L|^2}$ times the error of approximation for u_n from its limit. Because $|u_n - L| \to 0$, this proves

$$\lim_{n\to\infty} \frac{1}{u_n} = \frac{1}{L}.$$

■

The second group of limit rules of combination allow us to take two limits that we know and combine them with arithmetic. Notice how the limit rules correspond exactly with the arithmetic operations used to construct expression, as discussed in Subsection 2.2.2. The proofs of these theorems are more advanced and will not be given in this section.

Theorem 5.3.7 Sequence Limit of a Sum (SLC:Sum). *Given sequences $u = (u_n)$ with $u_n \to L$ and $v = (v_n)$ with $v_n \to M$, the sequence defined by the sum $w_n = u_n + v_n$ has limit*

$$\lim_{n\to\infty} [u_n + v_n] = L + M.$$

Theorem 5.3.8 Sequence Limit of a Difference (SLC:Diff). *Given sequences $u = (u_n)$ with $u_n \to L$ and $v = (v_n)$ with $v_n \to M$, the sequence defined by the difference $w_n = u_n - v_n$ has limit*

$$\lim_{n\to\infty} [u_n - v_n] = L - M.$$

Theorem 5.3.9 Sequence Limit of a Product (SLC:Prod). *Given sequences $u = (u_n)$ with $u_n \to L$ and $v = (v_n)$ with $v_n \to M$, the sequence defined by the product $w_n = u_n \cdot v_n$ has limit*

$$\lim_{n\to\infty} [u_n \cdot v_n] = L \cdot M.$$

Theorem 5.3.10 Sequence Limit of a Quotient (SLC:Quot). *Given sequences $u = (u_n)$ with $u_n \to L$ and $v = (v_n)$ with $v_n \to M$ and $M \neq 0$, the sequence defined by the sum $w_n = \dfrac{u_n}{v_n}$ has limit*

$$\lim_{n\to\infty} \frac{u_n}{v_n} = \frac{L}{M}.$$

In addition to algebraic operations combining sequences, we have operations associated with functions. This includes raising sequences to powers, applying exponential or logarithm functions, or using trigonometric functions. We will learn that each of these functions are continuous. Consequently, the following theorem will apply.

Theorem 5.3.11 Sequence Limit of a Continuous Function (SLC:CFxn). *Given a sequence $u = (u_n)$ with $u_n \to L$ and a function f that is continuous at L, the sequence defined by the sum $w_n = f(u_n)$ has limit*

$$\lim_{n\to\infty} f(u_n) = f(L).$$

5.3.3 Elementary Limit Rules for Functions

Having established the limit rules associated with sequences, we can apply those rules to create corresponding limit rules for functions.

The first collection of limit rules are some basic limits. We can think of them as our building blocks for more complicated limits. We begin by showing that constant functions and the identity functions are continuous.

Theorem 5.3.12 Limit of a Constant (LE:Const).

Hypothesis *k is a real number.*

Conclusion $\lim_{x\to a} k = k$.

Proof. For a constant function $f(x) = k$, the output sequence is a constant sequence regardless of the input sequence. ■

Theorem 5.3.13 Limit of the Identity (LE:Ident).

Hypothesis *none*

Conclusion $\lim_{x\to a} x = a$.

Proof. For identity function $f(x) = x$, the output sequence is will be the same as the input sequence. Since $x_k \to a$, the output sequence has limit a. ■

We include the limit of linear functions in our known limits of elementary functions.

Theorem 5.3.14 Limit of a Linear Function (LE:Line).

Hypothesis *m and b are real numbers.*

Conclusion $\lim_{x\to a}[mx + b] = ma + b$.

Proof. Given a sequence x_k with $x_k \neq a$ and $x_k \to a$, define the output sequence $y_k = mx_k + b$. This is a constant sum and constant multiple of x_k. By SLC:CM, we know $mx_k \to ma$. By SLC:CS, we then have $y_k = mx_k + b \to ma + b$. ■

5.3.4 Limit Rules of Combination

The second collection of limit rules tell us how we can take limits that we already know (starting with building blocks) and use them to compute more complicated limits. The first three rules take a single limit that is known to be valid and use arithmetic with a constant to find a new limit. Each of the theorem simply applies the corresponding limit rule for sequences on the sequence created by the function.

Theorem 5.3.15 Limit of a Constant Sum (LC:CS).

Hypothesis $\lim_{x\to a} f(x) = L$ *and k is a real number.*

Conclusion $\lim_{x\to a}[f(x) + k] = L + k$.

Theorem 5.3.16 Limit of a Constant Multiple (LC:CM).

Hypothesis $\lim_{x\to a} f(x) = L$ *and k is a real number.*

Conclusion $\lim_{x\to a}[k \cdot f(x)] = k \cdot L$.

Theorem 5.3.17 Limit of a Reciprocal or Multiplicative Inverse (LC:MInv).

Hypothesis: $\lim_{x\to a} f(x) = L$ *and* $L \neq 0$.

Conclusion: $\lim_{x\to a} \dfrac{1}{f(x)} = \dfrac{1}{L}$.

The next limit rules of combination allow us to take two limits that we know and combine them with arithmetic. In each of the cases, note that both limits in the hypothesis have $x \to a$ (i.e., x approaches the same value in both limits).

Theorem 5.3.18 Limit of a Sum (LC:Sum).

Hypothesis $\lim_{x\to a} f(x) = L$ *and* $\lim_{x\to a} g(x) = M$.

Conclusion $\lim_{x\to a}[f(x) + g(x)] = L + M$.

Theorem 5.3.19 Limit of a Difference (LC:Diff).

Hypothesis $\lim_{x\to a} f(x) = L$ *and* $\lim_{x\to a} g(x) = M$.

Conclusion $\lim_{x\to a}[f(x) - g(x)] = L - M$.

Theorem 5.3.20 Limit of a Product (LC:Prod).

Hypothesis $\lim_{x\to a} f(x) = L$ *and* $\lim_{x\to a} g(x) = M$.

Conclusion $\lim_{x\to a}[f(x) \cdot g(x)] = L \cdot M$.

Theorem 5.3.21 Limit of a Quotient (LC:Quot).

Hypothesis $\lim_{x\to a} f(x) = L$ *and* $\lim_{x\to a} g(x) = M$ *and* $M \neq 0$.

Conclusion $\lim_{x\to a} \dfrac{f(x)}{g(x)} = \dfrac{L}{M}$.

In addition to the arithmetic of functions, composition of functions plays an important role in algebra and calculus. So we need a limit rule associated with composition. Recall that for composition, the output of one function becomes the input to another function.

Theorem 5.3.22 Limit of a Continuous Composition (LC:Comp).

Hypothesis $\lim_{x\to a} f(x) = L$ *and* g *is continuous at* L*, or in other words,* $\lim_{u\to L} g(u) = g(L)$.

Conclusion $\lim_{x\to a} g \circ f(x) = \lim_{x\to a} g(f(x)) = g(L)$.

If f is continuous at a so that $\lim_{x\to a} f(x) = f(a)$ then we have $\lim_{x\to a} g \circ f(x) = g \circ f(a)$. In other words, the composition of continuous functions is a continuous function.

5.3.5 Justifying Limit Calculations

There are two ways in which limit rules are applied. One way is to provide formal justification of limit calculations, or in other words, to write a proof of limit statements. The other way limit rules are used is to break a computation down into recognizable and manageable parts. This section focuses on the process of formal justification.

A mathematical proof is essentially a sequence of statements, each of which is demonstrably true based only on previously stated knowledge and logical arguments. This means that when writing a proof or any careful justification, we must be careful that when we write something down we have previously

established all of the necessary conditions at a previous step. In order to avoid circular reasoning, we should avoid referring to something as true before we actually show it is true.

For justification of limit statements, this means that we start from the small building blocks that create our formula and put them together one step at a time until we can justify the limit statement we are trying to prove.

Example 5.3.23 Compute and justify $\lim_{x\to 2} 3x^2(2x-5)$.

Solution. Start by planning ahead. The formula $3x^2(2x-5)$ is a product of $3x^2$ and $2x-5$. This second factor $2x-5$ is a linear function so there is a limit rule for that piece. But $3x^2$ is not linear, and we recognize it as a product of 3 (a constant) and x^2. Finally, we see that $x^2 = x \cdot x$ is the product of the identity with itself. We will start with the elementary formulas and build them back up to the full function.

1. $\lim_{x\to 2} x = 2$ by LE:Ident.
2. $\lim_{x\to 2} x \cdot x = 2 \cdot 2 = 4$ by LC:Prod using limits of $f(x) = x$ (step 1) and $g(x) = x$ (step 1).
3. $\lim_{x\to 2} 3x^2 = 3 \cdot 4 = 12$ by LC:CM using constant $k = 3$ and limit of $f(x) = x^2$ (step 2).
4. $\lim_{x\to 2} 2x - 5 = 2(2) - 5 = -1$ by LE:Line ($m = 2$, $b = -5$).
5. $\lim_{x\to 2} 3x^2(2x-5) = 12(-1) = -12$ by LC:Prod using the limits found in step 3 and step 4.

□

Example 5.3.24 Compute and justify $\lim_{x\to 3} x^3 + 4x^2 - 3x + 1$.

Solution. It is important to note that limit rules of combination only combine two formulas at a time. In this calculation, we will need the limit of x^3. Writing this as $x^3 = x \cdot x \cdot x$ is not going to be as useful as writing $x^3 = x \cdot x^2$ because there are no rules to combine three limits at once. In addition, subtraction is always problematic, so it is best to rewrite subtraction as a sum,

$$x^3 + 4x^2 - 3x + 1 = x^3 + 4x^2 + -3x + 1.$$

1. $\lim_{x\to 3} x = 3$ by LE:Ident.
2. $\lim_{x\to 3} x \cdot x = 3 \cdot 3 = 9$ by LC:Prod using the limits in step 1 (twice).
3. $\lim_{x\to 3} x \cdot x^2 = 3(9) = 27$ by LC:Prod using the limits in step 1 and step 2.
4. $\lim_{x\to 3} 4x^2 = 4(9) = 36$ by LC:CM using $k = 4$ and the limit in step 2.
5. $\lim_{x\to 3} -3x + 1 = -3(3) + 1 = -8$ by LE:Line ($m = -3$, $b = 1$).
6. $\lim_{x\to 3} x^3 + 4x^2 = 27 + 36 = 63$ by LC:Sum using limits in step 3 and step 4.
7. $\lim_{x\to 3} x^3 + 4x^2 + -3x + 1 = 63 + -8$ by LC:Sum using limits in step 6 and step 5.

□

Because most expressions that we work with are defined strictly in terms of the basic arithmetic operations and elementary functions, the limit rules we have developed essentially allow us to replace the independent variable in the formula $f(x)$ with the limiting point $x \to c$. That is, whenever the expression involves basic arithmetic operations (addition, subtraction, multiplication, and division), we know that we *could* apply the limit rules step-by-step to justify

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

The exception is that our rule for quotients does not allow division by zero.

Theorem 5.3.25 *If $f(x)$ is an algebraic expression that involves only arithmetic operations, then $\lim_{x \to c} f(x) = f(c)$ so long as $f(c)$ is defined.*

Example 5.3.26 Determine $\lim_{x \to 2} \dfrac{2x+3}{x^2-5}$.

Solution. At first glance, we might worry that the theorem does not apply because x^2 is a power and not an arithmetic operation. However, because $x^2 = x \cdot x$ is a product, we have a function $f(x) = \dfrac{2x+3}{x \cdot x - 5}$ defined in terms of arithmetic operations. We evaluate $f(2)$:

$$f(2) = \frac{2(2)+3}{2^2-5} = -7.$$

Consequently, by Theorem 5.3.25, we have

$$\lim_{x \to 2} \frac{2x+3}{x^2-5} = -7.$$

□

We will learn in the next section how to deal with expressions where the value is not defined.

5.3.6 Summary

- The limit of a function $\lim_{x \to c} f(x)$ represents the value L that is the limit of the dependent sequence $f(x_n)$ for *every* independent sequence (x_n) that satisfies $x_n \neq c$ and $x_n \to c$. One-sided limits add the constraint that the sequence must stay below c ($x \to c^-$) or above c ($x \to c^+$).
- Numerically, a function limit $\lim_{x \to c} f(x)$ can be approximated by testing the value of the function for values of the independent variable following a sequence $x \to c$.
- Graphically, a function limit $\lim_{x \to c} f(x)$ corresponds to the y-value of the point in the plane that the sequence of points $(x_n, f(x_n))$ approaches from the left and from the right as $x_n \to c$. If the two branches (left vs right) approach different points, the two-sided limit does not exist.
- Limit rules associated with all of the arithmetic operations justify applying the same operations with limits.
- An argument justifying limits using limit rules must demonstrate that the component limits are known prior to combining them with a limit rule.

5.3.7 Exercises

Use appropriate tables to approximate the following function limits. For a two-sided limit, be sure that your work verifies that both sides approximate the same value

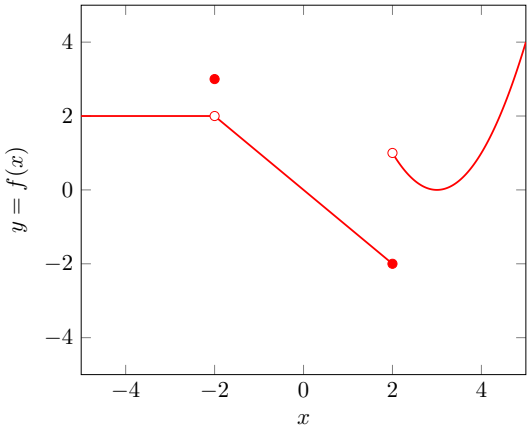
1. $\lim_{x\to 3^-} \dfrac{2^x - 8}{x - 3}$

2. $\lim_{x\to 3^+} \dfrac{2^x - 8}{x - 3}$

3. $\lim_{x\to 2} \dfrac{x^2 - 4}{2^x - 4}$

4. $\lim_{x\to 1} \dfrac{x^2 - 1}{|x - 1|}$

Consider the function f whose graph is shown below. Find the following values, if they exist, based on the graph and assuming the coordinates of shown points are integers.



5.

(a) $f(-2)$

(b) $\lim_{x\to -2^-} f(x)$

(c) $\lim_{x\to -2^+} f(x)$

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

6.

(a) $f(0)$

(b) $\lim_{x\to 0^-} f(x)$

(c) $\lim_{x\to 0^+} f(x)$

(d) $\lim_{x\to 0} f(x)$

7.

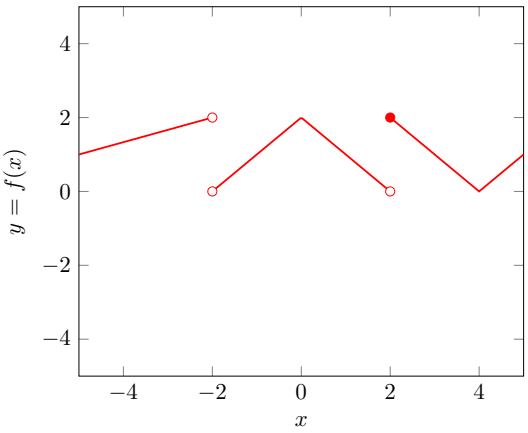
(a) $f(2)$

(b) $\lim_{x\to 2^-} f(x)$

(c) $\lim_{x \to 2^+} f(x)$

(d) $\lim_{x \to 2} f(x)$

Consider the function f whose graph is shown below. Find the following values, if they exist, based on the graph and assuming the coordinates of shown points are integers.



8.

(a) $f(-2)$

(b) $\lim_{x \to -2^-} f(x)$

(c) $\lim_{x \to -2^+} f(x)$

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

9.

(a) $f(1)$

(b) $\lim_{x \to 1^-} f(x)$

(c) $\lim_{x \to 1^+} f(x)$

(d) $\lim_{x \to 1} f(x)$

10.

(a) $f(2)$

(b) $\lim_{x \to 2^-} f(x)$

(c) $\lim_{x \to 2^+} f(x)$

(d) $\lim_{x \to 2} f(x)$

Compute and justify the value of each limit applying the limit rules for functions step-by-step.

11. $\lim_{x \to -3} \dfrac{4x+1}{2x+3}$

12. $\lim_{x \to 2} 3x^2 - 4x + 5$

13. $\lim_{x \to 4} \dfrac{5x^2}{2x^2(3x - 1)}$

14. $\lim_{x \to -2} x^3 - 4x^2 + 5x - 7$

5.4 Continuity of Functions

5.4.1 Overview

The elementary limit rules for functions tell us that the limit of an algebraic expression made from arithmetic operations will equal the value of the expression at the point in question, if that value exists. So why bother introducing limits at all if they are the same as function evaluation?

The fact of the matter is, they aren't the same thing at all. Recall that for piecewise functions, we can use limits to find the limiting value of a function to the left and to the right of a break point. Function evaluation would only allow us to look at the point itself. Having a function value agree with the limits is a characteristic of a function being continuous. A value for x where a function is not defined is an example of a discontinuity.

In this section, we consider the continuity of functions. We learn about removable and infinite discontinuities, which correspond to holes and vertical asymptotes in a graph. We learn to compute limits of functions at these discontinuities by looking at simplified, factored expressions. Sign analysis is used for infinite discontinuities to determine whether the discontinuity corresponds to unbounded positive or negative values.

5.4.2 Removable and Infinite Discontinuities

The intuitive idea of a continuous function is a function whose graph is connected. Sometimes, this is thought of as being able to draw the graph without lifting the pen. The technical definition of a continuity at a point, say at $x = c$, involves three parts. First, the limit on the left exists. This means that we can trace the graph on a branch with $x < c$. Second, the limit on the right exists. This means that we also can trace the graph on a branch with $x > c$. Third, both limits are equal to $f(c)$. This gives us the connection from the left branch to the right branch through the point.

Any time a function has a break, it has a discontinuity at that location. A break can be a simple hole, a jump between values, or an infinite discontinuity associated with a vertical asymptote. Discontinuities might also occur due to limits themselves not existing for any reason.

Consider two functions, $f(x) = \dfrac{1}{(x-3)(x+2)}$ and $g(x) = \dfrac{x^2-4x+3}{x-3}$. In both functions, the value of the function is not defined at $x = 3$; f and g are both **discontinuous** at $x = 3$. Consequently, the corresponding limits $\lim\limits_{x\to 3} f(x)$ and $\lim\limits_{x\to 3} g(x)$ can not be computed directly using the limit rules for functions.

If we look at the graphs of our functions, as shown below, we see that there is something fundamentally different about the behavior around $x = 3$. The function $f(x) = \dfrac{1}{(x-3)(x+2)}$ appears to have a vertical asymptote at $x = 3$. The function $g(x) = \dfrac{x^2-4x+3}{x-3}$ looks continuous, even though we know it has a break at $x = 3$.

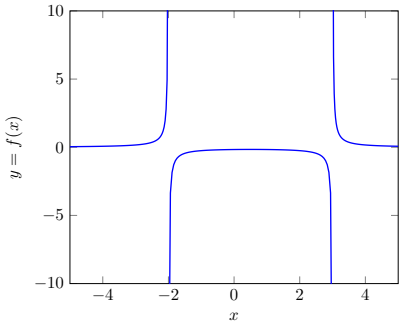


Figure 5.4.1 $y = \dfrac{1}{(x-3)(x+2)}$

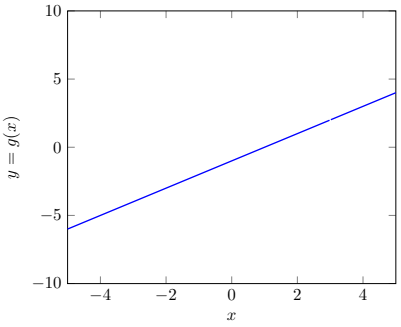


Figure 5.4.2 $y = \dfrac{x^2-4x+3}{x-3}$

If we factor the formula for $g(x)$, we discover that the formula simplifies.

$$\begin{aligned} g(x) &= \frac{x^2-4x+3}{x-3} \\ &= \frac{(x-3)(x-1)}{x-3} \\ &= x-1, \quad x \neq 3 \end{aligned}$$

Notice that we must include a domain restriction when we simplify. The original function is not defined for $x = 3$, but the simplified version is. To ensure the functions are the same, they must have the same domain. Because $x-1$ is *continuous* at $x = 3$, $g(x)$ has a **hole** at $x = 3$ and we call this a **removable discontinuity**. A vertical asymptote at a point corresponds to a **infinite discontinuity**.

Example 5.4.3 The function $f(x) = \dfrac{3x^2-x-2}{x-1}$ has a removable discontinuity at $x = 1$. What is the continuous function equivalent to $f(x)$?

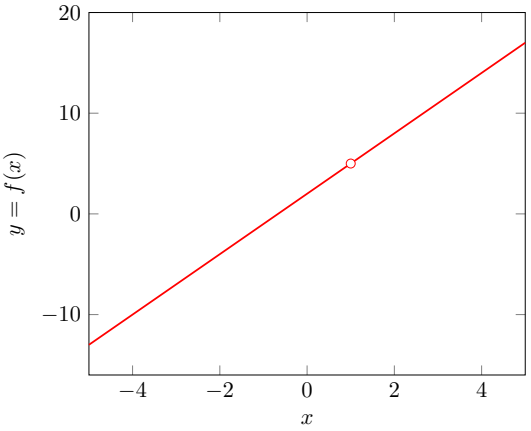
Solution. A polynomial, like $3x^2 - x - 2$, will have a factor of $x - 1$ if and only if that polynomial has a value of 0 when $x = 1$. So we can see if it will cancel a factor by checking $3(1)^2 - (1) - 2 = 0$. Knowing this factor, we can soon find $3x^2 - x - 2 = (x-1)(3x+2)$. For all $x \neq 1$, we have

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

We can only say this for $x \neq 1$ since the domain of f is $(-\infty, 1) \cup (1, \infty)$. That is,

$$f(x) = 3x+2, \quad x \neq 1.$$

Our function $f(x)$ has the same graph as $y = 3x + 2$ except it has a hole at $x = 1$.



□

The previous example illustrates a basic feature of **rational functions** (i.e., a ratio or quotient of two polynomials). That is that there will be a canceling factor if the numerator and denominator have a common zero.

Theorem 5.4.4 *A rational function* $f(x) = \dfrac{p(x)}{q(x)}$ *where* p *and* q *are polynomial functions has a domain defined by*

$$D = \{x : q(x) \neq 0\}.$$

Further, p *and* q *will have canceling common factors of the form* $(x-a)$ *where* a *is a constant if and only if* $p(a) = 0$ *and* $q(a) = 0$.

For rational functions, the only possible discontinuities are holes and infinite discontinuities at vertical asymptotes. Holes correspond to points that are not in the domain but can be removed by canceling common factors. Any other points of discontinuity must be vertical asymptotes.

Example 5.4.5 Describe the discontinuities of the function

$$f(x) = \frac{x^3 - 5x^2 + 6x}{x^2 + x - 6}.$$

Solution. The discontinuities are determined for a rational function by finding the zeros of the polynomial in the denominator, $q(x) = x^2 + x - 6$. We solve this by factoring:

$$q(x) = (x+3)(x-2).$$

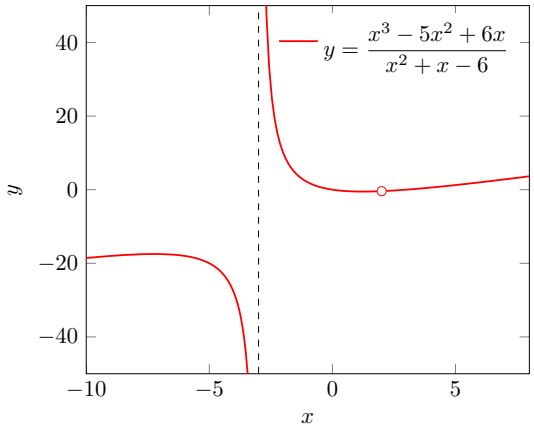
There are discontinuities (breaks in the graph) at $x = -3$ and at $x = 2$.

We determine the type of discontinuity by seeing if common factors cancel. The numerator $p(x) = x^3 - 5x^2 + 6x$ can be tested even before factoring. At $x = -3$, we have $p(-3) = -27 - 5(9) + 6(-3) = -90$ so that $x+3$ is not going to be a common factor. There must be a vertical asymptote at $x = -3$. At $x = 2$, we have $p(2) = 8 - 5(4) + 6(2) = 0$ so that there will be a common factor that cancels.

$$\begin{aligned} f(x) &= \frac{x^3 - 5x^2 + 6x}{x^2 + x - 6} \\ &= \frac{x(x^2 - 5x + 6)}{(x+3)(x-2)} \\ &= \frac{x(x-2)(x-3)}{(x+3)(x-2)} \end{aligned}$$

$$= \frac{x(x-3)}{x+3}, \quad x \neq 2.$$

Because the new formula has a natural domain $x \neq -3$, the discontinuity at $x = 2$ was removable. The graph has a hole at $x = 2$ and a vertical asymptote at $x = -3$. (Notice the addition of an explicit domain on the last step when we canceled, corresponding to the hole.)



□

5.4.3 Limits at Discontinuities

The limit rules do not apply when substitution would result in division by zero. These precisely occur at points of discontinuity. Suppose a rational function $f(x) = \frac{p(x)}{q(x)}$ has $p(c) = 0$ and $q(c) = 0$. Immediate substitution of $x = c$ into $f(x)$ would result in $\frac{0}{0}$, which we have earlier identified as an **indeterminate limit form**. Because $p(c) = 0$ and $q(c) = 0$, $p(x)$ and $q(x)$ have a common factor $x - c$. Cancellation of that factor gives $f(x)$ a simplified form, and we can try again to evaluate the limit.

Example 5.4.6 Evaluate $\lim_{x\to 2} \frac{x^2 - 5x + 6}{x^2 - 4}$.

Solution. The formula is defined in terms of elementary arithmetic, so we try to evaluate the expression by substituting $x = 2$.

$$\lim_{x\to 2} \frac{x^2 - 5x + 6}{x^2 - 4} \stackrel{?}{=} \frac{2^2 - 5(2) + 6}{2^2 - 4} = \frac{0}{0}$$

The limit has an indeterminate form. We can factor $x - 2$ from numerator and denominator and rewrite the expression.

$$\begin{aligned} f(x) &= \frac{x^2 - 5x + 6}{x^2 - 4} \\ &= \frac{(x-2)(x-3)}{(x-2)(x+2)} \\ &= \frac{x-3}{x+2}, \quad x \neq 2 \end{aligned}$$

Limits use the function to the side of the point in question. In this case, $f(x)$ uses the same formula on the left and the right of the discontinuity. Because the new formula is continuous, we can use substitution.

$$\lim_{x\to 2} \frac{x^2 - 5x + 6}{x^2 - 4} = \lim_{x\to 2} \frac{x-3}{x+2}$$

$$= \frac{2-3}{2+2} = -\frac{1}{4}$$

□

When a limit has a form $\frac{0}{0}$, we know to rewrite the formula in a simplified form. For sequences, we learned that if $a_n > 0$ and $a_n \to 0$, then $\frac{1}{a_n} \to +\infty$. That is, the reciprocal of a small positive number will be a large positive number. The smaller a_n becomes, the larger $\frac{1}{a_n}$ will be. Consequently, a rational function with a limit of the form $\frac{L}{0}$ has a vertical asymptote, and the limit will be unbounded. We use sign analysis to determine if the left- and right-limits are $+\infty$ or $-\infty$.

Example 5.4.7 Evaluate $\lim_{x\to -2} \dfrac{x^2-5x+6}{x^2-4}$.

Solution. In the example above, we already found

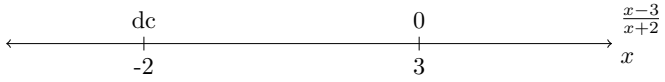
$$f(x) = \frac{x^2-5x+6}{x^2-4} = \frac{x-3}{x+2}, \quad x \neq 2.$$

Attempting substitution, we find

$$\lim_{x\to -2} f(x) \stackrel{?}{=} \frac{-2-3}{-2+2} = \frac{-5}{0}.$$

This is an undefined expression and indicates that $f(x)$ has an infinite discontinuity.

To find the limit as either $+\infty$ or $-\infty$, we do sign analysis on the simplified formula. The test intervals are separated by the *roots* and *discontinuities*. The roots are at solutions to $x - 3 = 0$; the discontinuities are at solutions to $x + 2 = 0$. We have a root at $x = 3$ and a discontinuity at $x = -2$, illustrated in the number line shown below.



For the limit, we need the signs of the function in each interval bordering the point $x = -2$. The intervals to test are $(-\infty, -2)$ and $(-2, 3)$.

$$f(-3) = \frac{-3-3}{-3+2} = 6$$
$$f(-1) = \frac{-1-3}{-1+2} = -4$$

We could update the number line with these signs.

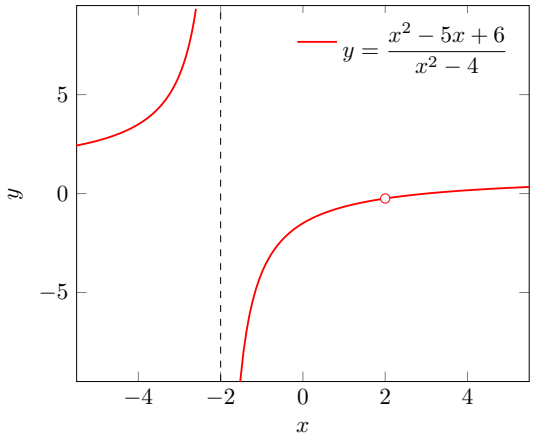
To the left of $x = -2$, we see that $f(x) > 0$ (positive), so a limit from the left at the vertical asymptote must be

$$\lim_{x\to -2^-} \frac{x-3}{x+2} = +\infty.$$

To the right of $x = -2$, we see that $f(x) < 0$ (negative), so a limit from the right at the vertical asymptote must be

$$\lim_{x\to -2+} \frac{x-3}{x+2} = -\infty.$$

On the graph of the function, shown below, we see that the graph is unbounded above ($+\infty$) to the left of the vertical asymptote and unbounded below ($-\infty$) to the right of the vertical asymptote.



□

Motivated by our example, we are ready for a definition of a removable discontinuity. A removable discontinuity occurs when the graph to the left and to the right of a discontinuity approach the same point, but the function itself is not defined to match.

Definition 5.4.8 A function f has a **removable discontinuity** at $x = c$ if $\lim_{x\to c} f(x)$ exists (left- and right-limits have same value) *and* $\lim_{x\to c} f(x) \neq f(c)$, either because they are different values or $f(c)$ does not exist. ◊

An infinite discontinuity occurs at any point where the function has an infinite limit.

Definition 5.4.9 A function f has an **infinite discontinuity** at $x = c$ if one or both of $\lim_{x\to c^-} f(x)$ and $\lim_{x\to c^+} f(x)$ is infinite. The graph $y = f(x)$ has a vertical asymptote $x = c$. ◊

A jump discontinuity occurs when the limits on the left and right of a point both exist but have different values. We usually see these with piecewise functions.

Definition 5.4.10 A function f has an **jump discontinuity** at $x = c$ if $\lim_{x\to c^-} f(x)$ and $\lim_{x\to c^+} f(x)$ both exist but $\lim_{x\to c^-} f(x) \neq \lim_{x\to c^+} f(x)$. The graph $y = f(x)$ has a vertical gap between the branches to the left and to the right of $x = c$. ◊

5.4.4 Continuity on Intervals

Having discussed the continuity of functions at individual points, we introduce the idea of describing continuity on intervals. We want to be able to say that the graph of the function is connected over an entire interval.

Recall that a limit of a function $\lim_{x\to c} f(x)$ is defined in terms of sequences $x_n \to c$ with $x_n \neq c$. When thinking about continuity on an interval, we also require that the sequences stay in the interval.

We begin with open intervals. An open interval (a, b) is the set $\{x : a < x < b\}$. Open intervals have the feature that for every value in the set, say $c \in (a, b)$, there will be a sub-interval (a, c) to the left of the point and another sub-interval (c, b) to the right of the point. In relation to a sequence with $x_n \to c$, we can deal with left- and right-limits inside the interval.

Definition 5.4.11 A function f is **continuous on the open interval** (a, b) if for every $c \in (a, b)$, f is continuous at $x = c$. $\Diamond$

Closed intervals are a little trickier. A closed interval $[a, b] = \{x : a \leq x \leq b\}$ includes the end points. For values c strictly between a and b, we know that there are subintervals to the left and the right of c. However, at $x = a$, the interval only contains points to the right; and at $x = b$, the interval only contains points to the left. Continuity of a function on a closed interval must take this into account.

Definition 5.4.12 A function f is **continuous on the closed interval** $[a, b]$ if for every $c \in (a, b)$, f is continuous at $x = c$ and $\lim_{x\to a^+} f(x) = f(a)$ and $\lim_{x\to b^-} f(x) = f(b)$. $\Diamond$

Continuity on an interval including only one end point requires one-sided continuity at that point using a limit that stays inside the interval. All of these definitions can be combined into a single definition.

Definition 5.4.13 A function f is **continuous on an interval** I if for every $c \in I$ and every sequence with values $x_n \in I$, $x_n \neq c$, and $x_n \to c$, we have $f(x_n) \to f(c)$. $\Diamond$

5.4.5 Extreme and Intermediate Value Theorems

There are two important theorems that describe what we know about functions that are continuous on closed intervals. The Extreme Value Theorem guarantees that any function that is continuous on a closed interval has a highest and lowest point within that interval. The Intermediate Value Theorem guarantees that a function that is continuous on a closed interval can not skip over any values between its values at the endpoints. The proofs for both of these theorems require advanced methods not taught at this level. We treat them essentially as axioms, statements that are true without proof.

Theorem 5.4.14 Extreme Value Theorem. *Suppose f is a function that is continuous on $[a, b]$. Then there must exist values $c_m, c_M \in [a, b]$ so that for any $x \in [a, b]$ we have*

$$f(c_m) \leq f(x) \leq f(c_M).$$

The values $f(c_m)$ and $f(c_M)$ are the minimum and maximum values, respectively, of the function f on $[a, b]$.

If a function is not continuous on $[a, b]$, then it does not necessarily have a maximum or minimum value. One way that this might happen is if f has a vertical asymptote within the interval. In that case, the values of f would be unbounded. Another way that this might happen is that f is bounded by what would be a maximum (or minimum) value but just doesn't reach it because of a sudden jump.

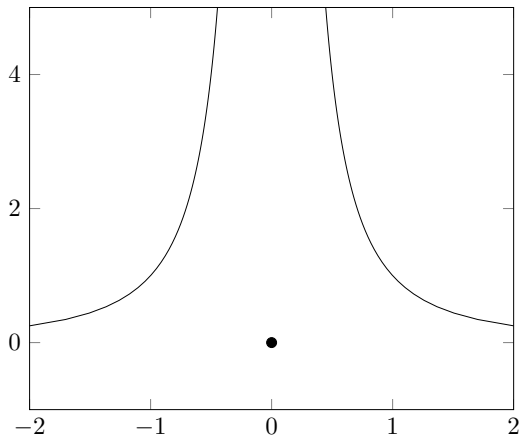
Example 5.4.15 Consider the function defined piecewise as

$$f(x) = \begin{cases} \frac{1}{x^2}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

This function has a non-removable discontinuity at $x = 0$, corresponding to a vertical asymptote. Because the formula has x^2 in the denominator (always positive), we have

$$\lim_{x\to 0} f(x) = +\infty.$$

This function is unbounded on the interval $[-1, 1]$ and has no maximum. It does have a minimum at $f(0) = 0$ since that is below the rest of the graph.

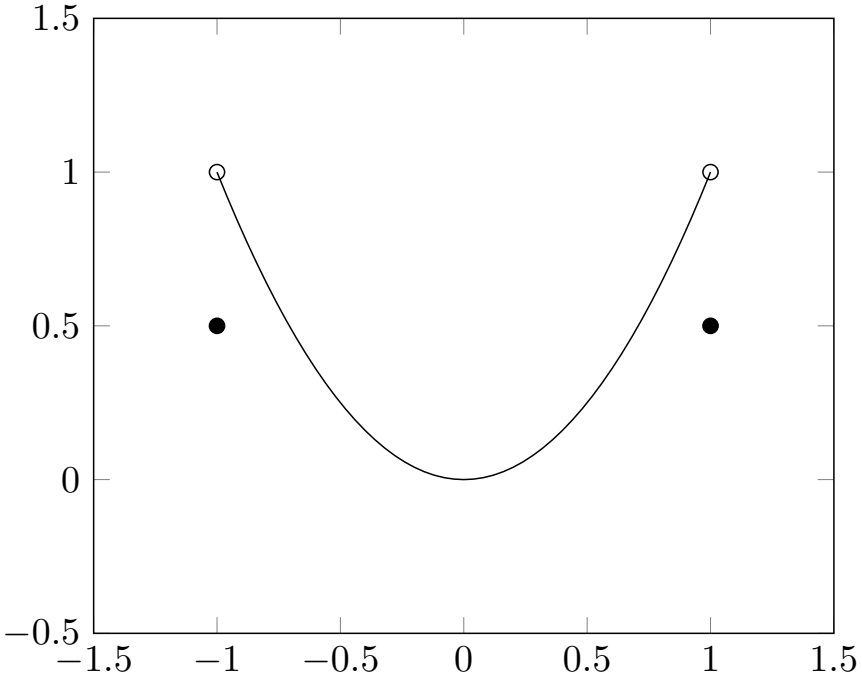


□

Example 5.4.16 Consider the function defined piecewise as

$$f(x) = \begin{cases} x^2, & -1 < x < 1, \\ \frac{1}{2}, & x = \pm 1. \end{cases}$$

This function has a removable discontinuities at $x = \pm 1$, where the limits are 1 but the values are $\frac{1}{2}$. In this case, f is continuous on $(-1, 1)$ but not on $[-1, 1]$. The maximum value should have been $y = 1$, but the graph never reaches that value because of the discontinuity. The function does have a minimum value at $f(0) = 0$.



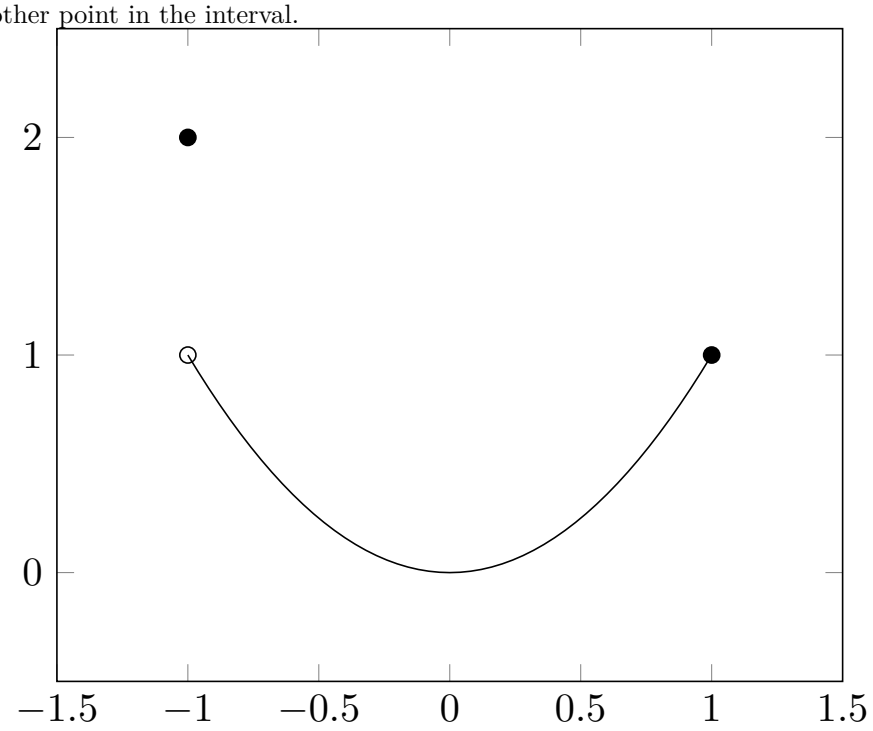
□

Example 5.4.17 Consider the function defined piecewise as

$$f(x) = \begin{cases} x^2, & -1 < x \leq 1, \\ 2, & x = -1. \end{cases}$$

This function has a removable discontinuity at $x = -1$. In this case, f is

continuous on $(-1, 1]$ but not on $[-1, 1]$. In spite of the discontinuity at $x = -1$, this function has a maximum value $f(-1) = 2$ because that value is above every other point in the interval.



□

The previous example is included to emphasize that a theorem gives conditions that guarantee something is true. But those conditions are not always required. The extreme value theorem gives conditions that guarantee a function will have a maximum value. There are no exceptions for a continuous function on a closed interval to have both maximum and minimum values. But there are discontinuous functions that have them as well. It is just that there are also discontinuous functions that do not have extreme values.

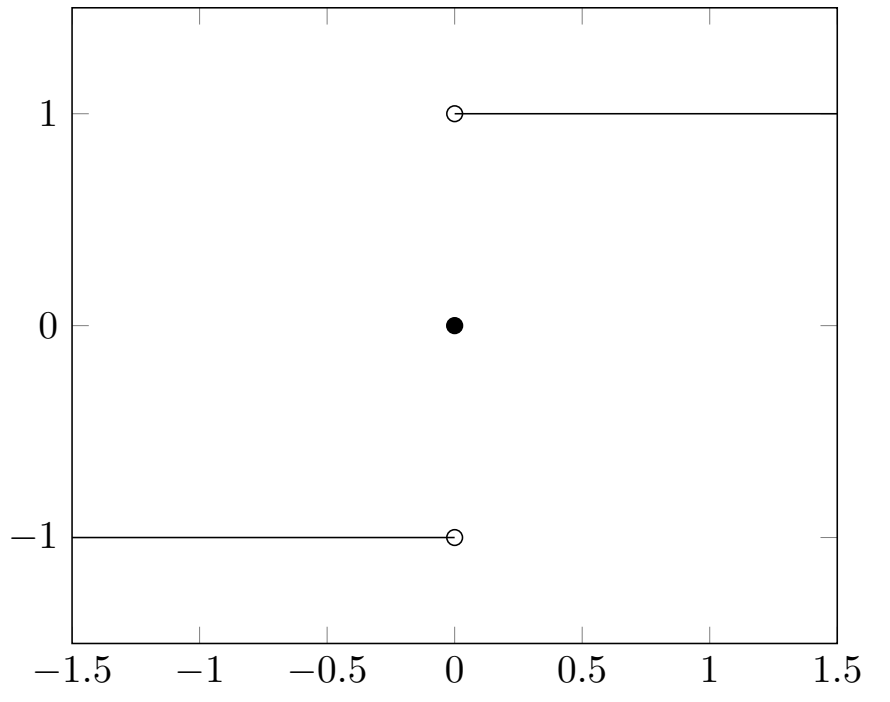
Theorem 5.4.18 Intermediate Value Theorem. *Suppose f is a function that is continuous on $[a, b]$. Then for every y between $f(a)$ and $f(b)$, there exists some $x \in (a, b)$ so that $f(x) = y$.*

The Intermediate Value Theorem guarantees that the graph of $y = f(x)$ intersects every horizontal line between $y = f(a)$ and $y = f(b)$ at least once for values of x between a and b. Because continuity is essentially connectedness, the only way for the graph to go from $y = f(a)$ to $y = f(b)$ is to cross through all intermediate values. A discontinuous function has the ability to jump across values without touching them.

Example 5.4.19 Consider the function defined piecewise as

$$f(x) = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0. \end{cases}$$

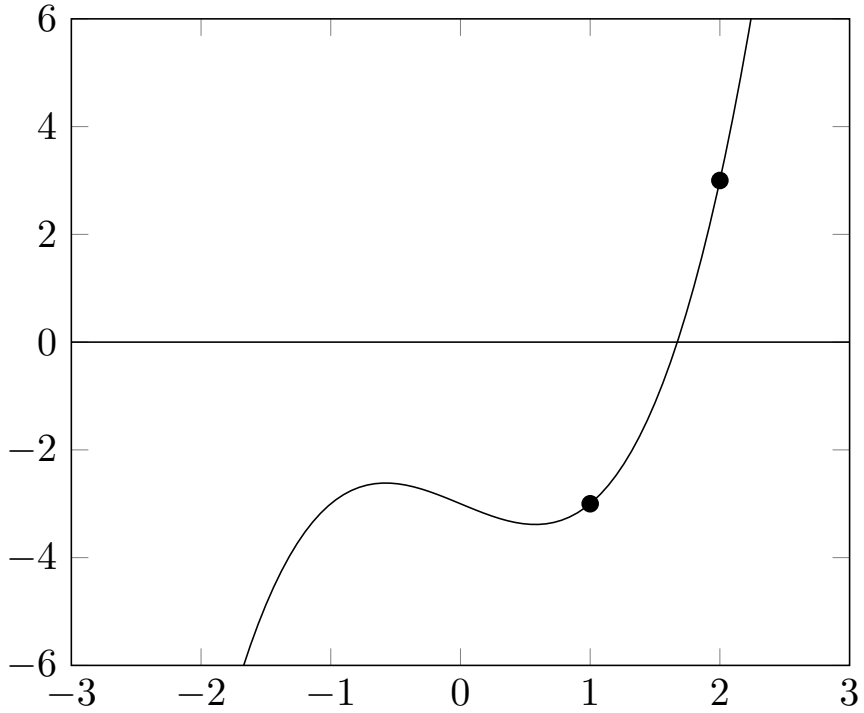
This function has a jump discontinuity at $x = 0$, and is otherwise constant. If we consider the interval $[-1, 1]$, the values at the endpoints are $f(-1) = -1$ and $f(1) = 1$. Except for $y = 0$, the function $y = f(x)$ has no solutions for $-1 < y < 1$ because of the jump.



□

The Intermediate Value Theorem allows us to know that a continuous function has a solution to an equation within a particular interval. If the interval is small, we have an approximation to the value of the solution. We say that the interval **brackets** the solution. Finding successively smaller bracketing intervals allows us to approximate the root to any needed precision. The Intermediate Value Theorem guarantees this works for continuous functions.

Example 5.4.20 The function $f(x) = x^3 - x - 3$ is continuous because it is a polynomial and defined everywhere. Because $f(1) = -3$ and $f(2) = 3$, we know that $f(x)$ must pass through every y-value between -3 and 3 for at least one value of x in the interval $(1, 2)$. In particular, if we are solving $f(x) = 0$, since $y = 0$ is between $f(1) = -3$ and $f(2) = 3$, we know that there is a solution x bracketed by the interval $[1, 2]$.



If we find a smaller interval, then we can know more precisely where the root occurs. In particular, since $f(1.6) = -0.504$ and $f(1.7) = 0.213$ and $y = 0$ is between those values, the Intermediate Value Theorem guarantees that our continuous function has a root bracketed by the interval $[1.6, 1.7]$. □

The Intermediate Value Theorem is our justification for performing sign analysis by testing intervals at single points. If we have solved for all of the roots (zeros) and all of the discontinuities of a function f, then f can not change sign on any interval containing none of the roots or discontinuities. Suppose that $f(a)$ and $f(b)$ have opposite sign with $a < b$. Then $y = 0$ is between $f(a)$ and $f(b)$. If a and b were chosen from an interval with no discontinuities, f must be continuous on $[a, b]$. The Intermediate Value Theorem would then guarantee that $f(x) = 0$ has a solution with $a < x < b$. Because the interval contained no roots, $f(a)$ and $f(b)$ must not have had opposite signs. Thus, f can never change sign on an interval containing no roots or discontinuities.

5.4.6 Summary

- A function defined by an algebraic formula has discontinuities at every point for which the formula is undefined.
- A rational function is defined as the quotient of two polynomials. Discontinuities of rational functions only occur at the zeros of the denominator. If the numerator and denominator have a zero at the same location $x = c$, then $x - c$ is a common factor that can be cancelled.
- A limit of the form $\frac{0}{0}$ is indeterminate. For rational functions with a limit of this form, we must factor and simplify to continue. If the limit ultimately exists (as a number), the discontinuity is removable and the limit corresponds to having a hole in the graph.
- A rational function with a limit of the form $\frac{L}{0}$ where $L \neq 0$ has an infinite discontinuity. The graph of such a function has a vertical asymptote. The

left- and right-side limits have signs that are based on sign analysis of the function in the intervals to the left and right of the point of interest.

- A function is continuous on an interval if it continuous at every point in the interval. If an end point is included in the interval, the function must be one-sided continuous from the side contained in the interval.
- The Extreme Value Theorem guarantees that whenever a function f is continuous on a closed interval $[a, b]$, there are points in the interval where f reaches its maximum and minimum (extreme) values restricted to that interval.
- The Intermediate Value Theorem guarantees that whenever a function f is continuous on a closed interval $[a, b]$, the equation $f(x) = y$ has a solution with $a < x < b$ for any y between $f(a)$ and $f(b)$.
- The Intermediate Value Theorem guarantees that a function can only change sign at its roots or discontinuities.

5.4.7 Exercises

Compute each of the following limits. If the limit is infinite, state both left- and right-side limits.

1. $\lim\limits_{x\to 3} \dfrac{2x-6}{x-3}$

2. $\lim\limits_{x\to 3} \dfrac{2x}{x-3}$

3. $\lim\limits_{x\to 2} \dfrac{x^2-x-2}{x-2}$

4. $\lim\limits_{x\to -1} \dfrac{x-2}{x^2-x-2}$

5. $\lim\limits_{x\to 2} \dfrac{x^2-2x-8}{x^2-4}$

6. $\lim\limits_{x\to -2} \dfrac{x^2-2x-8}{x^2-4}$

7. $\lim\limits_{x\to 4} \dfrac{x^2-2x-8}{x^2-4}$

8. $\lim\limits_{x\to 0} \dfrac{3x^2-5x+2}{2x^2-x-1}$

9. $\lim\limits_{x\to -\frac{1}{2}} \dfrac{3x^2-5x+2}{2x^2-x-1}$

10. $\lim\limits_{x\to 1} \dfrac{3x^2-5x+2}{2x^2-x-1}$

Classify the discontinuities for each function, if any. State the limits at each discontinuity.

11. $f(x) = \dfrac{3}{x^2-5}$

12. $f(x) = \dfrac{2x}{x^2+3x}$

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

14. $f(x) = \dfrac{x^3 + 7x^2 + 12x}{x^2 + 3x}$

15. $f(x) = \begin{cases} 3x, & x < 1, \\ 2, & x = 1, \\ 4 - x^2, & x > 1. \end{cases}$

16. $f(x) = \begin{cases} \frac{3}{x-3}, & x < 2, \\ 2x - 7, & x > 2. \end{cases}$

5.5 Instantaneous Rate of Change

5.5.1 Overview

Accumulation functions are defined in terms of their rate of accumulation. That is, we started by knowing the rate of accumulation $f'(x)$ and used that rate and an initial value to create the accumulation function

$$f(x) = f(a) + \int_a^x f'(z)\,dz.$$

Thus far, we only know a few elementary accumulation formulas for simple polynomials. What about other functions?

Perhaps the biggest breakthrough in the historical development of calculus was the recognition of a relationship between accumulation computed through definite integrals and the rate of change computed through derivatives. A definite integral represents the exact accumulation for a given rate as the independent variable goes between two points. A Riemann sum approximates this accumulation by summing increments that treat the rate of accumulation as if constant over short intervals. The definite integral is equal to the limit of the Riemann sums.

In this section, we introduce the derivative of a function. The derivative represents the instantaneous rate of change at a single point. We will introduce the average rate of change between two points. The average rate of change approximates the instantaneous rate of change when the two points are close together. The derivative will then equal the limit of the average rate of change.

5.5.2 Slope and the Average Rate of Change

Consider the point–slope equation of a line

$$y - b = m(x - a),$$

which is the equation of a line with slope m and passing through a point (a, b). When we solve for y, so that y is a function of x,

$$y = f(x) = b + m(x - a),$$

we can recognize that this is in the form of an accumulation function with constant rate,

$$f(x) = b + \int_a^x m\,dx.$$

That is, the slope acts as the rate of accumulation.

The slope of a line also represents a rate of change, meaning a ratio of covarying changes. Given two points on the line, (x_1, y_1) and (x_2, y_2), the slope is defined by the ratio

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

If we think of the point $f(a) = b$ as the initial value for $f(x)$, then the increment of change for f defined by

$$\Delta f = f(x) - f(a)$$

will always be proportional to the change in the input

$$\Delta x = x - a.$$

The slope is the proportionality constant representing this constant ratio of the changes, or rate of change.

Only in the case of a linear function do we actually find that the rate of change $\Delta f/\Delta x$ is a constant. When the rate of change varies, the value of the rate of change depends on the interval chosen. We call this the **average rate of change**.

Definition 5.5.1 The **average rate of change** of a function $f(x)$ going from a to b is defined as the ratio

$$\left.\frac{\Delta f}{\Delta x}\right|_{a,b} = \frac{f(b) - f(a)}{b - a}.$$

$\Diamond$

The average rate of change is equal to the slope of the line that joins the two points $(a, f(a))$ and $(b, f(b))$. That line is called the **secant line**. The order of the points for the average rate of change does not matter.

$$\begin{aligned}\left.\frac{\Delta f}{\Delta x}\right|_{b,a} &= \frac{f(a) - f(b)}{a - b} \\ &= \frac{-(f(b) - f(a))}{-(b - a)} \\ &= \frac{f(b) - f(a)}{b - a} \\ &= \left.\frac{\Delta f}{\Delta x}\right|_{b,a}\end{aligned}$$

Consequently, we often just refer to calculating the average rate of change over an interval $[a, b]$.

Example 5.5.2 Find the average rate of change of $f(x) = x^3 - 4x$ over the interval $[1, 2]$.

Solution. The average rate of change is the slope of the line joining the points $(1, f(1))$ and $(2, f(2))$. So we first need to calculate the function values:

$$f(1) = (1)^3 - 4(1) = -3; \qquad f(2) = (2)^3 - 4(2) = 0.$$

This allows us to compute the average rate of change:

$$\left.\frac{\Delta f}{\Delta x}\right|_{1,2} = \frac{f(2) - f(1)}{2 - 1} = \frac{0 - -3}{1} = 3.$$

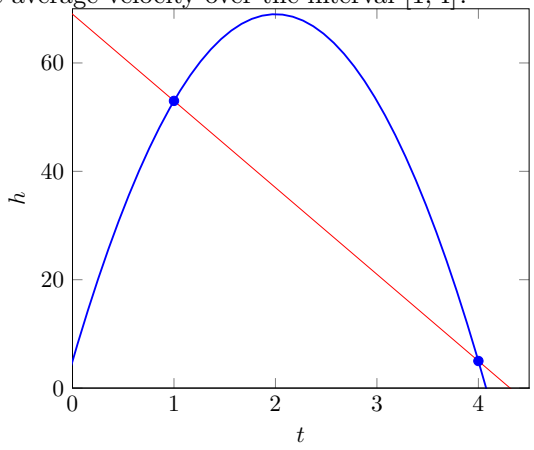
$\square$

A rate of change for a physical quantity has units corresponding to the ratio it describes. One of the most common examples is calculating the velocity of an object. Velocity measures the distance traveled per unit time, which means velocity is the rate of change of position with respect to time. How would we compute a velocity? Measure the position of the object at two times, measure how much the position changed, and divide by the amount of time that passed. The average velocity is the ratio of the change in position over the change in time.

Example 5.5.3 If the position (height, in feet) of an object above the ground is defined by the function of time (in seconds)

$$h(t) = 5 + 64t - 16t^2,$$

then what is the average velocity over the interval $[1, 4]$?



Solution. The average velocity is the slope of the line joining the points $(1, h(1))$ and $(4, h(4))$. So we first need to calculate the function values which gives the two positions:

$$h(1) = 5 + 64(1) - 16(1)^2 = 53; \qquad f(4) = 5 + 64(4) - 16(4)^2 = 5.$$

This allows us to compute the average velocity as an average rate of change:

$$\left.\frac{\Delta h}{\Delta t}\right|_{[1,4]} = \frac{h(4) - h(1)}{4 - 1} = \frac{5 - 53}{3} = \frac{-48}{3} = -16.$$

So the average velocity is -16 ft/s, since the height dropped by 48 feet during those 3 seconds. □

In the previous example,the graph of the height of the object as a function of time shows that the ball was initially going up and then came down. However, the average rate of change calculated gave a negative rate. Over the three minutes in the interval, the ball started going up and then fell enough so that the overall change was negative. If we wanted to know the speed of the object when $t = 1$, this interval is much too large to provide a good approximation.

5.5.3 Instantaneous Rate of Change

The instantaneous rate of change of a function is the rate of change at a particular point. When we think about a rate of change as representing a slope, our existing strategies require knowing two different points to compute a slope. We expect that the average rate of change between two points should approximate the instantaneous rate of change if the increment Δx is not too big. Further, we expect that the approximation should improve by making Δx smaller.

The figure below is an interactive graph of $f(x) = x^3 - 4x$ showing one point at $x = 1$ and a second point at $x = 1 + h$ where the value of h is controlled by a slider. The average rate of change and slope of the secant line between these two points is also calculated. Notice that when h is close to zero, $h \approx 0$, the secant line is closer and closer to a tangent line.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 5.5.4

Graphically, the instantaneous rate of change equals the slope of the tangent line at the point. A tangent line is defined in terms of a single point such that the line is the line that best approximates the function near that point. We define the instantaneous rate of change, or the derivative at a point, as a limit of the average rate of change.

Definition 5.5.5 The **instantaneous rate of change** of a function $f(x)$ at $x = a$ is the derivative at the point and is defined as the limit of the average rate of change when the width of the interval is made arbitrarily small:

$$\left.\frac{df}{dx}\right|_a = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h}$$

or

$$\left.\frac{df}{dx}\right|_a = \lim_{x\to a} \frac{f(x)-f(a)}{x-a}$$

◊

The limits defining the instantaneous rate of change or derivative, when they exist, represent the limiting value of the average rate of change for a sequence of points x_n that converge to a, $x_n \to a$. The first limit using h defines this sequence as $x_n = a + h_n$ where $h_n \to 0$. That is, h represents the value of Δx. This suggests defining the average rate of change in terms of a reference point a and the displacement h to the second point,

$$\left.\frac{\Delta f}{\Delta x}\right|_{a;h} = \frac{f(a+h)-f(a)}{h}.$$

The derivative is the limit of this average rate of change as $h \to 0$. For the limit to exist, the limit value of the average rate of change can not depend on which sequence we choose. We explore the definition by considering an example.

Example 5.5.6 Find the instantaneous rate of change of $f(x) = x^3 - 4x$ at 1.

Solution. In the earlier example, we found the average rate of change of f on the interval $[1, 2]$:

$$\left.\frac{\Delta f}{\Delta x}\right|_{[1,2]} = 3.$$

The width of this interval was $h = \Delta x = 2 - 1 = 1$. We need to consider a sequence of intervals that include $x = 1$ but have a width that is decreasing to zero. The following list of intervals is the start of one possible sequence.

$$[1, 2],\ [1, 1.1],\ [1, 1.01],\ [1, 1.001],\ \ldots$$

To make this calculation more manageable, it is sometimes easier to create an entry in a table that shows the value of the function at the second point that is needed to compute a slope. The table only shows the summary information. The average rate of change for each intervals was computed using the definition

$$\left.\frac{\Delta f}{\Delta x}\right|_{[1,x_n]} = \frac{f(x_n)--3}{h}$$

where $h = b_k - 1$.

Table 5.5.7 Illustration of using shrinking intervals to estimate the instantaneous rate of change.

n	$[1, x_n]$	$h = \Delta x$	x_n	$f(x_n)$	$\left.\frac{\Delta f}{\Delta x}\right\|_{[1,x_n]}$
1	$[1, 2]$	1	2	0	3
2	$[1, 1.1]$	0.1	1.1	-3.069	-0.69
3	$[1, 1.01]$	0.01	1.01	-3.009699	-0.9699
4	$[1, 1.001]$	0.001	1.001	-3.000996999	-0.996999

From this table, we can see that the spacing between the points h is approaching zero as the second point in the interval is approaching the point in question. As this happens, the sequence of values representing the average rate of change on these intervals appears to be approaching the value of -1. From the table, we would estimate that this is the instantaneous rate of change:

$$\left.\frac{df}{dx}\right|_1 \approx -1.$$

□

In the previous example, the limit of the average rate of change was an easily recognized value because it was an integer. When using a table and the limit is not an integer, you need to look for a decimal value that is approximated. This often works better if we consider a sequence of intervals on both sides of the point of interest and use the average of the two sides.

Example 5.5.8 Find the instantaneous rate of change of $f(x) = 3^x$ at 1.

Solution. We will create two sequences of decreasing width intervals, one on the left and the other on the right of 1. This will allow us to more easily recognize the decimal representation of the limiting value. In addition, we will use a variable h to represent the offset from 1 to the other endpoint of the interval. One sequence will involve negative values of h (on the left) and the other will involve positive values of h (on the right). Notice how in these tables, we require more decimal places for the value of $f(1 + h)$ in order to obtain the average rate of change to the same number of significant digits.

Table 5.5.9 Illustration of using shrinking intervals on the left to estimate the instantaneous rate of change.

h	$f(1 + h)$	$f(1)$	Δf	$\left.\frac{\Delta f}{\Delta x}\right\|_{1;h}$
-0.1	2.68788	3	-0.31212	3.1212
-0.01	2.967222	3	-0.032778	3.2778
-0.001	2.9967160	3	-0.0032940	3.2940

Table 5.5.10 Illustration of using shrinking intervals on the right to estimate the instantaneous rate of change.

h	$f(1 + h)$	$f(1)$	Δf	$\left.\frac{\Delta f}{\Delta x}\right\|_{1;h}$
0.1	3.34837	3	0.34837	3.4837
0.01	3.033140	3	0.033140	3.3140
0.001	3.0032976	3	0.0032976	3.2976

Looking at the columns for the average rate of change in these two tables, we notice that the values are rising when $h < 0$ as the size of the interval

shrinks but that the values are dropping when $h > 0$. The limiting value should therefore be somewhere between 3.2940 and 3.2976. The average of these values gives an even better approximation,

$$\left.\frac{df}{dx}\right|_1 \approx 3.2958.$$

□

A dynamic graph of this calculation is illustrated below.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 5.5.11

5.5.4 Using Algebra for Rates of Change

In the previous sections, we computed the instantaneous rate of change by looking for a limiting value in the average rate of change for a sequence of intervals that had decreasing width. It is important to realize that this is the fundamental definition of the instantaneous rate of change. In many cases involving basic algebraic functions, like polynomials, it is possible to determine the limiting value of this process algebraically. This is because we can find a formula for the average rate of change for any value of the spacing h and determine what will happen when $h \to 0$.

The process for using this strategy is to consider the same calculations that we used to form the tables, but to determine the formula for each step instead of numerical values for the particular values of h.

- Identify the function $f(x)$ and the point of interest a.
- Use composition to find the formula for $f(a+h)$ and expand.
- Compute the change in the function $\Delta f = f(a+h) - f(a)$ and simplify.
- Compute the average rate of change,

  $$\frac{\Delta f}{\Delta x} = \frac{f(a+h) - f(a)}{h},$$

  and use factoring in order to find a formula that does not divide by h.
- The limit or instantaneous rate of change is the value of this final formula when $h \to 0$. This works because simplifying the rate of change formula found a new expression that was continuous at the point $h = 0$.
- If it is not possible to rewrite without dividing by h, then you can use the formula itself in a table to see what happens when $h \to 0$ by testing a sequence of values for h that approach 0.

Take a moment to look at how the calculations described above compare to our process of approximating the derivative using a table.

Example 5.5.12 Find the instantaneous rate of change for $f(x) = x^2 + 3x$ at 2.

Solution. We will just follow the steps outlined above.

1. The function has been identified and the point given. We also need $f(2) = 2^2 + 3(2) = 10$ in later steps.

2. Find $f(2+h)$ and expand:

$$\begin{aligned} f(2+h) &= (2+h)^2 + 3(2+h) = (2+h)(2+h) + 3(2+h) \\ &= 4 + 4h + h^2 + 6 + 3h = 10 + 7h + h^2 \end{aligned}$$

3. Compute the change in function value, $\Delta f = f(2+h) - f(2)$:

$$\Delta f = f(2+h) - f(2) = (10 + 7h + h^2) - (10) = 7h + h^2$$

4. Compute the formula for the average rate of change:

$$\left.\frac{\Delta f}{\Delta x}\right|_{2;h} = \frac{7h + h^2}{h} = \frac{h(7+h)}{h} = 7 + h.$$

   The average rate of change is only defined when $h \neq 0$.

5. The limit of the average rate of change uses the simplified formula on both sides of $h = 0$. Because the reduced formula is continuous, the limiting value can be found by substitution:

$$\left.\frac{df}{dx}\right|_2 = \lim_{h\to 0} 7 + h = 7 + 0 = 7.$$

□

The definition of the derivative allows us to compute the rate of change for functions other than polynomials.

Example 5.5.13 Find the instantaneous rate of change for $f(x) = \frac{1}{3x+1}$ at 1.

Solution. We again just follow the steps outlined above.

1. The function has been identified and the point given. We also need $f(1) = \frac{1}{3(1)+1} = \frac{1}{4}$ in later steps.

2. Find $f(1+h)$ and expand:

$$\begin{aligned} f(1+h) &= \frac{1}{3(1+h)+1} = \frac{1}{3+3h+1} \\ &= \frac{1}{4+3h} \end{aligned}$$

3. Compute the change in function value, $\Delta f = f(1+h) - f(1)$:

$$\Delta f = f(1+h) - f(1) = \frac{1}{4+3h} - \frac{1}{4}.$$

   We will later need to simplify this, so let us find a common denominator, which requires multiplying each fraction's numerator and denominator by the missing factor:

$$\Delta f = \frac{4}{4(4+3h)} - \frac{4+3h}{4(4+3h)} = \frac{4-(4+3h)}{4(4+3h)} = \frac{-3h}{4(4+3h)}.$$

4. Compute the formula for the average rate of change:

$$\left.\frac{\Delta f}{\Delta x}\right|_{1;h} = \frac{\left(\frac{-3h}{4(4+3h)}\right)}{h}$$

$$= \frac{1}{h} \cdot \frac{-3h}{4(4+3h)}$$
$$= \frac{-3}{4(4+3h)},$$

defined when $h \neq 0$.

5. The instantaneous rate of change is the limit of the average rate of change. Because the reduced formula is continuous at $h = 0$, the limit can be computed using substitution

$$\left.\frac{df}{dx}\right|_1 = \lim_{h\to 0} \frac{-3}{4(4+3h)}$$
$$= \frac{-3}{4(4+3(0))} = \frac{-3}{16}.$$

□

5.5.5 Interpretation of the Rate of Change

The rate of change often has a physical interpretation. For example, if we know the position (e.g., height) as a function of time, then the rate of change corresponds to the velocity of the object. In chemistry, if we know the concentration of a reactant in solution as a function of time, then the rate of change of concentration describes the reaction rate. We can also have rates of change with respect to variables other than time. For example, in biology, the number of new fish born in a year (called recruitment) might be a function of the current population size (called the stock). The rate of change of the recruitment with respect to the stock measures how much the recruitment would change per unit increase in the stock. In economics, if we know how an equation relating the revenue that corresponds to the number of items being sold, then the rate of change of revenue with respect to the number of items is called the marginal revenue and corresponds to the amount of revenue change per extra item sold.

The units of a rate of change are determined by the units in the ratio. Since velocity is the rate of change of position with respect to time, the units of velocity are the units of length divided by the units of time, such as kilometers per hour or meters per second. Marginal revenue is the rate of change of revenue with respect to items sold, so the units would be a monetary unit per item, such as dollars per item.

Example 5.5.14 In a chemical reaction, the concentration of a reactant is measured as a function of time. If C represents the concentration measured grams per liter and t represents the time elapsed since the reaction began measured in seconds, then this function is the mapping $t \mapsto C$.

What are the units of $\frac{dC}{dt}$?

Solution. The quickest solution to answer this question says to take the units of the dependent variable C and divide by the units of the independent variable t. That would give grams per liter per second, or $\frac{\text{g}}{\text{L}\cdot\text{s}}$. To clarify why we divide the units in that way, recall that the rate of change is the ratio of the change in C (output) over the change in t (input). Because C has units of $\frac{\text{g}}{\text{L}}$, those will also be the units of ΔC. Dividing this by the change in time Δt, which has units of seconds s, we obtain the units of the rate of change. □

We can use the definition of the rate of change to calculate an instantaneous rate of change for a given model of a physical quantity.

Example 5.5.15 A population grows in time according to a model

$$P(t) = 400 \cdot 1.1^t,$$

where P is the population count and t is the time in years from the start of the model. How fast is the population growing when $t = 2$?

Solution. Because we do not yet know the rules of differentiation to find the algebraic rule for the derivative, we will use the definition of the derivative in terms of a limit. The derivative is defined as

$$\left.\frac{dP}{dt}\right|_2 = \lim_{h\to 0} \frac{P(2+h) - P(2)}{h}.$$

Substituting the formula for $P(t)$, we can write this as

$$\left.\frac{dP}{dt}\right|_2 = \lim_{h\to 0} \frac{400 \cdot 1.1^{(2+h)} - 400 \cdot 1.1^2}{h}.$$

We can now set up a table of values of this average rate of change for values of h with $h \to 0$. Our table will show a final value with 6 digits of accuracy so that we can visualize the limit converging.

Table 5.5.16 Table of $\frac{\Delta P}{\Delta t}$ for $h < 0$.

h	$\left.\frac{\Delta P}{\Delta t}\right\|_{2;h} = \frac{400 \cdot 1.1^{(2+h)} - 400 \cdot 1.1^2}{h}$
-0.1	$\frac{400 \cdot 1.1^{1.9} - 400 \cdot 1.1^2}{-0.1} \approx 45.9011$
-0.01	$\frac{400 \cdot 1.1^{1.99} - 400 \cdot 1.1^2}{-0.01} \approx 46.1082$
-0.001	$\frac{400 \cdot 1.1^{1.999} - 400 \cdot 1.1^2}{-0.001} \approx 46.1279$

Table 5.5.17 Table of $\frac{\Delta P}{\Delta t}$ for $h > 0$.

h	$\left.\frac{\Delta P}{\Delta t}\right\|_{2;h} = \frac{400 \cdot 1.1^{(2+h)} - 400 \cdot 1.1^2}{h}$
0.1	$\frac{400 \cdot 1.1^{2.1} - 400 \cdot 1.1^2}{0.1} \approx 46.3507$
0.01	$\frac{400 \cdot 1.1^{2.01} - 400 \cdot 1.1^2}{0.01} \approx 46.1521$
0.001	$\frac{400 \cdot 1.1^{2.001} - 400 \cdot 1.1^2}{0.001} \approx 46.1323$

Comparing the tables using values on the left ($h < 0$) and values on the right ($h > 0$), our calculations inform us that $\left.\frac{dP}{dt}\right|_2$ is between 46.1279 and 46.1323 individuals per year. The best estimate would be the average of the rates of change for $h = \pm 0.001$, or $\left.\frac{dP}{dt}\right|_2 \approx 46.1301$ individuals per year. □

Example 5.5.18 A company determines that the number of a particular item it can sell in a month depends on the sale price for that item according to the demand function

$$x = \frac{12000}{p+3}$$

where x measures the number of items sold per month and p measures the sale price of an item in dollars.

Find the monthly revenue for a sale price of $p = 5$ and determine rate of change of monthly revenue with respect to the sale price when $p = 5$. Compare this rate of change with the actual change in revenue for particular price changes.

Solution. The total revenue R is equal to the number of items sold times the price of each item. That is,

$$R = x \cdot p.$$

Subsituting the demand function in place of x, we find

$$p \mapsto R = \frac{12000p}{p+3}.$$

We can find the monthly revenue using $p = 5$ to obtain

$$p = 5 \ \mapsto \ R = \frac{12000(5)}{5+3} = 7500.$$

The monthly revenue will be \$7500 when the item price is $p = 5$.

To find the rate of change of R with respect to p, we write down the definition of the derivative.

$$\begin{aligned}\left.\frac{dR}{dp}\right|_5 &= \lim_{h\to 0} \frac{R(5+h) - R(5)}{h} \\ &= \lim_{h\to 0} \frac{\frac{12000(5+h)}{5+h+3} - 7500}{h}\end{aligned}$$

We can simplify this formula using algebra.

$$\begin{aligned}\left.\frac{dR}{dp}\right|_5 &= \lim_{h\to 0} \frac{\frac{60000+12000h}{8+h} - 7500}{h} \\ &= \lim_{h\to 0} \frac{\frac{60000+12000h}{8+h} - \frac{7500(8+h)}{8+h}}{h} \\ &= \lim_{h\to 0} \frac{\frac{60000+12000h}{8+h} - \frac{60000+7500h)}{8+h}}{h} \\ &= \lim_{h\to 0} \frac{\left(\frac{4500h}{8+h}\right)}{h}\end{aligned}$$

We should never use a fraction within a fraction. Dividing by h is equivalent to multiplying by $1/h$. This allows us to rewrite our formula for the average rate of change in a way that can be simplified before finding the limit.

$$\begin{aligned}\left.\frac{dR}{dp}\right|_5 &= \lim_{h\to 0} \frac{4500h}{8+h} \cdot \frac{1}{h} \\ &= \lim_{h\to 0} \frac{4500}{8+h} \\ &= \frac{4500}{8+0} \\ &= 562.50\end{aligned}$$

This rate of change tells us that starting at a unit price of $p = 5$, the monthly revenue is increasing at a rate of \$562.50 for every \$1.00 increase in price. Because the derivative is an instantaneous rate of change, it will not match the average rate of change. For example, if we set the unit price at $p = 6$, we get a monthly revenue of $R = 8000$. The average rate of change for the monthly revenue with respect to price would be

$$\left.\frac{\Delta R}{\Delta p}\right|_{5,6} = \frac{R(6) - R(5)}{6-5} = \frac{8000 - 7500}{1} = 500,$$

which is lower than the instantaneous rate of change. However, if we consider a smaller change in unit price, say $h = 0.1$, to get $p = 5.10$, we now find

$$\left.\frac{\Delta R}{\Delta p}\right|_{5,5.10} = \frac{R(5.10) - R(5)}{0.1} \approx \frac{7555.55 - 7500}{1} = 555.5,$$

which is much closer to the instantaneous rate of change. Even smaller changes in the price will result in an average rate of change that is a closer approximation to the instantaneous rate of change. □

5.5.6 Summary

- A rate of change is the ratio of the change in a dependent variable (output) over the change in the independent variable (input).
- The **average rate of change** of $x \mapsto y$ between $x = a$ and $x = b$ is

$$\left.\frac{\Delta y}{\Delta x}\right|_{a,b} = \frac{y(b) - y(a)}{b - a}$$

  and represents the slope of the secant line or chord connecting points $(a, y(a))$ and $(b, y(b))$.
- The **instantaneous rate of change** of $x \mapsto y$ at a point $x = a$ is called the **derivative** at $x = a$, defined by a limit of the average rate of change where the second point approaches the first point:

$$\left.\frac{dy}{dx}\right|_{a} = \lim_{h \to 0} \frac{y(a + h) - y(a)}{h}.$$

  The instantaneous rate of change represents the slope of the tangent line to the curve at the point $(a, y(a))$.
- The units of measurement of a physical rate of change are the units of the dependent variable divided by the units of the independent variable.

5.5.7 Exercises

For the given function, compute the average rate of change over the given interval. Write the equation of the corresponding secant line.

1. $f(x) = x^2 - 2x$ on $[1, 3]$.

2. $x \mapsto y = \dfrac{3}{x}$ on $[1, 2]$.

3. $y(x) = 2^x$ on $[2, 4]$.

Use a table to approximate the derivative of the given function at the specified point. Include enough data to ensure that your approximation has five digits of accuracy.

4. Find $\left.\frac{dy}{dx}\right|_1$ for $y(x) = x^2 + x$.

5. Find $\left.\frac{dQ}{dt}\right|_1$ for $Q(t) = 3^t$.

6. Find $\left.\frac{df}{dx}\right|_2$ for $f(x) = \sqrt{x}$.

Use the definition of the derivative and compute the resulting limit exactly to find the exact instantaneous rate of change for the specified function. Write the equation of the corresponding tangent line.

7. Find $\left.\frac{dy}{dx}\right|_1$ for $y(x) = x^2 + x$.

8. Find $\left.\frac{df}{dx}\right|_2$ for $y(x) = x^2 - 3x$.

9. Find $\left.\frac{df}{dx}\right|_{-1}$ for $y(x) = 2x^2 + 3x$.

10. Find $\left.\frac{dP}{dt}\right|_2$ for $P(t) = t^3 - 2t$.

11. Find $\left.\frac{dz}{dr}\right|_3$ for $z(r) = \dfrac{3}{r}$.

12. Find $\left.\frac{dR}{dz}\right|_2$ for $R(z) = \dfrac{3}{2z+1}$.

Applications

13. The height h of an object dropped from a height of 100 feet as a function of time t in seconds since the object was dropped satisfies a model

$$h(t) = 100 - 16t^2.$$

(a) Find the average velocity over the interval $[1, 2]$.

(b) Find the instantaneous velocity at $t = 1$.

Be sure to use appropriate units. What is the interpretation of the sign of these values?

14. A vehicle accelerates from a stop at time $t = 0$ (in seconds) to highway speed at $t = 6$. The velocity v of the vehicle (in miles per hour) is a function of t given by

$$v(t) = -\frac{5}{9}t^3 + 5t^2, \quad 0 \le t \le 6.$$

The rate of change of velocity is called acceleration.

(a) Find the velocity at $t = 0$ and $t = 6$.

(b) Determine the average acceleration over the interval $[0, 6]$.

(c) Find the velocity at $t = 1$ and $t = 2$.

(d) Determine the average acceleration over the interval $[1, 2]$.

(e) Determine the instantaneous acceleration at $t = 1$.

Be sure to use appropriate units. Was the vehicle accelerating at a constant rate?

15. Researchers studying tree growth in Germany found a relationship between the age of the tree (years) and its total weight (kg) for sycamore maples in a particular forest. If a represents the age of a tree and W is its total weight, the relationship $a \mapsto W$ was modeled using regression with a quadratic polynomial as

$$W = 31.6601 - 7.6351a + 0.4334a^2.$$

(Albert et al, 2014)

(a) Find the mass of a fifteen year old tree and of a twenty year old tree.

(b) Determine the average rate of change of the mass of a tree with respect to age over the interval $[15, 20]$.

(c) Determine the instantaneous rate of change of the mass of a tree with respect to age at the age $a = 15$.

Be sure to use appropriate units.

16. Kinesin is a motor protein that facilitates transport along the axon of a neuron. Researchers recorded the velocity of single kinesin molecules pulling microscopic glass beads subject to a constant resistive force of approximately 1 pN with varying ATP concentrations. The velocity of transport along a microtubule V ($\frac{\text{nm}}{\text{s}}$) depended on the concentration of ATP C ($\mu\text{mol} \cdot \text{L}$) and was modeled with a Michaelis-Menten equation

$$V = \frac{85C}{C + 814}.$$

(Schnitzer et al., 2000)

(a) Find the velocity of transport when ATP has concentrations of 100 $\frac{\mu\text{mol}}{\text{L}}$ and 200 $\frac{\mu\text{mol}}{\text{L}}$

(b) Determine the average rate of change of the transport velocity with respect to concentration over the interval $[100, 200]$.

(c) Determine the instantaneous rate of change of the transport velocity with respect to concentration at a concentration of 100 $\frac{\mu\text{mol}}{\text{L}}$.

Be sure to use appropriate units.

5.6 The Fundamental Theorem of Calculus, Part One

5.6.1 Overview

When we introduced the definite integral, we also learned about accumulation functions. An accumulation function is a function A defined as a definite integral from a fixed lower limit a to a variable upper limit where the integrand is a given function f,

$$A(x) = A(a) + \int_a^x f(z)\,dz.$$

The function f was called the rate of accumulation for the function A, and we wrote $A'(x) = f(x)$. Then we defined another rate of change, the instantaneous rate of change, with a corresponding function called the derivative. For a function $F(x)$, the derivative was defined by a limit,

$$\frac{dF}{dx}(x) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h}.$$

This section establishes a relation between these two concepts of the rate of change. The Fundamental Theorem of Calculus proves that the derivative of an accumulation function exactly matches the rate of accumulation at whenever the rate of accumulation is continuous. That is, the instantaneous rate of change of a quantity, which graphically gives the slope of the tangent line on the graph, is exactly the same as the value of the rate of accumulation when the function is expressed as an accumulation using a definite integral. The proof of the fundamental theorem relies on properties of continuous functions as well as properties of limits.

5.6.2 Illustration of an Example

To illustrate the concept that we will prove, let us consider a simple polynomial function

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

Using our rules of accumulation, we know that $f(x)$ can be written as an accumulation,

$$f(x) = 5 + \int_0^x 3z^2 - 3\,dz.$$

What happens if we compute the derivative using the definition?

We start with some preparatory algebra based on $f(x) = x^3 - 3x + 5$.

$$\begin{aligned} f(x+h) &= (x+h)^3 - 3(x+h) + 5 \\ &= (x+h)(x+h)(x+h) - 3(x+h) + 5 \\ &= (x^2 + 2xh + h^2)(x+h) - 3x - 3h + 5 \\ &= x^3 + 3x^2h + 3xh^2 + h^3 - 3x - 3h + 5 \end{aligned}$$

$$\begin{aligned} f(x+h) - f(x) &= (x^3 + 3x^2h + 3xh^2 + h^3 - 3x - 3h + 5) - (x^3 - 3x + 5) \\ &= 3x^2h + 3xh^2 + h^3 - 3h \\ &= h(3x^2 + 3xh + h^2 - 3) \end{aligned}$$

The derivative can be computed using the limit:

$$\begin{aligned}\frac{df}{dx}(x) &= \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}\\ &= \lim_{h\to 0}\frac{h(3x^2+3xh+h^2-3)}{h}\\ &= \lim_{h\to 0} 3x^2+3xh+h^2-3\\ &= 3x^2+3x(0)+(0)^2-3\\ &= 3x^2-3.\end{aligned}$$

The derivative exactly matches the rate of accumulation.

Our ultimate goal in this section is to show that this will always happen for accumulation functions. To reach this goal, we require some additional concepts.

5.6.3 Average Value of a Function

The derivative computes the limit of the average rate of change. In preparation for the Fundamental Theorem of Calculus, we need a relation between the idea of average value and the definite integral.

Consider how we compute the average value of a list of numbers. We add up all the values and divide by the number of values in the list. When thinking about a function f on an interval $[a,b]$, there are infinitely many different values to consider. We need a different way to think about it. We define the average value of a function using a limit of a standard average value.

Let f be a piecewise continuous function on $[a,b]$ so that the definite integral $\int_a^b f(x)\,dx$ will be defined. Consider a uniform partition of the interval $[a,b]$ with $\Delta x = \frac{b-a}{n}$ and $x_k = a + k\cdot\Delta x$, just as we defined when creating a Riemann sum. We *approximate* the average value of f on the interval $[a,b]$, which is represented by the symbol $\langle f\rangle_{[a,b]}$, by finding the average of the values $f(x_k)$ for $k = 1,2,\ldots,n$:

$$\langle f\rangle_{[a,b]} \approx \frac{1}{n}\sum_{k=1}^{n} f(x_k).$$

The approximation is improved with larger and larger values of n, so the actual average value will be the limit of the average as $n\to\infty$:

$$\langle f\rangle_{[a,b]} = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} f(x_k).$$

The average value defined by this limit looks remarkably similar to the limit of a Riemann sum that would define a definite integral. In particular,

$$\int_a^b f(x)\,dx = \lim_{n\to\infty}\sum_{k=1}^{n} f(x_k)\frac{b-a}{n}.$$

Comparing our two limits reveals that the definite integral is the same as the average value multiplied by $b-a$. This will be our formal definition of the average value of a function.

Definition 5.6.1 For a function $f(x)$ which has an integral on an interval

$[a, b]$, the **average value** of $f(x)$ on $[a, b]$ is defined by

$$\langle f(x)\rangle_{[a,b]} = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

$\Diamond$

When we think of the definite integral as the total signed area between the graph $y = f(x)$ and the axis $y = 0$, the average value can be interpreted as the value of a constant function that would have the same signed area. This is a consequence of writing

$$\int_a^b f(x)\,dx = (b-a)\cdot\langle f(x)\rangle_{[a,b]}.$$

The average value multiplied by $b-a$ equals the total signed area of $f(x)$ from a to b, so we can think of $(b-a)\cdot\langle f(x)\rangle_{[a,b]}$ as the signed area of a rectangle with a vertical position given by the average value.

Imagine the region below the graph $y = f(x)$ and between the lines $x = a$ and $x = b$ as if it were frozen water. If the ice melted but was trapped between the vertical lines $x = a$ and $x = b$, the high regions of the ice would melt and fill the valleys until the water level was flat. The resulting flat value is equivalent to the average value of the original function. Any regions of area above the average value line are used to fill an equivalent amount of area missing below the line.

Example 5.6.2 Consider finding the average value of $f(x) = x^2 - 2$ on the interval $[0, 2]$. The average value is computed by dividing the definite integral of $f(x) = x^2 - 2$ as x goes from 0 to 2 by the length of the interval.

$$\begin{aligned}\langle x^2\rangle_{[0,2]} &= \frac{1}{2-0}\int_0^2 x^2 - 2\,dx\\ &= \frac{1}{2}\left[\frac{1}{3}x^3 - 2x\right]_0^2\\ &= \frac{1}{2}\left(\frac{8}{3} - 2(2)\right) - \frac{1}{2}\left(\frac{0}{3} - 2(0)\right)\\ &= \frac{1}{2}\left(\frac{-4}{3}\right) = -\frac{2}{3}\end{aligned}$$

Having found the average value, we now create a graph of $y = f(x) = x^2 - 2$ together with the graph $y = \langle f(x)\rangle_{[0,2]} = -\frac{2}{3}$, as shown in Figure 5.6.3. The graph of the average value is the horizontal line. We can see that the area above the average value is matched by the unshaded area below the average value line and above the function.

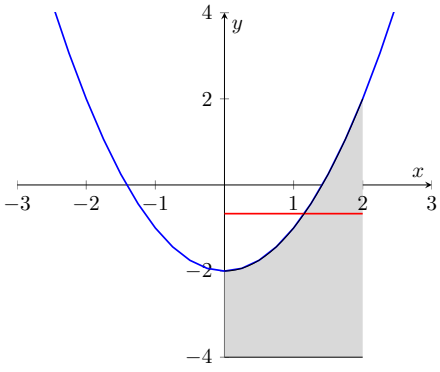


Figure 5.6.3 The graph of $y = x^2 - 2$ along with its average value over $[0, 2]$.

□

We need one more theorem before we discuss the Fundamental Theorem. That theorem is called the Mean Value Theorem for Definite Integrals. The phrase **mean value** is equivalent to **average value**, just as the mean of a set of numbers is equivalent to the average of those numbers. We have previously pointed out that the average or mean value of a function over an interval is equal to the constant value (horizontal function) that would have the same integral.

In Figure 5.6.3, we can see that the function actually crosses the line representing average value. As an equation, this point of intersection corresponds to a solution of the equation

$$f(x) = \langle f(x) \rangle_{[0,2]} \qquad \Leftrightarrow \qquad x^2 - 2 = -\frac{2}{3}.$$

When a function is continuous, such an intersection point will always occur. If a function is not continuous, it is possible for the function to skip over its average value without such an intersection.

Theorem 5.6.4 Mean Value Theorem for Definite Integrals. *If a function f is continuous on $[a, b]$, then there is a value $c \in (a, b)$ so that*

$$f(c) = \langle f \rangle_{[a,b]} = \frac{1}{b-a} \int_a^b f(x)\, dx.$$

Equivalently, the equation can be rewritten

$$f(c) \cdot (b - a) = \int_a^b f(x)\, dx.$$

Proof. Because f is continuous on $[a, b]$ (a closed interval), the Theorem 5.4.14 guarantees that f has an absolute maximum value $f_{\max}$ and an absolute minimum value $f_{\min}$ inside the interval. The average value must be between the maximum and minimum values. Indeed, for all $x \in [a, b]$ we have

$$f_{\min} \leq f(x) \leq f_{\max}.$$

Definite integrals preserve the ordering so that

$$f_{\min} \cdot (b - a) \leq \int_a^b f(x)\, dx \leq f_{\max} \cdot (b - a),$$

and dividing by the length of the interval $b - a$ we have

$$f_{\min} \leq \langle f(x) \rangle_{[a,b]} \leq f_{\max}.$$

The Theorem 5.4.18 then guarantees that between the x-values where the extremes occur we must have at least one value $x = c$ where $f(x) = \langle f(x) \rangle_{[a,b]}$. ■

The Mean Value Theorem for Definite Integrals will give us a tool with which we can replace a definite integral by a corresponding value of the integrand or rate function at some point within the interval times the length of that interval.

5.6.4 The Fundamental Theorem of Calculus

Now, are you ready to be blown away? Having learned what the average value of a function over an interval represents—the constant height that would

give the same signed area over the interval— we can discover that there is a relationship between the ideas of average rate of change and average value.

Let's think about an accumulation function, A with its corresponding rate of accumulation A'. What is the average rate of change of $A(x)$ as x goes from a to b?

$$\frac{\Delta A}{\Delta x}\Big|_{a,b} = \frac{A(b)-A(a)}{b-a} = \frac{1}{b-a}\left(A(b)-A(a)\right).$$

Because A is an accumulation, the change ΔA can be rewritten using an integral and

$$\frac{\Delta A}{\Delta x}\Big|_{a,b} = \frac{1}{b-a}\int_a^b A'(x)\,dx.$$

But that is just the average value of the rate of accumulation function A'. Two completely different ideas of average value end up measuring the very same thing.

Theorem 5.6.5 *The average rate of change (the difference quotient) of an accumulation function $A(x)$ is exactly equal to the average value (the integral divided by interval length) of the corresponding rate of accumulation $A'(x)$:*

$$\frac{\Delta A}{\Delta x}\Big|_{a,b} = \langle A' \rangle_{[a,b]}.$$

Pay attention, however, that we are thinking about two different functions. The average value of the rate of accumulation is based on the integral of the rate $A'(x)$. The average rate of change uses the rate of change based on the difference quotient using $A(x)$. The equivalence of these two averages provides exactly what is necessary to compute the derivative of an accumulation function.

Theorem 5.6.6 The Fundamental Theorem of Calculus, Part One (FTC1). *Given any function $f(x)$ that is continuous on an interval I and a value $a \in I$. The accumulation function*

$$A(x) = \int_a^x f(z)\,dz$$

is differentiable and $\frac{dA}{dx}(x) = f(x)$ for all $x \in I$.

Proof. Given the accumulation function $A(x)$ and its associated integrand $f(x)$, we consider the average rate of change of A between x and $x+h$. By Theorem 5.6.5, we can rewrite this in terms of the value of the rate function f,

$$\frac{\Delta A}{\Delta x}\Big|_{x,x+h} = \frac{A(x+h)-A(x)}{h} = f(c_h),$$

for some value c_h between x and $x+h$. The symbol c_h includes the h to emphasize that this value depends on h.

Now consider a sequence of values $h \to 0$. Because c_h is between x and $x+h$ for all h, the corresonding sequence c_h also converges, $c_h \to x$. As f is a continuous function,

$$\frac{dA}{dx}(x) = \lim_{h\to 0}\frac{\Delta A}{\Delta x}\Big|_{x,x+h} = \lim_{h\to 0} f(c_h) = f(x).$$

■

What have we shown? Earlier, when discussing accumulation functions,

$$A(x) = A(a) + \int_a^x f(z)\,dz,$$

we learned to *identify* the rate function based on its appearance within the definite integral and we wrote $A'(x) = f(x)$ as a way of describing this association. In that association, the prime (apostrophe) of A' was telling us to *identify the appropriate rate of accumulation.*

Now we have learned about another concept, the derivative, which is defined as the limiting value of the average rate of change of a function f from the input value of interest x and a second point $x + h$ as $h \to 0$. This is a fundamentally different concept from accumulation defined as the limit of a Riemann sum. Nevertheless, when we compute the derivative of an accumulation function, we recover exactly that function's corresponding rate of accumulation, so long as the rate of accumulation is a continuous function.

The rate of accumulation and the derivative are really different perspectives of the same function. This surprisingly deep relationship between definite integrals and derivatives will continue to develop.

5.6.5 Summary

- For simple polynomials which we previously learned to express as accumulation functions, the rate of accumulation seems miraculously to agree with the derivative of the polynomial.
- The Fundamental Theorem of Calculus (FTC1) shows us that this isn't circumstance but will always happen when the rate of accumulation is a continuous function. That is, the derivative of an accumulation function will equal the corresponding rate function. When we write $f'(x)$, that means *both* the rate of accumulation when $f(x)$ is an accumulation function *and* the derivative of a function $f(x)$ because those are ultimately the same thing.
- We can compute the **average value of a function** on an interval $[a, b]$ using a definite integral,

$$\langle f(x) \rangle_{[a,b]} = \frac{1}{b-a} \int_a^b f(x)\, dx.$$

  The integral replaces the idea of adding a list of values and dividing by the length of the interval replaces the idea of dividing by the number of values being added.
- The average rate of change of an accumulation function $f(x)$ and the average value of the rate of accumulation $f'(x)$ for that function are equal to each other.
- The Mean Value Theorem for Definite Integrals guarantees that for a continuous function, the equation $f(c) = \langle f(x) \rangle_{[a,b]}$ has a solution for some value $c \in (a, b)$. It allows us to substitute

$$\int_a^b f(x)\, dx = f(c) \cdot (b - a)$$

  for some c between a and b, but does not tell us how to find c.

5.6.6 Exercises

Find the average value of the given function over the given interval. Sketch a graph of the function and the average value over the interval. Then solve for c so that $f(c) = \langle f(x) \rangle_{[a,b]}$, if it exists.

1. $f(x) = 4 + 2x$ with $[a, b] = [0, 2]$.

2. $f(x) = 4x - x^2$ with $[a, b] = [0, 2]$.

3. $f(x) = 4x - x^2$ with $[a, b] = [0, 4]$.

4. $f(x) = \begin{cases} 1, & 0 \le x < 2 \\ 3, & 2 \le x \le 3 \end{cases}$ with $[a, b] = [0, 3]$. You will need to split the integral into two intervals. (Hint: Think about the graph geometrically.)

5. $f(x) = \begin{cases} 2x, & 0 \le x < 2 \\ 6 - x, & 2 \le x \le 4 \end{cases}$ with $[a, b] = [0, 4]$. You will need to split the integral into two intervals. (Hint: Think about the graph geometrically.)

6. $f(x) = \begin{cases} x, & 0 \le x < 2 \\ 5 - x, & 2 \le x \le 3 \end{cases}$ with $[a, b] = [0, 3]$. You will need to split the integral into two intervals. (Hint: Think about the graph geometrically.)

Applying the Fundamental Theorem of Calculus.

7. Find $\frac{dF}{dx}(x)$ where $F(x) = 10 + \int_1^x z^2 - 3z \, dz$. Then give the equation of the tangent line at $x = 1$.

8. Find $\frac{dG}{dx}(x)$ where $G(x) = \int_0^x \frac{1}{z^2 + 4} \, dz$. Then give the equation of the tangent line at $x = 0$.

9. Find $\frac{dH}{dx}(x)$ where $H(x) = 3 + \int_2^x z e^{-z^2} \, dz$. Then give the equation of the tangent line at $x = 2$.

Applications of Average Value

10. The density (kilograms per meter) of a rod that is two meters long depends on position along the rod according to the equation

$$\rho(x) = 2 - 0.25x, \quad 0 \le x \le 2.$$

Find the average density of the rod.

11. A car accelerates from 0 to 64 miles per hour over eight seconds so that the velocity of the car is a function of time given by

$$v(t) = 16t - t^2, \qquad 0 \le t \le 8.$$

What is the average velocity of the car during those eight seconds? How far does the car travel? (Hint: Use Theorem 5.6.5 and pay attention to units.)

12. During a rainstorm, the rate R (inches per hour) at which rain fell varied according to the following relation,

$$R(t) = \begin{cases} 2t, & 0 \le t < 0.25, \\ 0.5, & 0.25 \le t < 0.5, \end{cases}$$

where t is measured in hours. What is the average rate at which rain fell during the storm? What was the total amount of rain that fell

during the storm? (Hint: Use Theorem 5.6.5)

13. When the state police measure vehicle speed from aircraft, one approach of determining a car's speed is to time how long it takes to travel a fixed distance, say a quarter mile. Suppose that you were timed and the state police recorded 11.25 seconds. They charge you with speeding at 80 mph. If the police never actually recorded your exact speed, how can they guarantee that you must have been speeding? (Hint: Use Theorem 12.5.5

Part II

Chapter 6

Accumulation and Integrals

6.1 Accumulation Functions and the Definite Integral

Overview. The concept of the **definite integral** can be motivated by the notion of accumulated change. When we learn about linear functions, the idea of a constant slope or rate of change serves as the fundamental concept. For a definite integral, we generalize this notion to a changing rate.

In this section, we begin with an example of linear functions and piecewise linear functions as models of accumulation. Using these examples, we establish some basic principles that we want to hold in general. These principles become the fundamental properties of the definite integral.

6.1.1 Linear Functions as Accumulation

The word accumulation is defined as "the acquisition or gradual gathering of something" (Oxford Dictionary, , accessed August 27, 2019). Consider a tank of water that has water added at a constant rate of $5\,\frac{\text{L}}{\text{min}}$. At the start of our observation, the tank contains 200 L of water. We wish to think of the amount of water in the tank as an accumulation of the water that has flowed into the tank as a function of time.

Quantities that have a constant rate of change are modeled with linear functions, and the rate of change is used as the slope. If V is the volume of water that the tank contains (in liters) and t is the time of observation (in minutes), then the state of the tank is given by (t, V). The equation that models the accumulation is then given by

$$V = 200 + 5t.$$

The V-intercept of 200 represents the starting value (when $t = 0$), and the product $5t$ represents the accumulation of additional water that is added during the interval of time $(0, t)$.

The point-slope equation of a line similarly captures the idea of accumulation. Suppose after 10 minutes, the water flowing into the tank stops and water begins to drain at a rate of $15\,\frac{\text{L}}{\text{min}}$. We can use our earlier model to find the volume of water after the 10 minutes of filling has completed,

$$V = 200 + 5(10) = 250.$$

This becomes a new initial value for the tank relative to the draining, which corresponds to a negative rate of change or slope. For $t > 10$, we have a new model,

$$V = 250 - 15(t - 10).$$

The expression $-15(t - 10)$ represents the accumulated loss of water. We multiply the rate -15 by the *increment* of time $t - 10$, since that is how long the tank was left to drain.

Putting our models together, we obtain a piecewise function that represents the accumulation of water in the tank.

$$V = \begin{cases} 200 + 5t, & 0 \le t \le 10, \\ 250 - 15(t - 10), & t > 10. \end{cases}$$

We can think of the rate of accumulation R as another variable, which is also piecewise,

$$R = \begin{cases} 5, & 0 < t < 10, \\ -15, & t > 10. \end{cases}$$

We do not define R when $t = 10$ because of the ambiguity in how the transition occurs. Because R is the rate of accumulation corresponding to the accumulation V, we write $R = V'$ (read V-prime). We will later learn that R is the derivative of V at points where R is continuous.

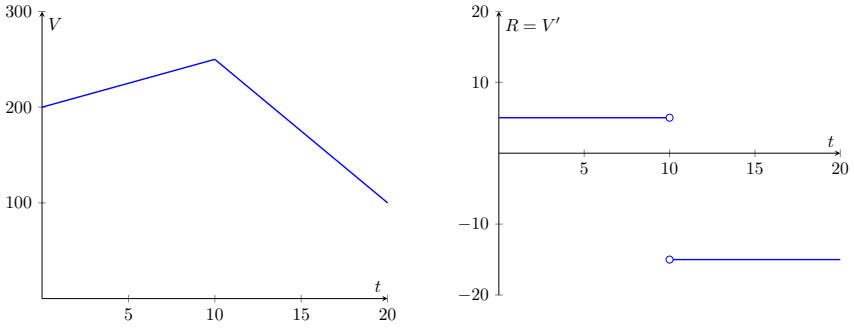


Figure 6.1.1 The volume of the tank of water that fills for 10 minutes and then drains, and the corresponding rate of accumulation, as functions of time.

Given any piecewise constant rate of accumulation $f(x) = A'(x)$ for an accumulation $A(x)$, we can easily compute the formulas for $A(x)$ as a piecewise linear function. We repeatedly apply the point-slope equation of a line and require that $A(x)$ is continuous at each transition point. This will then help motivate some general properties that will relate to the definite integral.

Example 6.1.2 Suppose $f(x) = A'(x)$ is defined as

$$f(x) = \begin{cases} 3, & x < 2, \\ -2, & 2 < x < 5, \\ 5, & x > 5. \end{cases}$$

If $A(0) = 2$, find $A(x)$ as a piecewise function.

Solution. Because the initial value is given as $A(0) = 2$, we begin our construction at $x = 0$. This point on the domain is inside the interval $x < 2$, so we start with a rate $A' = 3$. The formula for $A(x)$ with $x < 2$ is therefore

$$A(x) = 2 + 3x, \quad x < 2.$$

So that $A(x)$ is continuous, we must have $A(2) = 2 + 3(2) = 8$.

Having found the value of $A(x)$ on the interval $(-\infty, 2]$, we next consider the interval $(2, 5)$ where $A' = -2$. Using our value $A(2) = 8$ as an initial value, we can write

$$A(x) = 8 + -2(x - 2), \quad 2 < x < 5.$$

To have continuity at $x = 5$, we require $A(5) = 8 + -2(3) = 2$. Repeating the process on the last interval, $(5, \infty)$, where $A' = 5$, we obtain

$$A(x) = 2 + 5(x - 5), \quad x > 5.$$

Putting the pieces together, we obtain our final piecewise representation of $A(x)$:

$$A(x) = \begin{cases} 2 + 3x, & x \leq 2, \\ 8 + -2(x - 2), & 2 < x \leq 5, \\ 2 + 5(x - 5), & x > 5. \end{cases}$$

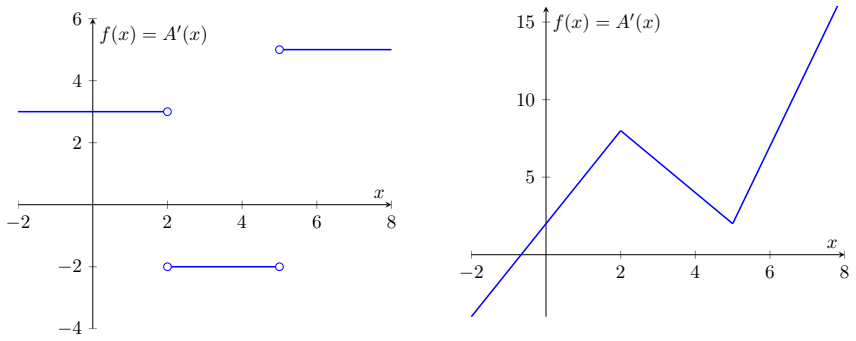


Figure 6.1.3 Graphs of the piecewise constant rate of accumulation $f(x) = A'(x)$ and the piecewise linear accumulation $A(x)$.

□

6.1.2 A Geometric Interpretation of Accumulation

In our earlier example (Example 6.1.2), we had the point $A(0) = 2$ and a rate $f(x) = 3$ for $x < 2$. When we used the point-slope formula to find $A(2)$, we had

$$A(2) = 2 + 3(2) = 8.$$

Then, having found $A(2) = 8$ and knowing $f(x) = -2$ for $2 < x < 5$, we were able to compute $A(5)$ as

$$A(5) = 8 + -2(5 - 2) = 2.$$

In each case, we took a known starting value ($A(0) = 2$ or $A(5) = 8$) and then added an *increment of change*. With a constant rate of accumulation, these increments were calculated as the rate of change times the increment of change in the independent variable, Δx.

Expressing the increment of change as a product of two values has a useful geometric interpretation. The most basic geometric idea that is calculated as a product of two numbers is area. Can we interpret the increment of change as an area? Almost. An area is always a positive number, but our second increment of change $-2(3) = -6$ was a negative value. So we modify our idea to **signed area**.

How does the area geometrically appear? Consider the graph of the rate of accumulation, $y = f(x)$. The rate of change corresponds to the height of the graph from the axis. We should soon recognize that there are rectangles from which we can find the signed area. When the graph is below the axis, we have a signed height that is considered negative. When the graph is above the axis, the signed height is positive. The increment of x depends on which direction we are going. We compute

$$\Delta x = x_{\text{end}} - x_{\text{start}}.$$

Consequently, if the increment moves to the right, we have $\Delta x > 0$; if the increment moves to the left, we have $\Delta x < 0$. The signed area is simply the product of the signed height times the signed increment of x.

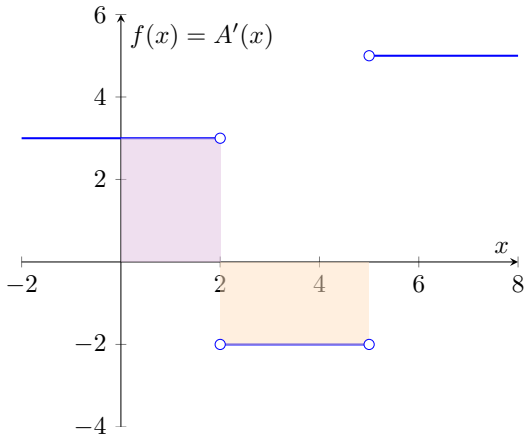


Figure 6.1.4 A graph of the rate of accumulation function $y = f(x)$ showing the increments of change as signed areas, shaded in color. The increment as x goes from $x = 0$ to $x = 2$ is shaded in purple. The increment as x goes from $x = 2$ to $x = 5$ is shaded in orange.

We will generalize the idea of accumulation from constant rates of change to arbitrary rates of change using signed area. Suppose we are given the graph of a function $f(x)$ that is a rate of accumulation $f(x) = A'(x)$ for some quantity $A(x)$. We will require the function to be piecewise continuous and have no infinite limits. The increment of change for $A(x)$ as x goes from $x = a$ to $x = b$ will be equal to the sum of the signed areas formed by the regions between $y = f(x)$ and the axis $y = 0$. This increment of change will be represented by a definite integral,

$$\Delta A = A(b) - A(a) = \int_a^b f(x)\,dx.$$

The notation for a definite integral is meant to be suggestive of this interpretation. The integral symbol $\int$ is drawn to look like the letter S to represent summation. The limits of integration a and b indicate the value for x where we start on the bottom the the value for x where we end on the top. What do we add? The increments $f(x)\,dx$ that are being accumulated. The expression $f(x)$ is the function giving the rate of accumulation and symbolically represents the signed height of incremental rectangles. The symbol dx is called the **infinitesimal** and symbolically represents the signed increment of the independent variable or width of the rectangle. When $b > a$, we are integrating to the right and $dx > 0$; when $b < a$, we are integrating to the left (reverse) and $dx < 0$.

When the shape of the graph of $f(x)$ uses straight line segments or other simple geometric shapes, we can calculate the signed area using simple geometric formulas.

Area Formulas for Common Geometric Regions.

- Rectangle, length ℓ and width w.

$$A = \ell\, w$$

- Triangle, base b and height h (perpendicular to base).

$$A = \tfrac{1}{2} b\, h$$

- Parallelogram, base b and height h (perpendicular to base).
$$A = b\,h$$
- Trapezoid, parallel sides b_1 and b_2 and height h (perpendicular to parallel sides).
$$A = \tfrac{1}{2}(b_1 + b_2)\,h$$
- Circle, radius r.
$$A = \pi r^2$$

We now have two different interpretations of the definite integral. First, the definite integral is the total accumulated change where the function in the integral is the rate of accumulation $f(x)$. Second, the definite integral is the sum of the signed areas between the graph of $f(x)$ and the axis. If we can calculate the definite integral, such as by geometric formulas for area, then we can interpret that value as the total accumulated change. This can allow us to compute additional values of the accumulation function:

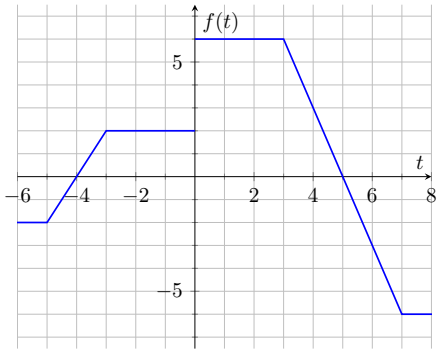
$$A(b) = A(a) + \int_a^b f(x)\,dx.$$

Notice how this equation for the accumulation function is similar to the point-slope equation of a linear function with slope m,

$$f(x) = f(a) + m(x - a).$$

There is an initial value. To that initial value, we add the increment of change. For a linear function, the rate of accumulation is constant and the increment is $m(x - a)$ For non-constant rates of accumulation, the increment of change is given by an integral.

Example 6.1.5 Consider the graph of the function $f(t)$ shown below. Suppose that $f(t)$ is the rate of accumulation for $A(t)$, $f(t) = A'(t)$. If we know $A(2) = 5$, find the values for $A(6)$ and $A(-3)$.



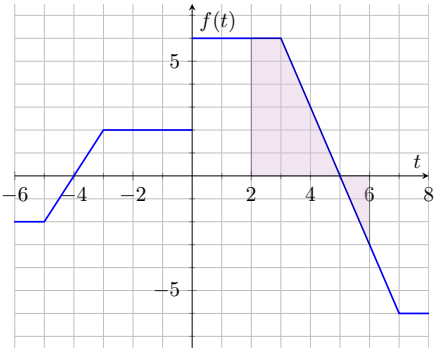
Solution. To find $A(6)$, we start at the known value $A(2) = 5$ and add the accumulated increment of change as t goes from 2 to 6. That is, using the notation of a definite integral, we have

$$A(6) = A(2) + \int_2^6 f(t)\,dt.$$

We will calculate this definite integral using geometric shapes.

We identify our shapes by considering the regions vertically between the graph of $f(t)$ and the t-axis. At $t = 2$, the graph is above the axis and remains

above the axis until $t = 5$. Our first region consists of a trapezoid bounded by $t = 2$ on the left, the t-axis below, and the graph of $f(t)$ above and to the right. From $t = 5$ to $t = 6$, the graph of $f(t)$ is below the axis. The second region consists of a triangle bounded by the t-axis on the top, the line $t = 6$ on the right, and the graph of $f(t)$ on the left. These shapes are illustrated in the graph below as shaded regions.



Having identified the relevant regions, we now calculate their signed area. We first recall that the direction of t is left to right, so that horizontal signed lengths are positive. Vertical signed lengths depend on whether we are above (positive) or below (negative) the axis. The trapezoid between $t = 2$ and $t = 5$ has parallel bases of signed length 1 (top) and 3 (bottom) and a perpendicular height of 5 units. The resulting area for this trapezoid is

$$\text{area}_1 = \tfrac{1}{2}(1 + 3)(5) = 10.$$

The triangle has a base of signed length 1 and a height of signed length -3, with corresponding area

$$\text{area}_2 = \tfrac{1}{2}(1)(-3) = -\tfrac{3}{2}.$$

Note that we could instead have used a single rectangle for $t = 2$ to $t = 3$ and a triangle for $t = 3$ to $t = 5$ in place of the trapezoid.

The total accumulated increment of change is the sum of the signed areas,

$$\int_2^6 f(t)\,dt = 10 + -\tfrac{3}{2} = \tfrac{17}{2}.$$

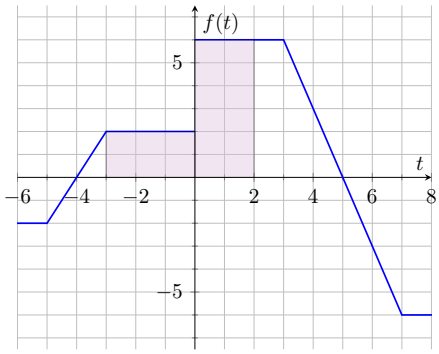
Consequently, we find

$$A(6) = A(2) + \int_2^6 f(t)\,dt = 5 + \tfrac{17}{2} = \tfrac{27}{2}.$$

To find $A(-3)$, we again start at the known value $A(2) = 5$ and add the accumulated increment of change as t goes from 2 to -3. Using a definite integral, we have

$$A(-3) = A(2) + \int_2^{-3} f(t)\,dt.$$

Because t is going backwards, our increments of t will be negative. When we calculate geometric signed areas, illustrated in the graph below, our horizontal edges will have negative signed lengths.



This time, the regions consist of two rectangles. The signed area for the rectangle from $t = 2$ to $t = 0$, formed by a horizontal edge with signed length -2 and a vertical edge with signed length 6, is

$$\text{area}_1 = (-2)(6) = -12.$$

The signed area for the second rectangle from $t = 0$ to $t = -3$, formed by a horizontal edge with signed length -3 and a vertical edge with signed length 2, is

$$\text{area}_2 = (-3)(2) = -6.$$

The total accumulated increment is the sum,

$$\int_2^{-3} f(t)\,dt = -12 - 6 = -18.$$

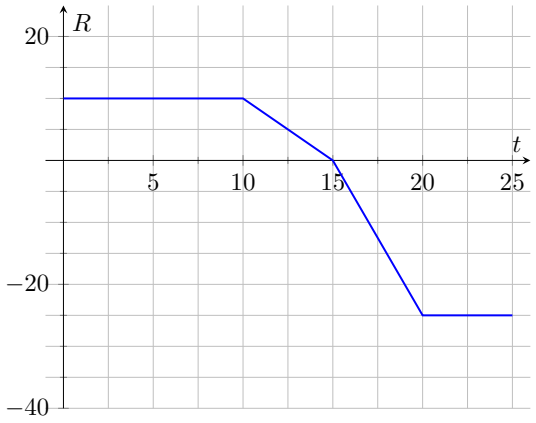
Consequently, we find

$$A(-3) = A(2) + \int_2^{-3} f(t)\,dt = 5 - 18 = -13.$$

□

Example 6.1.6 A large tank of water initially contains 400 liters of water. For ten minutes, water is added at a constant rate of 10 $\frac{\text{L}}{\text{min}}$. The rate of water flow then steadily declines for the next five minutes from 10 $\frac{\text{L}}{\text{min}}$ to 0 $\frac{\text{L}}{\text{min}}$. At this point, a pump starts draining the tank, ramping its progress over five minutes so that the rate of draining goes from 0 $\frac{\text{L}}{\text{min}}$ to 25 $\frac{\text{L}}{\text{min}}$. The pump then drains water at this steady rate of 25 $\frac{\text{L}}{\text{min}}$ for another five minutes. How much water is in the tank at the end of this procedure?

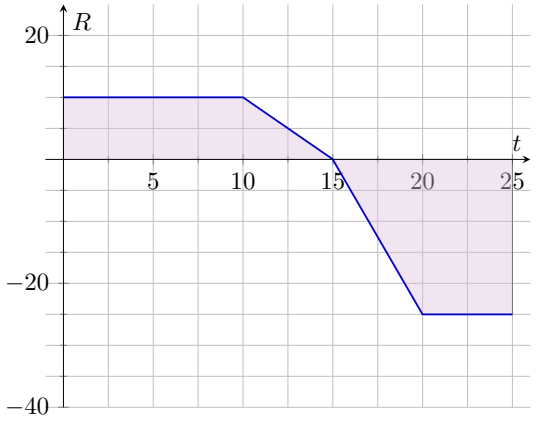
Solution. We start by describing the state variables. Let V represent the volume of water in the tank, measured in liters. Let R represent the rate of accumulation of water, which will be positive when water is flowing into the tank and negative when it is pumped out, measured in liters per minute. Let t represent the time since the situation begins, measured in minutes. As we read the description of the problem, we should note that most of the information is describing the rate of accumulation $R = V'$ for the volume of water in the tank. We can sketch a graph of R from the description, and we make the assumption that the description implies that this graph should be formed with line segments.



Now that we have a graph of the rate of accumulation $R(t) = V'(t)$, we can use geometric methods to calculate the area of regions to compute the total increment of change in the volume.

$$V(25) = V(0) + \int_0^{25} R(t)\,dt.$$

We consider the shaded regions in the graph below.



We could interpret the regions as either two trapezoids or as rectangles and triangles. The signed area of the trapezoid above the axis corresponding to times $t = 0$ to $t = 15$ is

$$\text{area}_1 = \tfrac{1}{2}(10 + 15)(10) = 125,$$

meaning that there were 125 liters added to the tank during the first 15 minutes. The signed area of the trapezoid below the axis corresponding to times $t = 15$ to $t = 25$ is

$$\text{area}_2 = \tfrac{1}{2}(10 + 5)(-25) = -\tfrac{375}{2} = -187.5.$$

This means that there were 187.5 liters drained from the tank during the last 10 minutes. Combining the results gives us the definite integral and overall increment of change in the volume of the tank. Starting with our initial tank level, we have

$$V(25) = V(0) + \int_0^{25} R(t)\,dt = 400 + (125 - 187.5) = 337.5.$$

The tank ends with 337.5 liters. □

6.1.3 Summary

- Piecewise constant rates of change correspond to continuous, piecewise linear functions.
- Given a function $f(x)$ that provides the rate of change for another function $A(x)$, the function A is called the accumulation function with rate f and the function f is called the rate of accumulation for A. We write $f(x) = A'(x)$.
- The change in an accumation function $A(x)$ as x goes from $x = a$ to $x = b$, calculated from A by $A(b) - A(a)$, is represented by a **definite integral** of its rate of accumulation,

$$A(b) - A(a) = \int_a^b f(x)\,dx.$$

- The definite integral calculates a sum of increments, each represented by the product $f(x)\,dx$.
- The geometric interpretation of the definite integral is the sum of the increments of signed area of regions bounded by the graph of the rate of accumlation $f(x)$ and the x-axis.

6.1.4 Exercises

1.

6.2 Properties of Definite Integrals

Overview. Motivated by the properties of total accumulated change and of area, the definite integral inherits several significant properties. These properties are stated as theorems. We will be interested in applying the results of the theorems. However, to prove these properties is outside the scope of this text. We essentially think of these properties as the axioms of definite integrals, basic properties which must always be true.

Because accumulation functions are defined in terms of definite integrals, we also develop properties of accumulation functions in terms of our knowledge of the rate of accumulation. We will learn how the sign of the rate of accumulation determines if the accumulation function is increasing or decreasing. We will also learn how the concavity of the accumulation function is related to the rate of accumulation.

6.2.1 Integrability

Before we talk about the properties of the definite integral, we need to establish some terminology about when the definite integral is even defined. From our introduction, we know that the definite integral will be defined if we can describe the geometric region as a finite number of rectangles and triangles. Such shapes will occur for functions that are defined piecewise constant or piecewise linear. However, more complicated functions might have potential issues.

One of the most significant developments of modern mathematics was developing an understanding of when functions can be integrated or not. Mathematicians would take an interpretation of the definite integral and then construct bizarre functions for which that interpretation would break. Then they would create new definitions for integrals that worked over progressively more complex circumstances. For our purposes, we will focus on the definition of the integral using limits of Riemann sums.

Definition 6.2.1 Integrability. A function $f : [a, b] \to \mathbb{R}$ is **integrable** (or, more simply, integrable) on $[a, b]$ if $\displaystyle\int_a^b f(x)\,dx$ is defined. $\Diamond$

The actual interpretation of when the definite integral is defined is described in (((Unresolved xref, reference "section-riemann-sums"; check spelling or use "provisional" attribute))) . The scope of mathematics for this text is not concerned with determining which functions are or are not integrable, with one exception. Continuous functions are integrable.

Theorem 6.2.2 Continuous Functions are Integrable. *If $f : [a, b] \to \mathbb{R}$ is continuous, then f is integrable on $[a, b]$.*

In fact, we get a result a little better than this. The function can have a finite number of discontinuities on the interval as long as left- and right-limits exist at all of the discontinuities. We do not allow infinite discontinuities at this stage of learning, as that will require a concept called improper integrals.

6.2.2 Splitting Properties

Consider any region in the plane for which we can find its area. Suppose we could cut the region, like we might cut a shape in two parts with scissors. The area of the original region would be the sum of the areas of the subregions created by our cut. This fact of geometric area motivates the **splitting** property of the definite integral.

Splitting properties are motivated by considering adjacent intervals, say $[a, b]$ and $[b, c]$, and requiring that the definite integral on $[a, c]$ is the sum of the integrals over the two pieces,

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx.$$

As described, this would seem to require that $a < b < c$. However, the definite integral is defined in a way that the order does not matter, so long as we replace our idea about intervals into directed excursions through an interval.

Recalling that the definite integral is motivated as the mathematical tool to compute the total change in a quantity as the accumulated change resulting from its rate of change, this result could be interpreted as saying, "The total change in Q as x goes from a to c is equal to the change in Q as x goes first from a to b plus the change as x then goes from b to c."

Theorem 6.2.3 Splitting Property of Definite Integrals. *Suppose that f is integrable on an interval that contains a, b and c. Then*

$$\int_a^b f(x)dx + \int_b^c f(x)dx = \int_a^c f(x)dx.$$

Similarly, if x does not change, then the dependent quantity Q should also not change, regardless of the function defining the rate of change. This is the motivation for the next theorem.

Theorem 6.2.4 Integral on an Empty Interval. *For any function f,*

$$\int_a^a f(x)\,dx = 0.$$

Combining these theorems, we obtain a reversal property of definite integrals. If we switch the order of the limits of integration, then the value of the definite integral must change sign.

Theorem 6.2.5 Integral in Reverse. *For any integrable function f,*

$$\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx.$$

Proof. Because an integral starting and ending at a must equal zero, if we go from a to b and then back, there must be no overall change:

$$\int_a^b f(x)\,dx + \int_b^a f(x)\,dx = \int_a^a f(x)\,dx = 0.$$

This means the two integrals are additive inverses to each other. ■

When we interpret a definite integral as signed area, we must take into account the direction of integration. The usual integration over an interval $[a, b]$ is interpreted as going from $x = a$ to $x = b$. Regions above the axis generate positive signed area; regions below the axis generate negative signed area. However, if we were to reverse direction, going from $x = b$ to $x = a$, then the signs are reversed. This potential stumbling block can be mitigated if we remember to think of dx as having a sign as well. Integrals going from left to right result in $dx > 0$; integrals going from right to left have $dx < 0$.

Example 6.2.6 Suppose that we know $\int_0^4 f(x)\,dx = 6$ and $\int_3^4 f(x)\,dx = 10$. Find $\int_0^3 f(x)\,dx$.

Solution. We use the splitting property of definite integrals. The interval $[0, 4]$ can be split into $[0, 3]$ and $[3, 4]$ so that

$$\int_0^4 f(x)\,dx = \int_0^3 f(x)\,dx + \int_3^4 f(x)\,dx.$$

We know two of the integrals and can solve for the third:

$$6 = \int_0^3 f(x)\,dx + 10 \qquad \Leftrightarrow \qquad \int_0^3 f(x)\,dx = -4.$$

An alternate approach for finding the integral is to start with the integral that is wanted, using the interval $[0, 3]$, so that we start at 0 and end at 3. We will use the splitting property using out-of-order points and go from 0 to 4 and then from 4 to 3:

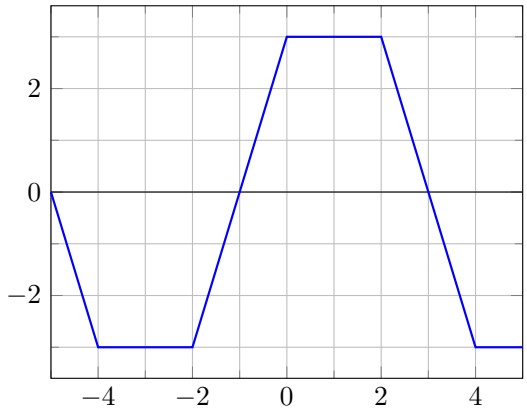
$$\int_0^3 f(x)\,dx = \int_0^4 f(x)\,dx + \int_4^3 f(x)\,dx.$$

The second integral is in a reversed order. If we switch the order to go left-to-right, then the integral is subtracted instead of added:

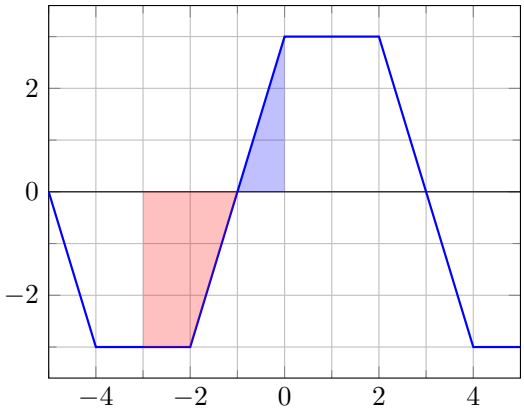
$$\int_0^3 f(x)\,dx = \int_0^4 f(x)\,dx - \int_3^4 f(x)\,dx = 6 - 10 = -4.$$

□

Example 6.2.7 Suppose that the graph below shows $y = f(x)$. Use the graph to find $\displaystyle\int_0^{-3} f(x)\,dx$.



Solution. Because the graph consists of straight lines, we can use geometry to calculate areas and use signed area to determine values of definite integrals. Shading the region between the graph $y = f(x)$ and the axis $y = 0$ and between $x = -3$ and $x = 0$, we get the figure shown below.



The region between $x = -3$ and $x = -1$ is a trapezoid that has area $\frac{1}{2}(1+2)(3) = \frac{9}{2}$. The region between $x = -1$ and $x = 0$ is a triangle with area $\frac{1}{2}(1)(3) = \frac{3}{2}$. Signed area corresponds to an integral from left-to-right so that

$$\int_{-3}^{0} f(x)\,dx = -\frac{9}{2} + \frac{3}{2} = -\frac{6}{2} = -3.$$

The integral of interest uses the opposite order, and so has the opposite sign:

$$\int_{0}^{-3} f(x)\,dx = -\int_{-3}^{0} f(x)\,dx = -(-3) = 3.$$

□

6.2.3 Summary

- **Integrability**: A function $f(x)$ is **integrable** on an interval if the definite integral $\displaystyle\int_a^b f(x)\,dx$ can be computed for any values a, b in the interval.

  We will need a clear definition for the definite integral to be more specific.

- Continuity implies integrability. If $f(x)$ is continuous on an interval, then it is guaranteed to also be integrable on that interval.

  In fact, so long as $f(x)$ is piecewise continuous with a finite number of discontinuities and $f(x)$ has one-sided limits at all of those discontinuities, then $f(x)$ will still be integrable. (The value at the end points don't matter to integrals.)

- **Splitting**: A definite integral $\displaystyle\int_a^b f(x)\,dx$ involves the independent variable x going through the values from $x = a$ (start) to $x = b$ (end). The splitting principle means that you can consider going from $x = a$ and take a diversion to any other point $x = c$, and then continue from $x = c$ to $x = b$ and get the same result,

  $$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx,$$

  so long as $f(x)$ is integrable on an interval that includes all three points a, b, c.

6.3 Calculating Integrals Using Accumulations

Overview. We may be accustomed from past experience to expect that every mathematics problem is meant to have a nice simple formula. Definite integrals are an example of a mathematical calculation that is the opposite. If someone were to write down a randomly created definite integral, there would typically be no formula using elementary functions that calculate its value. However, there are some rules that we can use in particular situations.

This section focuses on some elementary formulas that are known for definite integrals. These formulas are called accumulation formulas. In light of the Fundamental Theorem of Calculus, the formulas also correspond to what is known as antiderivative. The most basic accumulation rule is for simple power functions.

Definite integrals satisfy a sum rule and constant multiple rule. These rules allow us to compute definite integrals whose rates of accumulation are formed from simpler rates through summation or through multiplication by a constant. Together with the elementary accumulation formulas that we learn, these rules allow us to compute a useful collection of definite integrals.

6.3.1 Accumulation of Power Functions

The elementary power functions $\text{pow}_p(x) = x^p$ have a relatively simple associated accumulation function. One way to learn about the accumulation formula is to try the simplest examples and look to see if a pattern arises. This process is called forming a conjecture. Observed patterns are not always true patterns, so we will eventually need a more rigorous approach.

An accumulation function for a given integrable rate function $f(x)$ can be constructed by calculating the definite integral from a particular constant starting point to the value of the independent variable. We write

$$A(x) = \int_a^x f(z)\,dz,$$

where we use a variable of integration that is different from the x used as the independent variable.

The simplest rate function is a constant function, corresponding to a power $p = 0$, $f(x) = 1$. An elementary accumulation function would be the integral from 0 to x, corresponding to the signed area of a rectangle. Because the rectangle has a width height of $f(x) = 1$ and a width

$$A(x) = \int_0^x 1\,dz = x.$$

We know that

$$\int_a^b 1\,dx = A(b) - A(a) = b - a.$$

We can also easily find the definite integral of the identity function $f(x) = x$, which corresponds to a power $p = 1$. Because the graph of $y = f(z) = z$ is linear, the definite integral $\displaystyle\int_0^x z\,dz$ corresponds to the signed area of a triangle. Since the triangle has a base x and a height $f(x) = x$, the definite integral that depends on end point x has a value

$$A(x) = \int_0^x z\,dz = \frac{1}{2}x^2.$$

The definite integral from $x = a$ to $x = b$ can be computed as the change in accumulation,

$$\int_a^b z\,dz = A(b) - A(a) = \frac{1}{2}b^2 - \frac{1}{2}a^2.$$

Higher powers become more complicated and required increasingly complicated results. For the quadratic power function $f(x) = x^2$ with $p = 2$, the graph takes the form of a parabola. The Greek philosopher and mathematician Archimedes used a method of exhaustion to determine the area enclosed by a parabola. In the context of an accumulation function, we have

$$A(x) = \int_0^x z^2\,dz = \frac{1}{3}x^3.$$

For a cubic power function $f(x) = x^3$ ($p = 3$), an 11th century Arab mathematician Ibn al-Haytham during the Islamic Golden Age found a result that shows

$$A(x) = \int_0^x z^3\,dz = \frac{1}{4}x^4.$$

We can recreate Archimedes's and Ibn al-Haytham's results using limits of Riemann sums.

Checkpoint 6.3.1 Show that $\displaystyle\int_0^x z^2\,dz = \frac{1}{3}x^3$ and $\displaystyle\int_0^x z^3\,dz = \frac{1}{4}x^4$.

Solution. For both integrals, the interval of integration goes from $a = 0$ to $b = x$. When we make a partition with n uniform subintervals, the width of each subinterval will be

$$\Delta x = \frac{x-0}{n} = \frac{x}{n}.$$

In addition, the points in the partition will be defined by

$$x_k = \frac{kx}{n}, \quad k = 0, 1, \ldots, n.$$

The integral for $f(z) = z^2$ uses values $f(x_k) = \frac{k^2x^2}{n^2}$ to give a Riemann sum

$$\begin{aligned}\int_0^x z^2\,dz &\approx \sum_{k=1}^n f(x_k)\Delta x \\ &= \sum_{k=1}^n \frac{k^2x^2}{n^2}\cdot\frac{x}{n} \\ &= \sum_{k=1}^n \frac{k^2x^3}{n^3} \\ &= \frac{x^3}{n^3}\cdot\sum_{k=1}^n k^2\end{aligned}$$

When we use the closed formula for the sum of squares and take a limit, we find

$$\begin{aligned}\int_0^x z^2\,dz &= \lim_{n\to\infty}\sum_{k=1}^n f(x_k)\Delta x \\ &= \lim_{n\to\infty}\frac{x^3}{n^3}\sum_{k=1}^n k^2 \\ &= \lim_{n\to\infty}\frac{x^3}{n^3}\cdot\frac{n(n+1)(2n+1)}{6}\end{aligned}$$

$$\begin{aligned}
&= \lim_{n\to\infty} \frac{x^3}{6} \cdot \frac{2n^3+3n^2+n}{n^3} \\
&= \lim_{n\to\infty} \frac{x^3}{6} \cdot (2 + \frac{3}{n} + \frac{1}{n^2}) \\
&= \frac{x^3}{6} \cdot (2+0+0) \\
&= \frac{x^3}{3}
\end{aligned}$$

The integral for $f(z) = z^3$ is similar but with $f(x_k) = \frac{k^3x^3}{n^3}$ to give a Riemann sum

$$\begin{aligned}
\int_0^x z^3\,dz &\approx \sum_{k=1}^{n} f(x_k)\Delta x \\
&= \sum_{k=1}^{n} \frac{k^3x^3}{n^3} \cdot \frac{x}{n} \\
&= \sum_{k=1}^{n} \frac{k^3x^4}{n^4} \\
&= \frac{x^4}{n^4} \cdot \sum_{k=1}^{n} k^3
\end{aligned}$$

Applying the formula for the sum of cubes and a limit, we find

$$\begin{aligned}
\int_0^x z^3\,dz &= \lim_{n\to\infty} \sum_{k=1}^{n} f(x_k)\Delta x \\
&= \lim_{n\to\infty} \frac{x^4}{n^4} \sum_{k=1}^{n} k^3 \\
&= \lim_{n\to\infty} \frac{x^4}{n^4} \cdot \frac{n^2(n+1)^2}{4} \\
&= \lim_{n\to\infty} \frac{x^4}{4} \cdot \frac{n^4+2n^3+n^2}{n^4} \\
&= \lim_{n\to\infty} \frac{x^4}{4} \cdot (1 + \frac{2}{n} + \frac{1}{n^2}) \\
&= \frac{x^4}{4} \cdot (1+0+0) \\
&= \frac{x^4}{4}
\end{aligned}$$

We start to look for possible patterns that might generalize these results. If we write the results side-by-side, we observe that each accumulation is also a power and with a pattern in the coefficient.

$$\begin{aligned}
\int_0^x z^0\,dz &= x^1 \\
\int_0^x z^1\,dz &= \frac{1}{2}x^2 \\
\int_0^x z^2\,dz &= \frac{1}{3}x^3 \\
\int_0^x z^3\,dz &= \frac{1}{4}x^4
\end{aligned}$$

The power of the accumulation is always one value higher than the power of the rate function. The coefficient multiplying the accumulation is then the reciprocal of that higher power. We summarize our conjecture as a power rule for integration.

Conjecture 6.3.2 *For any power $p > 0$, we will have*

$$\int_0^x z^p \, dz = \frac{1}{p+1} x^{p+1}.$$

Our conjecture will ultimately be shown to be correct. However, we will need some additional tools before we know how to prove that it is true. Riemann sums will not help because we would need a summation formula for each possible case. We will need a more general concept of accumulation called an antiderivative. The power rule will, in fact, give us an antiderivative for power functions of any power $p \neq -1$.

6.3.2 Accumulation Rules of Combination

Having found accumulation formulas for elementary power functions, we would like to find the definite integrals for more complicated functions. There are two basic rules of combination for definite integrals—the constant multiple and the sum rules.

Theorem 6.3.3 *If $f(x)$ is integrable on an interval containing a, b and k is a constant, then*

$$\int_a^b k\, f(x) \, dx = k \cdot \int_a^b f(x) \, dx.$$

Proof. Because $f(x)$ is integrable, we know

$$\lim_{n\to\infty} \sum_{k=1}^{n} f(x_k)\Delta x = \int_a^b f(x) \, dx.$$

If we then compute the Riemann sum for a rate $kf(x)$, each term is multiplied by k. Both the summation operator and the limit operator have a constant multiple rule, allowing the constant k to factor its way out.

$$\begin{aligned}
\int_a^b k\, f(x) \, dx &= \lim_{n\to\infty} \sum_{k=1}^{n} kf(x_k)\Delta x \\
&= \lim_{n\to\infty} k \sum_{k=1}^{n} f(x_k)\Delta x \\
&= k \lim_{n\to\infty} \sum_{k=1}^{n} f(x_k)\Delta x \\
&= k \int_a^b f(x) \, dx
\end{aligned}$$

■

Geometrically, the constant multiple rule is about rescaling the graph of $f(x)$ by a factor k. Because the width of the region is the same but every point has been stretched vertically by k, the signed area must be multiplied by k. The proof of the rule results because both limits and summation formulas have a constant multiple rule.

Theorem 6.3.4 *If $f(x)$ and $g(x)$ are each integrable on an interval containing a, b, then*

$$\int_a^b f(x) + g(x)\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx.$$

Proof. Because $f(x)$ and $g(x)$ are integrable, we know

$$\lim_{n\to\infty} \sum_{k=1}^{n} f(x_k)\Delta x = \int_a^b f(x)\,dx$$

and

$$\lim_{n\to\infty} \sum_{k=1}^{n} g(x_k)\Delta x = \int_a^b g(x)\,dx$$

If we then compute the Riemann sum for rate formed by the sum $f(x) + g(x)$, then the sum rules for the summation operator and for the limit operator allow us to add the integrals.

$$\begin{aligned}
\int_a^b f(x) + g(x)\,dx &= \lim_{n\to\infty} \sum_{k=1}^{n} \left(f(x_k) + g(x_k)\right)\Delta x \\
&= \lim_{n\to\infty} \left(\sum_{k=1}^{n} f(x_k)\Delta x + \sum_{k=1}^{n} g(x_k)\Delta x \right) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} f(x_k)\Delta x + \lim_{n\to\infty} \sum_{k=1}^{n} g(x_k)\Delta x \\
&= \int_a^b f(x)\,dx + \int_a^b g(x)\,dx.
\end{aligned}$$

■

Geometrically, the sum rule is thinking of the Riemann sum rectangles with rate $f(x)+g(x)$ and width Δx as being formed by two rectangles of width Δx. One rectangle has height $f(x)$ and the other has height $g(x)$. The total area is the sum $f(x)\Delta x + g(x)\Delta x$. When we add all of the increments, we can group all of the parts from $f(x)\Delta x$ to give a sum matching the integral $\displaystyle\int_a^b f(x)\,dx$. The sum of the other terms give $\displaystyle\int_a^b g(x)\,dx$.

With our two rules of arithmetic and the power rule for accumulations, we can compute the definite integrals of polynomials.

Example 6.3.5 Find $\displaystyle\int_2^4 3x^2 - 4\,dx$.

Solution. We start by looking at the integrand. The rate function $f(x) = 3x^2 - 4$ is a linear combination of the elementary powers x^2 and $x^0 = 1$, $f(x) = 3 \cdot x^2 + -4 \cdot 1$. The power rule gives us accumulation formulas for the powers, so we know

$$\begin{aligned}
\int_2^4 x^2\,dx &= \frac{1}{3}(4^3) - \frac{1}{3}(2^3) \\
&= \frac{64}{3} - \frac{8}{3} = \frac{56}{3}
\end{aligned}$$

$$\int_2^4 1\,dx = 4 - 2 = 2$$

Using the sum and constant multiple rules, we find

$$\begin{aligned}\int_2^4 3x^2 - 4\,dx &= 3\int_2^4 x^2\,dx + -4\int_2^4 1\,dx \\ &= 3\cdot\frac{56}{3} + -4\cdot 2 \\ &= 56 - 8 = 48\end{aligned}$$

□

In the previous example, we used the constant multiple and sum rule to break a definite integral of a linear combination into a corresponding linear combination of the values of the component definite integrals. However, it is usually more convenient to create an accumulation function that is a linear combination of the corresponding accumulations of the component rates. That is, a rate function $f(x) = 3x^2 - 4$ must have a corresponding accumulation function

$$A(x) = 3\cdot\frac{1}{3}x^3 + -4\cdot x = x^3 - 4x,$$

based on the accumulation functions for the rates x^2 and 1. The definite integral is then just the change in this new accumulation function.

Example 6.3.6 Find $\displaystyle\int_{-2}^{2} 2x^2 - 3x + 4\,dx$.

Solution. The rate function in then integrand is $f(x) = 2x^2 - 3x + 4$. The accumulations for elementary powers are $\frac{1}{3}x^3$ for x^2, $\frac{1}{2}x^2$ for x, and x for 1. Consequently, the accumulation for $f(x)$ is

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

The definite integral of the rate will equal the change in the accumulation,

$$\begin{aligned}\int_{-2}^{2} 2x^2 - 3x + 4\,dx &= A(2) - A(-2) \\ &= \left(\frac{2}{3}(2)^3 - \frac{3}{2}(2)^2 + 4(2)\right) - \left(\frac{2}{3}(-2)^3 - \frac{3}{2}(-2)^2 + 4(-2)\right) \\ &= \left(\frac{16}{3} - 6 + 8\right) - \left(\frac{-16}{3} - 6 - 8\right) \\ &= \frac{32}{3} + 16 = \frac{32+48}{3} = \frac{80}{3}\end{aligned}$$

□

6.3.3 Summary

-

6.4 Riemann Sums

Overview. When a rate of change is a simple function (piecewise constant), we can compute the definite integral as a summation of the increments. Each increment is the product of the rate of change times the width of the subinterval in the partition. When a rate of change is not simple (varying), we can approximate the total change by using simple functions that are either above or below the true rate. If we can make these approximations as good as we desire, then there is a limiting value and that value is defined as the definite integral.

The approximations to the definite integral using simple functions are called **Riemann sums**. In this section, we will learn how to create Riemann sums using a uniform partition. The Riemann sum will depend on the number of increments. The definite integral will be the limit of this sum as the number of increments goes to infinity.

6.4.1 Uniform Partitions

Recall that a partition P of an interval $[a, b]$ is an increasing, finite sequence $P = (x_0, x_1, \ldots, x_n)$ with $x_0 = a$ and $x_n = b$ and $x_{k-1} < x_k$. Adjacent terms in the sequence define subintervals, $I_k = [x_{k-1}, x_k]$, which has a width increment of size $\nabla x_k = x_k - x_{k-1}$. A uniform partition of an interval $[a, b]$ is a partition in which all increments are the same size.

Definition 6.4.1 Uniform Partition. The **uniform partition** of an interval $[a, b]$ of size n is the partition with equal increments

$$\nabla x_k = \Delta x = \frac{b-a}{n}.$$

The partition points are defined by the arithmetic sequence

$$x_k = a + k \cdot \Delta x, \quad k = 0, 1, \ldots, n.$$

The kth subinterval is $I_k = [a + (k-1)\Delta x, a + k\Delta x]$. ◊

The definition of the uniform partition suggests the basic steps required to create such a partition.

- Identify the interval $[a, b]$ and the size of partition n.
- Compute the partition increment size

  $$\Delta x = \frac{b-a}{n},$$

  which is the total width of the interval divided by the number of subintervals.
- Create the partition points by using an arithmetic sequence with initial value $x_0 = a$ and increment size Δx (just calculated).

Example 6.4.2 Find the uniform partition of $[1, 4]$ of size $n = 8$.

Solution. The interval uses $a = 1$, $b = 4$ and $n = 8$. Consequently we can compute

$$\Delta x = \frac{b-a}{n} = \frac{4-1}{8} = \frac{3}{8}.$$

Next, we define the partition points,

$$x_k = a + k\Delta x = 1 + k \cdot \frac{3}{8} = 1 + \frac{3}{8}k.$$

In particular, the partition includes the points shown in the table below.

k	x_k
0	1
1	$1 + \frac{3}{8} = 1\frac{3}{8}$
2	$1 + \frac{6}{8} = 1\frac{6}{8}$
3	$1 + \frac{9}{8} = 2\frac{1}{8}$
4	$1 + \frac{12}{8} = 2\frac{4}{8}$
5	$1 + \frac{15}{8} = 2\frac{7}{8}$
6	$1 + \frac{18}{8} = 3\frac{2}{8}$
7	$1 + \frac{21}{8} = 3\frac{5}{8}$
8	$1 + \frac{24}{8} = 4$

□

One of the tasks required in computing Riemann sums will involve evaluating a function at these partition points. This is yet another example of the importance of composition of functions in that we replace the independent variable (input) of the function with a formula for the partition point of interest.

Example 6.4.3 Evaluate $f(x_k)$ where $f(x) = x^2$ and x_k is a point in the uniform partition of $[1, 4]$ of size n.

Solution. We start by defining the partition. The interval $[1, 4]$ means that $a = 1$ and $b = 4$, as in the previous example. However, the size of the partition n is not specified, so we use the variable itself to compute the increment size,

$$\Delta x = \frac{b-a}{n} = \frac{4-1}{n} = \frac{3}{n}.$$

Once the increment is known, we can define the partition points which is an arithmetic sequence with $x_0 = 1$ and increments Δx to define

$$x_k = 1 + k\Delta x = 1 + k \cdot \frac{3}{n} = 1 + \frac{3k}{n}.$$

Once the formula for the partition point is known, we use composition with $f(x) = x^2$ and then expand the formula.

$$\begin{aligned} f(x_k) &= f(1 + \frac{3k}{n}) \\ &= [1 + \frac{3k}{n}]^2 \\ &= (1 + \frac{3k}{n})(1 + \frac{3k}{n}) \\ &= 1 + \frac{6k}{n} + \frac{9k^2}{n^2}. \end{aligned}$$

□

From this section, you should be able to write down the formula for the points in a uniform partition of an interval whether the size of the partition is given as a specific number or as an unspecified value. Using this formula, you should also be able to use that formula to evaluate a function at the partition points.

6.4.2 Uniform Riemann Sums

Recall that a simple function is a piecewise function that is constant on each subinterval defined by the partition. We know how to compute the accumulated change (definite integral) for every simple function. Suppose that we had any function $f(x)$ representing a rate of change of some quantity Q and we wanted to determine the resulting increment of change

$$Q(b) - Q(a) = \int_a^b f(x)\,dx$$

as x changes from $x = a$ to $x = b$. A **Riemann sum** approximates this definite integral by approximating the function $f(x)$ by a simple function defined on a partition of $[a, b]$.

A Riemann sum involves two steps: specifying the partition and choosing the simple function defined on the partition. The most common choice for a partition is a uniform partition. The simple function is defined by choosing a constant function value on each resulting subinterval. A Riemann sum requires that we choose the value to match the true function $f(x)$ at some point within the closed subinterval $[x_{k-1}, x_k]$. Different rules for how to choose the point define some common methods.

Left-Hand Rule The simple function uses the value at the left end point, $f(x_{k-1})$.

Right-Hand Rule The simple function uses the value at the right end point, $f(x_k)$.

Mid-Point Rule The simple function uses the value at the midpoint of the interval, $f(\frac{x_{k-1}+x_k}{2})$.

Trapezoid Rule The simple function uses the average of the values at the end-points, $\dfrac{f(x_{k-1}) + f(x_k)}{2}$.

Lower-Sum Rule The simple function uses the minimum value of the function on the subinterval, $\min(f(x) : x \in [x_{k-1}, x_k])$.

Upper-Sum Rule The simple function uses the maximum value of the function on the subinterval, $\max(f(x) : x \in [x_{k-1}, x_k])$.

The left-hand rule and the right-hand rule are the simplest rules to work with algebraically. We will focus on practicing using those rules. The trapezoid rule typically is a much better approximation and is preferred when using a computer. The lower-sum and upper-sum rules provide error bounds on our approximation. The lower-sum always underestimates the definite integral; the upper-sum always overestimates the value. If we know both the lower-sum and upper-sum, then the true value must be between them.

To compute a Riemann sum using a particular choice of simple function, we usually do not define the approximating simple function separately. We just compute the approximating definite integral based on that simple function. For clarity, our first example will define the function directly.

Example 6.4.4 Approximate $\displaystyle\int_2^5 x^2 dx$ using the left-hand rule with a uniform partition of size $n = 4$.

Solution. The first step is to define the partition. Our interval is $[a, b] =$

$[2, 5]$. Consequently, the increment size of the partition will be

$$\Delta x = \frac{5-2}{4} = \frac{3}{4}.$$

The partition points are defined by

$$x_k = 2 + \frac{3k}{4}, \quad k = 0, 1, 2, 3, 4.$$

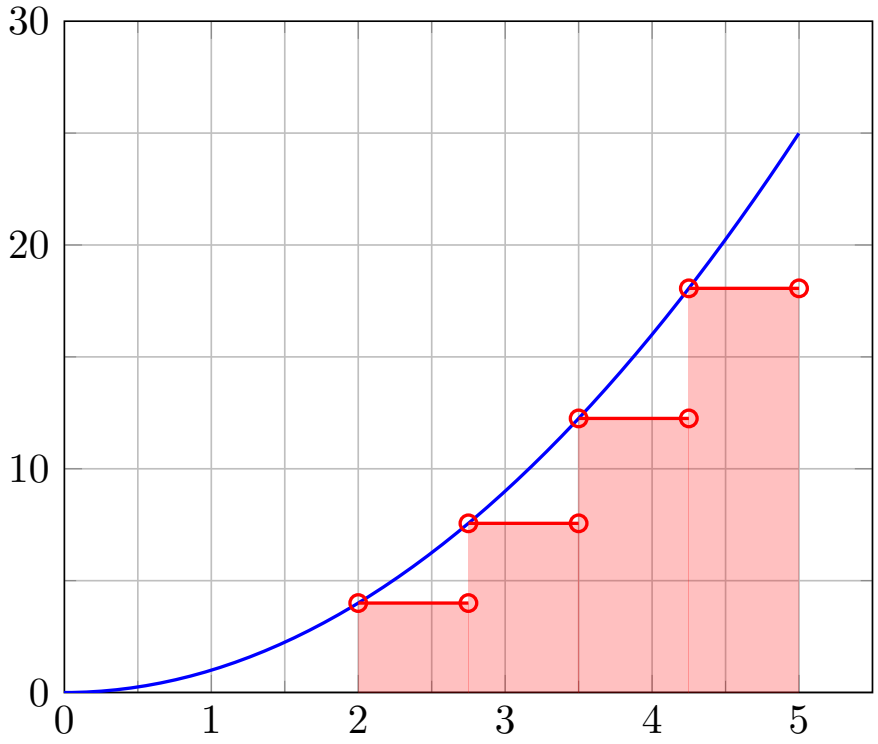
In particular, the partition is defined by the sequence

$$x = (2, 2.75, 3.5, 4.25, 5).$$

The second step is to define the simple function, $\mathcal{L}_f(x)$. Using the left-hand rule means that we will use the value of $f(x) = x^2$ on each subinterval $[x_{k-1}, x_k]$ by the value of $f(x_{k-1})$.

$$\mathcal{L}_f(x) = \begin{cases} f(2) = 2^2 = 4, & 2 < x < 2.75, \\ f(2.75) = 2.75^2 = 7.5625, & 2.75 < x < 3.5, \\ f(3.5) = 3.5^2 = 12.25, & 3.5 < x < 4.25, \\ f(4.25) = 4.25^2 = 18.0625, & 4.25 < x < 5. \end{cases}$$

A graph of $y = f(x)$ and the approximating simple function $y = \mathcal{L}_f(x)$ is shown below. The shaded region corresponds to the definite integral represented by the Riemann sum.



The Riemann sum is the definite integral of the approximating simple function. Notice how the limits of the integral correspond to the interval $[2, 5]$ while the limits of the sum correspond to counting the subintervals in the partition.

$$\int_2^5 \mathcal{L}_f(x)\, dx = \sum_{k=1}^{4} f(x_{k-1})\Delta x$$

$$
\begin{aligned}
&= f(2)\cdot\frac{3}{4} + f(2.75)\cdot\frac{3}{4} + f(3.5)\cdot\frac{3}{4} + f(4.25)\cdot\frac{3}{4}\\
&= 4(0.75) + 7.5625(0.75) + 12.25(0.75) + 18.0625(0.75)\\
&= 31.40625
\end{aligned}
$$

□

In usual practice, the only steps we really need are identifying the partition, determining the value of the function on each subinterval, and then computing the Riemann sum, which corresponds to the definite integral of the simple function. Writing down the piecewise formula for the simple function is not actually necessary. A table often makes the computation simpler.

Example 6.4.5 Use a Riemann sum with the right-hand rule and a uniform partition of size $n = 5$ to approximate $\displaystyle\int_0^2 \frac{1}{x+1}dx$.

Solution. Start by identifying the partition. First determine the increment size,

$$\Delta x = \frac{2-0}{5} = \frac{2}{5} = 0.4.$$

Use the increment size to find the partition,

$$x = (0, 0.4, 0.8, 1.2, 1.6, 2.0).$$

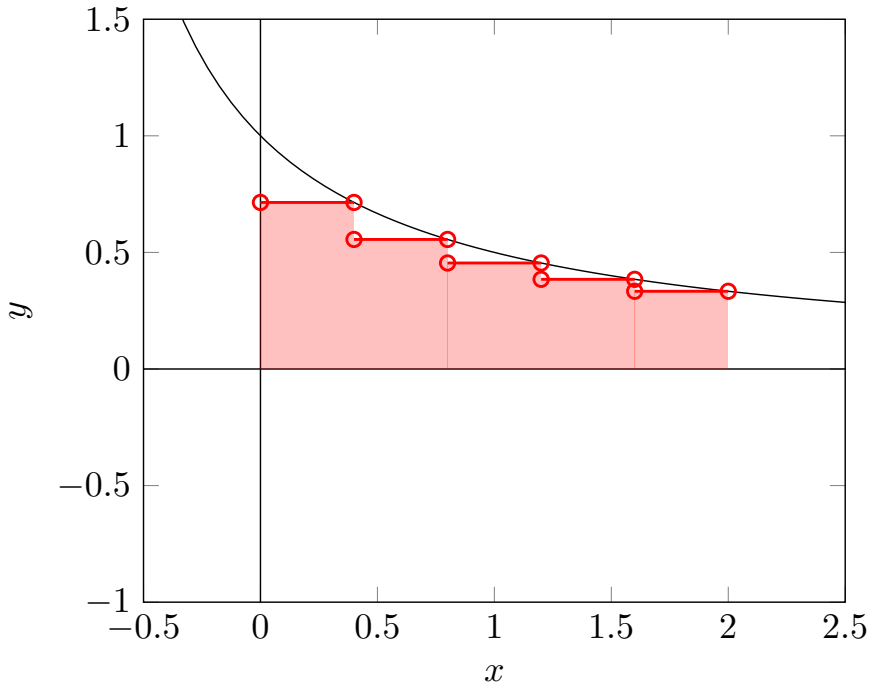
Once the partition is identified, calculate the value of the integrand function $f(x) = \dfrac{1}{x+1}$ at the right endpoint of each subinterval. The table below summarizes these computations.

k (index)	$[x_{k-1}, x_k]$ (interval)	$f(x_k)$ (value)
1	$[0, 0.4]$	$f(0.4) = \frac{1}{1.4} \approx 0.7143$
2	$[0.4, 0.8]$	$f(0.8) = \frac{1}{1.8} \approx 0.5556$
3	$[0.8, 1.2]$	$f(1.2) = \frac{1}{2.2} \approx 0.4546$
4	$[1.2, 1.6]$	$f(1.6) = \frac{1}{2.6} \approx 0.3846$
5	$[1.6, 2]$	$f(2) = \frac{1}{3} \approx 0.3333$

Knowing the constant values on each subinterval, if we called the simple function using the right endpoint $\mathcal{R}_f(x)$, then we have the Riemann sum

$$
\begin{aligned}
\int_0^2 \mathcal{R}_f(x)\,dx &= \sum_{k=1}^{5} f(x_k)\Delta x\\
&\approx 0.7143(0.4) + 0.5556(0.4) + 0.4546(0.4) + 0.3846(0.4) + 0.3333(0.4)\\
&\approx 0.9770
\end{aligned}
$$

The graph below shows the original function $y = f(x)$ and the simple function $y = \mathcal{R}_f(x)$ that was used in the Riemann sum.



□

The previous two examples illustrated very specific Riemann sums, where the size of the partition was specified as a small number. In order to compute definite integrals using limits of Riemann sums, we need to find an explicit formula for a Riemann sum involving a partition of unspecified size n.

The basic steps for these problems are as follows.

- Create a formula for the partition with increment

$$\Delta x = \frac{b-a}{n}$$

and partition points defined by an arithmetic sequence

$$x_k = a + k\,\Delta x.$$

- Evaluate the integrand function $f(x)$ at the appropriate choice, usually at an end point such as $f(x_{k-1})$ (left) or $f(x_k)$ (right), and expand the formula as necessary.

- Write down the Riemann sum using summation notation. Apply the properties of summation and the summation formulas to find an explicit formula for the Riemann sum in terms of n. The typical representation of the Riemann sum uses the form

$$\sum_{k=1}^{n} f(x_k^*)\Delta x,$$

where $f(x_k^*)$ is the function value chosen for the kth subinterval of the partition depending on which rule is chosen.

- To find the actual definite integral, take a limit of the explicit formula as $n \to \infty$. That is, the definite integral is computed as

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{k=1}^{n} f(x_k^*)\Delta x,$$

Example 6.4.6 Use a Riemann sum with the right-hand rule and a uniform partition of size n to approximate $\int_{-1}^{5}(5-2x)dx$.

Solution. To find the partition of the interval $[a,b]=[-1,5]$, we compute the partition increment size

$$\Delta x=\frac{5--1}{n}=\frac{6}{n}.$$

The partition points are defined using an arithmetic sequence

$$x_k=-1+k\cdot\frac{6}{n}=-1+\frac{6k}{n}.$$

The partition defines the kth subinterval $[x_{k-1},x_k]$ such that the right-hand rule will evaluate the integrand $f(x)=5-2x$ at the point x_k,

$$\begin{aligned}f(x_k)&=5-2x_k=5-2(-1+\frac{6k}{n})\\&=5+2-\frac{12k}{n}=7-\frac{12k}{n}\end{aligned}$$

The Riemann sum is equal to the sum of increments computed as the integrand function (rate) times the partition increment width. That is, if we use the function $\mathcal{R}_f(x)$ as the simple function using the right-hand end points of the intervals, then the Riemann sum is

$$\int_{-1}^{5}\mathcal{R}_f(x)dx=\sum_{k=1}^{n}f(x_k)\cdot\Delta x.$$

Using the value we found above and $\Delta x=\frac{6}{n}$, this gives

$$\begin{aligned}\int_{-1}^{5}\mathcal{R}_f(x)dx&=\sum_{k=1}^{n}(7-\frac{12k}{n})\cdot\frac{6}{n}\\&=\sum_{k=1}^{n}(\frac{42}{n}-\frac{72k}{n^2})\\&\overset{\text{Linearity}}{=}\frac{1}{n}\sum_{k=1}^{n}42-\frac{72}{n^2}\sum_{k=1}^{n}k\\&=\frac{1}{n}\cdot(42n)-\frac{72}{n^2}\cdot\frac{n(n+1)}{2}\\&=42-\frac{36(n+1)}{n}\end{aligned}$$

This final formula is the value of the Riemann sum using the right-hand rule.

The limit of the Riemann sum is the value of the actual definite integral of interest. That is, for this problem, we have

$$\begin{aligned}\int_{-1}^{5}(5-2x)dx&=\lim_{n\to\infty}[42-\frac{36(n+1)}{n}]\\&=\lim_{n\to\infty}[42-36\cdot\frac{n(1+\frac{1}{n})}{n}]\\&=42-36\cdot1=6.\end{aligned}$$

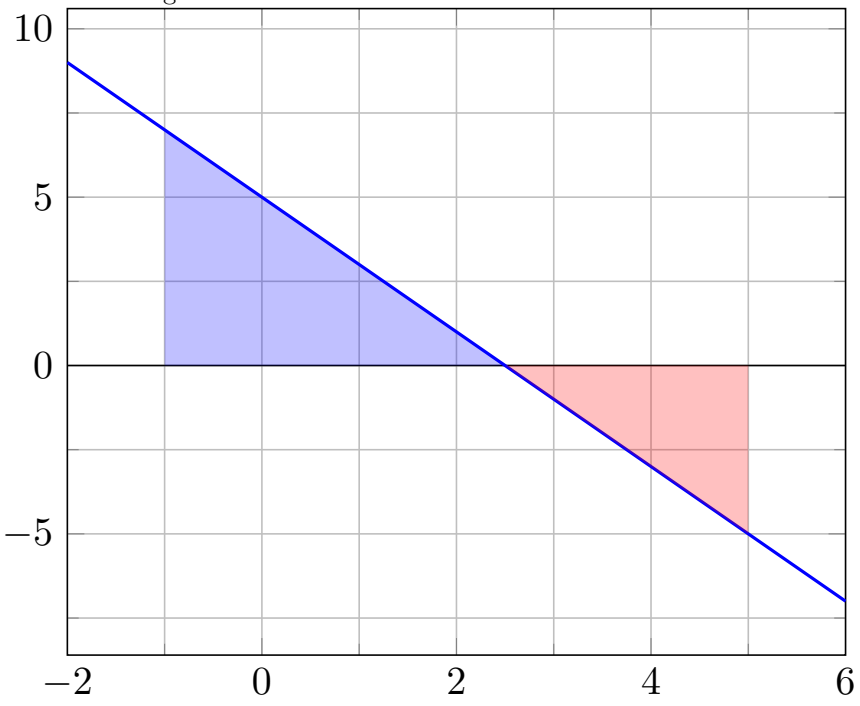
Because the graph $y = f(x)$ (shown below) is linear, we can compute the corresponding signed area using the area of triangles and compare our calculation. The graph crosses the axis when $f(x) = 0$ which occurs at $x = 2.5$. So we split the definite integral into two pieces,

$$\int_{-1}^{5} f(x)dx = \int_{-1}^{2.5} f(x)dx + \int_{2.5}^{5} f(x)dx.$$

The first region on interval $[-1, 2.5]$ is a triangle above the axis with height 7 and width 3.5 so that the area of the region is $\frac{1}{2}(7)(3.5) = 12.25$. The second region on interval $[2.5, 5]$ is a triangle below the axis with height 5 (since $f(5) = -5$) and base width $5 - 2.5 = 2.5$. The area of the second triangle is $\frac{1}{2}(5)(2.5) = 6.25$ but corresponds to a signed area of -6.25 (because below the axis). So

$$\int_{-1}^{5} f(x)dx = 12.25 + -6.25 = 6.$$

Thus, the limit of the Riemann sum exactly agrees with the geometric calculation of total signed area.



□

6.4.3 Summary

- The definite integral $\int_a^b f(x)\,dx$ of a general function $f(x)$ can be approximated by a **Riemann sum**, which is the definite integral of a simple function that approximates $f(x)$ as a piecewise constant function.

- A **partition of an interval** of size n is a finite sequence of values $P = (x_0, x_1, \ldots, x_n)$ which defines n subintervals $I_k = [x_{k-1}, x_k]$. The lengths of the subintervals are the increments $\nabla x_k = x_k - x_{k-1}$.

A **uniform partition of the interval** $[a, b]$ of size n has equal increments $\nabla x_k = \Delta x = \dfrac{b-a}{n}$. The sequence of points is arithmetic with formula

$$x_k = a + k\,\Delta x.$$

- A Riemann sum of $\displaystyle\int_a^b f(x)\,dx$ on a partition P identifies values in each subinterval, $x_k^* \in [x_{k-1}, x_k]$, uses the values $f(x_k^*)$ to define a simple function, and computes the integral of the simple function as a simple sum,

$$\sum_{k=1}^{n} f(x_k^*)\nabla x_k.$$

  Most calculations use simple rules to identify the points of evaluation.

  - Left-Hand Rule: Choose $x_k^* = x_{k-1}$ (left end-point).
  - Right-Hand Rule: Choose $x_k^* = x_k$ (right end-point).
  - Mid-Point Rule: Choose $x_k^* = \dfrac{x_{k-1} + x_k}{2}$ (mid-point).
  - Trapezoid Rule: Choose x_k^* so that $f(x_k^*) = \dfrac{f(x_{k-1}) + f(x_k)}{2}$ (average height of sides).
  - Lower-Sum Rule: Choose x_k^* so that $f(x_k^*) = \min(f(x) : x \in [x_{k-1}, x_k]$ (minimum value).
  - Upper-Sum Rule: Choose x_k^* so that $f(x_k^*) = \max(f(x) : x \in [x_{k-1}, x_k]$ (maximum value).

- The definite integral is the limit of all Riemann sums as the partition size grows $n \to \infty$ and the increments shrink $\Delta x \to 0$. In particular, for a uniform partition,

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{k=1}^{n} f(x_k^*)\Delta x.$$

- To approximate a definite integral $\displaystyle\int_a^b f(x)\,dx$ for a specific size Riemann sum (i.e., for a specific value of n), we apply the following steps.

  1. Find the specific partition points.
  2. Identify the evaluation points x_k^* according to the rule being used.
  3. Calculate the specific values of the integrand $f(x_k^*)$.
  4. Add the increments of the Riemann sum, multiplying each rate value $f(x_k^*)$ by the increment ∇x_k,

$$\sum_{k=1}^{n} f(x_k^*)\,\nabla x_k.$$

- To express a definite integral $\displaystyle\int_a^b f(x)\,dx$ as the limit of uniform Riemann sums, we apply the following steps.

  1. Calculate the uniform increment $\Delta x = \frac{b-a}{n}$.

2. Find and simplify the explicit formula for the partition points

$$x_k = a + k\,\Delta x.$$

3. Compute $f(x_k^*)$, usually using $x_k^* = x_k$ (right-hand rule), using function substitution (composition).
4. Write down the limit of the Riemann sum, remembering to multiply the rate value $f(x_k^*)$ by Δx,

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{k=1}^{n} f(x_k^*)\Delta x.$$

To compute the value of the definite integral using the limit of Riemann sums, we first compute the formula for the Riemann sum in terms of n,

$$\sum_{k=1}^{n} f(x_k^*)\,\Delta x,$$

and then evaluate the limit of that formula.

6.4.4 Exercises

1. Pending.

Chapter 7

Other Stuff Not Yet Placed

Part III

Chapter 8

Modeling Rates of Change

8.1 Extreme Values

We have learned earlier that when a function $f(x)$ can be written as an accumulation function, we can describe the behavior of the function in terms of its rate of accumulation $f'(x)$. We use sign analysis of $f'(x)$ to find the intervals of monotonicity of $f(x)$. And if $f'(x)$ can also be written as an accumulation function with rate $f''(x)$, the sign analysis of $f''(x)$ determines the intervals of concavity of $f(x)$.

In this section, we apply this information to describe the extreme values of a function. By identifying points where $f'(x)$ changes sign, we can find local maximum and minimum values. Global extreme values require comparing local extremes with end behaviors.

8.1.1 Local Extreme Values

When the derivative $f'(x)$ changes sign at a point where $f(x)$ is continuous, the function has a local or relative extreme value. We begin by focusing on what we mean by a local extreme value. A local extreme is a point where the function reaches either its highest or lowest point *on an interval around that point*. The function might exceed the value on some other interval, but the value needs to be the extreme in a neighborhood of the point.

Definition 8.1.1 Local (Relative) Extreme Values. A function $f(x)$ has a **local maximum** at a point $x = c$ in the domain if there is an interval (a, b) with $c \in (a, b)$ so that $f(x) \leq f(c)$ for all $x \in (a, b)$.

A function $f(x)$ has a **local minimum** at a point $x = c$ in the domain if there is an interval (a, b) with $c \in (a, b)$ so that $f(c) \leq f(x)$ for all $x \in (a, b)$. ◊

The reason that the definition describes these extreme values as local extremes is that the function might go higher or lower at some point outside of the interval. A local or relative extreme value is only at the highest or lowest points relative to its immediate neighbors. The following graph illustrates a function with multiple local extreme values.

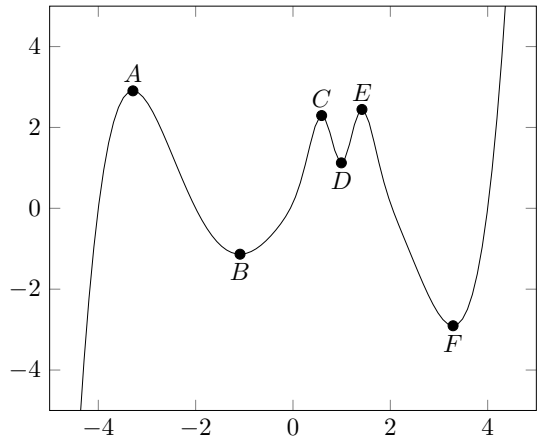


Figure 8.1.2 Illustration of a function with local extremes and no global extremes.

The function shows local maxima at points labeled A, C, and E and local minima at points labeled B, D, and F. Of the three maxima, the value at A is the greatest. Because the graph continues to increase after F, however, the function reaches values higher than all of the local maxima. Similarly, the

minimum at F is the lowest of the three lolal minima but the function reach values even lower at points to the left of A.

One observation we should make is that the local extreme values occurred where the function transitioned between an interval of increasing to an interval of decreasing. Such points are called turning points. Sign analysis using the first derivative can often identify these turning points, so we use sign analysis to find local extreme values. The applicable theorem is called the first derivative test for extreme values.

Theorem 8.1.3 First Derivative Test. *Suppose that $f'(x)$ exists on an interval (a, b), possibly except at $x = c$ with $a < c < b$ and that $f(x)$ is continuous at $x = c$.*

- *If $f'(x) < 0$ for $x \in (a, c)$ and $f'(x) > 0$ for $x \in (c, b)$, then $f(x)$ is decreasing on $(a, c]$ and increasing on $[c, b)$ so that f has a **local minimum** at $x = c$.*
- *If $f'(x) > 0$ for $x \in (a, c)$ and $f'(x) < 0$ for $x \in (c, b)$, then $f(x)$ is increasing on $(a, c]$ and decreasing on $[c, b)$ so that f has a **local maximum** at $x = c$.*
- *If $f'(x)$ does not change sign, then f does* not *have a local extreme value at $x = c$.*

Because $f'(x)$ most frequently changes sign at points where $f'(x) = 0$, we call such points the **critical points** of $f(x)$. When we have a more precise definition of the derivative, we will learn that critical points also need to include points where $f'(x)$ is not defined.

Example 8.1.4 Find the local extreme values of $f(x) = x^3 - 6x^2 - 36x + 5$.

Solution. The first step in a question about extreme values is to compute the rate of change $f'(x)$.

$$f'(x) = 3x^2 - 12x - 36.$$

To apply the First Derivative Test, we need to complete sign analysis. Because $f'(x)$ is defined and continuous everywhere, our critical points of f are the roots of $f'(x)$. We solve the equation by factoring:

$$\begin{aligned} 3x^2 - 12x - 36 &= 0 \\ 3(x^2 - 4x - 12) &= 0 \\ 3(x - 6)(x + 2) &= 0. \end{aligned}$$

There are two roots of f', $x = 6$ and $x = -2$, which are the critical points of f.

Next, we perform sign analysis using the roots as the end points of the test intervals, which are $(-\infty, -2)$, $(-2, 6)$, and $(6, \infty)$. Using the factored function $f'(x) = 3(x + 2)(x - 6)$ makes it easier to find the signs without necessarily computing the final value. We can just look at the signs of each factor:

$$\begin{aligned} f'(-3) &= 3(-3 + 2)(-3 - 6) = (+)(-)(-) = +, \\ f'(0) &= 3(0 + 2)(0 - 6) = (+)(+)(-) = -, \\ f'(8) &= 3(8 + 2)(8 - 6) = (+)(+)(+) = +. \end{aligned}$$

We now know f is *increasing* on $(-\infty, -2]$, *decreasing* on $[-2, 6]$, and *increasing* on $[6, \infty)$.

By the First Derivative Test, we now know that f has a local maximum at $x = -2$ and a local minimum at $x = 6$. The y-coordinate of the local maximum

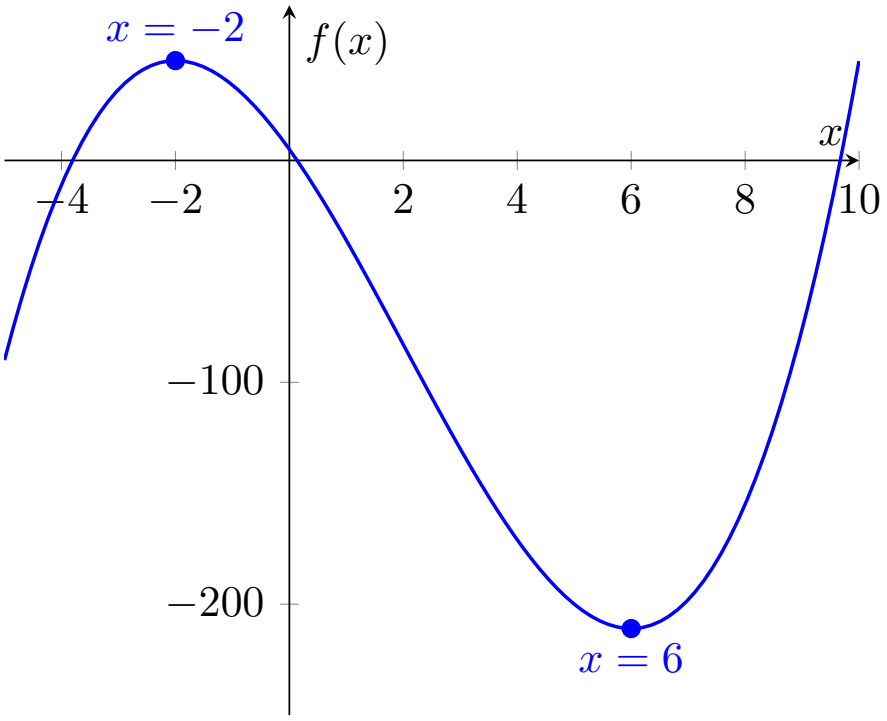
can be found using the formula for $f(x)$:

$$f(-2) = (-2)^3 - 6(-2)^2 - 36(-2) + 5 = 45.$$

This is guaranteed to be the maximum value over the interval $(-\infty, 6]$. The y-coordinate of the local minimum can also be found:

$$f(6) = (6)^3 - 6(6)^2 - 36(6) + 5 = -211,$$

which is guaranteed to be the minimum over the interval $[-2, \infty)$. If we wanted to graph this function and show the local extrema, we would know that our window would need to include the x-values of $x = -2$ and $x = 6$ as well these y-values.



□

Although we do not yet know all of the rules that would allow us to compute derivatives, with the help of technology we can analyze many other functions.

Example 8.1.5 Use technology to find the derivative of the function

$$f(x) = \frac{x}{x^2 + 3}.$$

Then describe the local extreme values of $f(x)$.

Solution. In SageMath, we find the derivative formula using the `diff` command, which stands for the verb *differentiate*. The following script will define our function for SageMath and then ask it to show us the derivative where x is the independent variable.

```
# Define our function.
f(x)=x/(x^2+3)
# Show the derivative.
show( diff(f(x), x) )
```

```
-2x^2/(x^2+3)^2 + 1/(x^2+3)
```

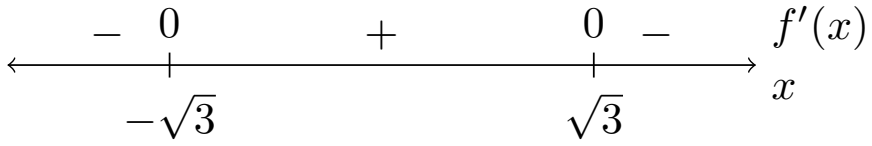
We see that $f(x)$ has a derivative

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

We can simplify this if we get a common denominator.

$$\begin{aligned} f'(x) &= -\frac{2x^2}{(x^2+3)^2} + \frac{(x^2+3)}{(x^2+3)^2} \\ &= \frac{-2x^2+x^2+3}{(x^2+3)^2} \\ &= \frac{3-x^2}{(x^2+3)^2} \end{aligned}$$

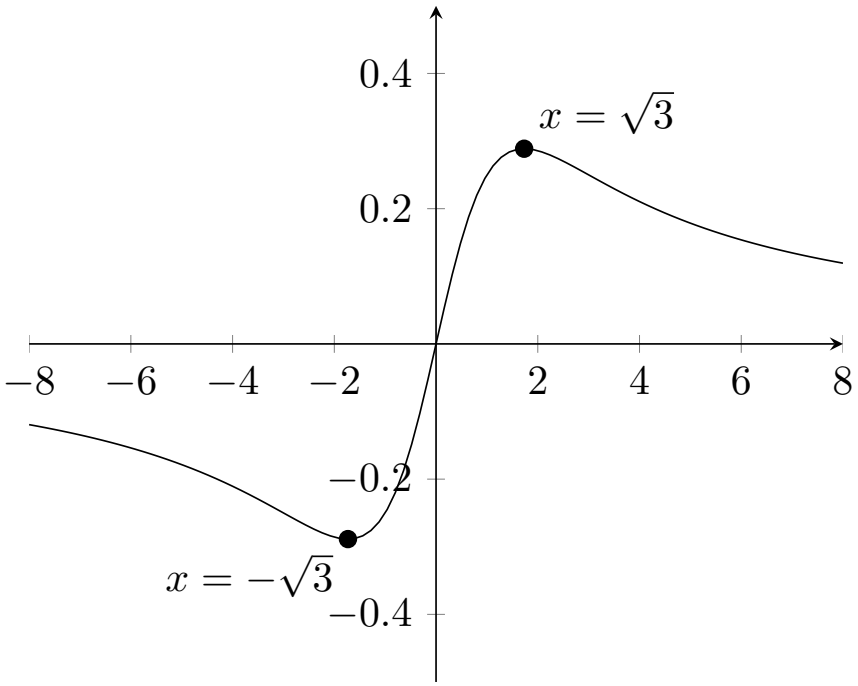
The denominator of $f'(x)$ is never zero because $x^2+3 \geq 3$ will never equal zero. So the sign can only change where $3-x^2=0$ which occurs at two values, $x = \pm\sqrt{3}$. There are three intervals of interest to test, $(-\infty, -\sqrt{3})$, $(-\sqrt{3}, \sqrt{3})$, and $(\sqrt{3}, \infty)$. We can find the signs of $f'(x)$ using the values $x=-2$, $x=0$ and $x=2$. The signs are summarized in the number line summary below.



We finish by interpreting our results.

- Because $f'(x) < 0$ on $(-\infty, -\sqrt{3})$ and $f'(x) > 0$ on $(-\sqrt{3}, \sqrt{3})$, we know $f(x)$ has a local minimum at $x = -\sqrt{3}$. (Minimum over the interval $(-\infty, +\sqrt{3})$)
- Because $f'(x) > 0$ on $(-\sqrt{3}, \sqrt{3})$ and $f'(x) < 0$ on $(\sqrt{3}, \infty)$, we know $f(x)$ has a local maximum at $x = \sqrt{3}$. (Maximum over the interval $(-\sqrt{3}, \infty)$)

A graph of the function is illustrated below.



□

Having discussed how the first derivative $f'(x)$ allows us to identify local extreme values of $f(x)$, we should note that the second derivative $f''(x)$ will allow us to identify local extreme values of $f'(x)$. These points are the **inflection points** of the function f, where the concavity of f changes. Inflection points are significant as being extreme values in that they represent points where the rate of accumulation or rate of change reaches either a maximum or minimum rate.

8.1.2 Global Extreme Values

In Figure 8.1.2, we saw that when a function has local extreme values, there could still be other points that are not local extremes that exceed the extremes. This leads to the idea of **global extreme values**.

Definition 8.1.6 Suppose the function f has domain D.

- f has a **global maximum** at $c \in D$ if $f(c) \geq f(x)$ for all $x \in D$.
- f has a **global minimum** at $c \in D$ if $f(c) \leq f(x)$ for all $x \in D$.

◊

To identify global extremes of a function, we first need to find all of the local extreme values. Then we use additional information to test whether the function manages to reach beyond those values. The sign analysis used to analyze local extrema does give us some information about the intervals immediately to the left and right of an extremum. For example, we know that a local maximum will be greater than all points in the immediately adjacent intervals, but we may not know how far down the function goes.

Finishing the analysis for global extremes generally involves computing limits of the function on intervals not already accounted for by the local extremes. If a limit has a real (finite) value, the function values approach that limit but may not actually reach the limit as an actual function value. When a function is approaches a value in a limit that *would be* an extreme value but never reaches it, we call that value a **bound** rather than an extreme.

Example 8.1.7 Find the global extreme values of $f(x) = 4x^2 - x^3$ restricted to each of the following domains.

1. $D = (-\infty, \infty)$
2. $D_1 = [-1, 3]$
3. $D_2 = (-2, 4]$
4. $D_3 = [-2, 2)$

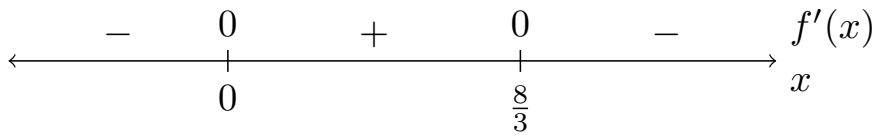
Solution. This question considers finding global extrema for a function when it is restricted to different domains. Regardless of which domain we use, we will need to find the local extreme values. Local extreme values are identified from sign analysis of $f'(x)$. We find

$$f'(x) = 4(2x) - (3x^2) = 8x - 3x^2.$$

Sign analysis begins by finding the roots, where $f'(x) = 0$.

$$\begin{aligned} 8x - 3x^2 &= 0 \\ x(8 - 3x) &= 0 \\ x = 0, \quad 8 - 3x &= 0 \\ x = 0, \quad x &= \frac{8}{3} \end{aligned}$$

Because $f'(x)$ is continuous, the roots determine the test intervals: $(-\infty, 0)$, $(0, \frac{8}{3})$, and $(\frac{8}{3}, \infty)$. Testing one value of x from each interval in $f'(x) = x(8 - 3x)$, we find the signs summarized on the following number line.



We can interpret the sign analysis of $f'(x)$ as characterizing the monotonicity of $f(x)$.

- $f(x)$ is decreasing on $(-\infty, 0)$.
- $f(x)$ is increasing on $(0, \frac{8}{3})$.
- $f(x)$ is decreasing on $(\frac{8}{3}, \infty)$.

Because $f(x)$ is decreasing on the left of $x = 0$ and increasing on the right, $f(x)$ has a *local minimum* at $x = 0$. The value of the function at this minimum is

$$f(0) = 4(0)^2 - (0)^3 = 0.$$

Similarly, because $f(x)$ is increasing on the left of $x = \frac{8}{3}$ and decreasing on the right, $f(x)$ has a *local maximum* at $x = \frac{8}{3}$. The value of the function at this maximum is

$$f(\frac{8}{3}) = 4(\frac{8}{3})^2 - (\frac{8}{3})^3 = \frac{256}{27} \approx 9.4815.$$

From our sign analysis of $f'(x)$, we know that $f(0)$ is the minimum value for the interval $(-\infty, \frac{8}{3}]$ and that $f(\frac{8}{3})$ is the maximum value for the interval $[\frac{8}{3}, \infty)$. To complete the analysis of global extreme values, we need use limits to compare $f(0)$ with points on the interval $(\frac{8}{3}, \infty)$ and $f(\frac{8}{3})$ with points on

the interval $(-\infty, 0)$. To evaluate limits involving $\pm\infty$, we need to factor out the dominant power,

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

before using the limit arithmetic of infinity.

$$\begin{aligned}
\lim_{x\to-\infty} f(x) &= \lim_{x\to-\infty} x^3(\frac{4}{x} - 1) \\
&= (-\infty)^3 \cdot (\frac{4}{-\infty} - 1) \\
&= -\infty \cdot (0 - 1) = +\infty \\
\lim_{x\to+\infty} f(x) &= \lim_{x\to+\infty} x^3(\frac{4}{x} - 1) \\
&= (\infty)^3 \cdot (\frac{4}{\infty} - 1) \\
&= \infty \cdot (0 - 1) = -\infty
\end{aligned}$$

Let us now address the question of the global extreme values on each of the requested restricted domains.

1. Find the global extremes on the interval $(-\infty, \infty)$.

   We have learned that on the interval $(-\infty, 0)$, f reaches all values between $f(0) = 0$ and $\lim_{x\to-\infty} f(x) = \infty$. Clearly, $f(x)$ has *no maximum value* because it is unbounded above. Similarly, we learned that on the interval $(\frac{8}{3}, \infty)$, $f(x)$ reaches all values between $-\infty$ and $f(\frac{8}{3}) = \frac{256}{27}$ and is unbounded below. So $f(x)$ has *no minimum value*. The range of f has been shown to be $(-\infty, \infty)$ and f has no global extreme values on $(-\infty, \infty)$.

2. Find the global extremes on the interval $D_1 = [-1, 3]$.

   We know f is decreasing on $[-1, 0)$, a subset of $(-\infty, 0)$, so the maximum value on that interval is

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

   We also know f is decreasing on $(\frac{8}{3}, 3]$ so the minimum value on that interval is

   $$f(3) = 4(3)^2 - (3)^3 = 9.$$

   Comparing these to the local minimum $f(0) = 0$ and the local maximum $f(\frac{8}{3}) = \frac{256}{27}$, we see that $f(0) = 0$ is the *global minimum* and $f(\frac{8}{3}) = \frac{256}{27}$ is the *global maximum* for the interval $D_1 = [-1, 3]$. When restricted to this domain, the range of f becomes $[0, \frac{256}{27}]$.

3. Find the global extremes on the interval $D_2 = (-2, 4]$.

   We know f is decreasing on $(-2, 0)$, so that the left end-point provides an upper bound

   $$\lim_{x\to-2^+} f(x) = f(-2) = 4(-2)^2 - (-2)^3 = 24.$$

   This is not a maximum value because $x = -2$ is not included in the domain. Because f is decreasing on $(\frac{8}{3}, 4]$ the minimum value on that interval is

   $$f(4) = 4(4)^2 - (4)^3 = 0.$$

The global minimum occurs in two locations,

$$f(0) = f(4) = 0.$$

The value $f(-2) = 24$ is not a global maximum because $x = -2$ is not included in the domain. However, f does include all values up to that value through the limit so that f has an upper bound of 24. When f is restricted to $D_2 = (-2, 4]$, the range is $[0, 24)$.

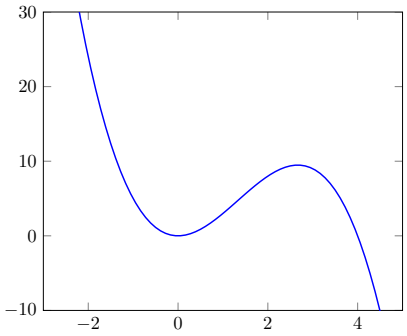
4. Find the global extremes on the interval $D_3 = [-2, 2)$.

   From the work above, we know $f(-2) = 24$ is the maximum value on $[-2, 0)$. Because the right end point $x = 2$ is to the left of the local maximum at $x = \frac{8}{3}$, we need to consider the interval of monotonicity $(0, 2)$. f is increasing on this interval and bounded above by the limit

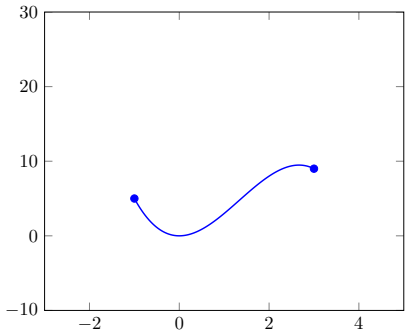
   $$\lim_{x \to 2^-} f(x) = f(2) = 4(2)^2 - 2^3 = 8.$$

   Our work shows that f has a global minimum at $x = 0$ with $f(0) = 0$ and a global maximum at $x = -2$ with $f(-2) = 24$. The range of f restricted to $D_3 = [-2, 2)$ is $[0, 24]$.

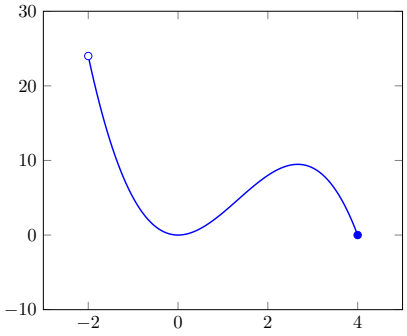
The following figure illustrates the graph $y = f(x)$ restricted to each domain. Be sure to compare the analysis that identified the global extremes with the graphs.



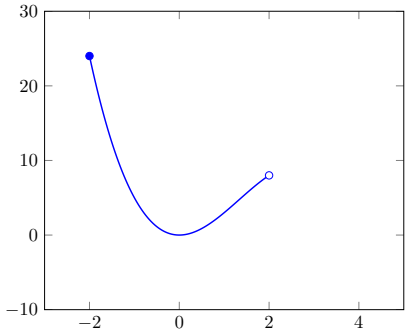
(a) Domain: $(-\infty, \infty)$



(b) Domain: $[-1, 3]$



(c) Domain: $(-2, 4]$



(d) Domain: $[-2, 2)$

Figure 8.1.8

□

The process to find global extrema is summarized as the following algorithm.

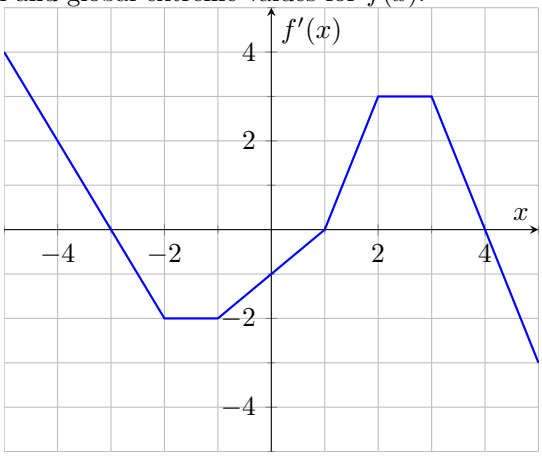
Algorithm 8.1.9 Finding Global Extreme Values. *To find the global extrema of a function $f(x)$ that is continuous on an interval:*

- *Determine the derivative $f'(x)$ and perform sign analysis.*
- *Identify all of the local extreme values and compute the value of $f(x)$ at each extreme.*
- *Find extremes or bounds for $f(x)$ at the end points using values or limits, respectively.*
- *Identify the highest and lowest values out of the local extremes and the end points.*
- *If an extreme value occurs at a point in the interval, that value is a global maximum/minimum. If an extreme value occurs as a limit at an excluded end point, that value is a bound but not a global extremum.*

8.1.3 Extreme Values involving Accumulation

In our examples above, we worked with functions with explicit formulas. However, most steps in the analysis involved only knowing information about the derivative. Here we consider examples where the derivative is given as the rate of accumulation and we do not know the explicit formula for the function in question.

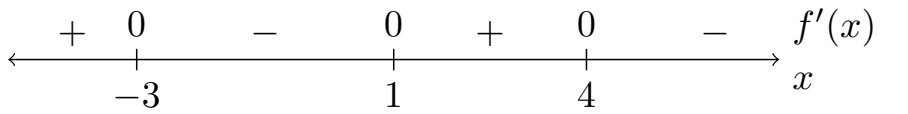
Example 8.1.10 Suppose f is defined as an accumulation function with an initial value $f(0) = 10$ and a rate of accumulation f' shown in the graph below. We assume the graph continues as shown outside of the viewing window. Determine local and global extreme values for $f(x)$.



Solution. Our function of interest is defined by an accumulation

$$f(x) = 10 + \int_0^x f'(z)\,dz.$$

We can find local extrema in the same way as before. However, instead of solving an equation $f'(x) = 0$, we can look at the graph to both find the roots and the signs of f'. The graph crosses the x-axis at $x = -3$, $x = 1$ and $x = 5$. The signs of $f'(x)$ are identified in the following sign analysis number line summary.



The Theorem 8.1.3 allows us to conclude that $f(x)$ has a local maximum at $x = -3$, a local minimum at $x = 1$, and another local maximum at $x = 4$. It is possible to decide which maximum has a higher value by considering the signed area of the graph. In particular, because $f'(x)$ has linear segments, we can compute the areas in question using elementary geometry to find

$$\int_{-3}^{1} f'(x)\,dx = -5,$$
$$\int_{1}^{4} f'(x)\,dx = 6.$$

Using the splitting property of definite integrals, this implies

$$f(4) - f(-3) = \int_{-3}^{4} f'(x)\,dx = -5 + 6 = 1.$$

Consequently, $f(4) = f(-3) + 1$ and f has a higher value at $x = 4$ than at $x = -3$.

We can find actual values if we use the initial value,

$$f(-3) = 10 + \int_{0}^{-3} f'(z)\,dz.$$

Because the integral goes right to left, we have $dz < 0$ and the signed area will be negated. Again using geometry, we find

$$f(-3) = 10 + -(-4.5) = 14.5.$$

Using this point and the integrals above, we can quickly find

$$f(1) = f(-3) + \int_{-3}^{1} f'(z)\,dz = 14.5 - 5 = 9.5,$$
$$f(4) = f(1) + \int_{1}^{4} f'(z)\,dz = 9.5 + 6 = 15.5.$$

To find global extrema, we need to think about what happens to the left and right of these local extrema. The sign analysis of $f'(x)$ shows that f is increasing on $(-\infty, -3)$. Thus, $f(-3)$ is the maximum value on $(-\infty, 3]$. The value of f is unbounded below on this interval,

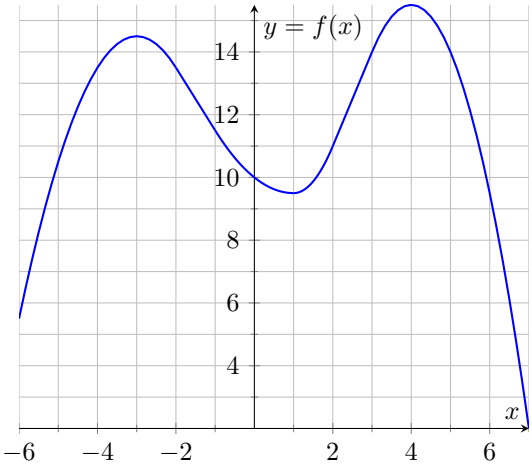
$$\lim_{x \to -\infty} f(x) = -\infty.$$

We can see this by considering the integral $\displaystyle\int_{-3}^{x}$ for $x < -3$. Because the integral goes right to left, $dz < 0$, and $f'(z) > 0$ on this interval, the integral becomes more and more negative the further x goes to the left. With a similar argument using $\displaystyle\int_{4}^{x} f'(z)\,dz$, we find

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

and f is again unbounded below on the interval $(4, \infty)$.

We conclude that f has a global maximum $f(4) = 13.5$ and no global minimum. A graph of $y = f(x)$ is shown below.



□

8.1.4 Summary

- A function f has a local maximum (plural: maxima) at $x = c$ if $f(c) \geq f(x)$ in a neighborhood of c. The function f has a local minimum (plural: minima) at $x = c$ if $f(c) \leq f(x)$ in a neighborhood of c.
- A function f has a global maximum at $x = c$ if $f(c) \geq f(x)$ for all x in the domain. The function f has a global minimum at $x = c$ if $f(c) \leq f(x)$ for all x in the domain.
- We can use sign analysis of the derivative f' to find local extreme values. Points where $f'(x)$ changes sign are local extrema. This is called the First Derivative Test.
- Global extrema can occur at local extrema or at boundaries of intervals. We need to compare the value of the function at each of the local extrema with the end points. If an end point is not included, the value of the limit at that point can serve as a bound for the function but would not be an actual extreme value.

8.1.5 Exercises

For each function, identify all local extrema.

1. $f(x) = x^3 - 9x^2 - 48x + 60$

2. $f(x) = 120x + 3x^2 - 2x^3$

3. $f(x) = x^4 - 8x^2$

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

For each function, find the global extrema on the given intervals.

5. $f(x) = 4x - x^2$ on (i) $D = [-1, 6]$, (ii) $D = (1, 4)$, and $(-\infty, \infty)$

6. $f(x) = x^2 + 3x$ on (i) $D = [-1, 2]$, (ii) $D = (-4, 1]$, and $(-\infty, \infty)$

7. $f(x) = x^3 - 9x^2 - 48x + 60$ on (i) $D = [-5, 5]$, (ii) $D = [-10, 10)$, and $(-\infty, \infty)$

8. $f(x) = x^4 - 12x^3 + 28x^2 - 17$ on (i) $D = [-1, 3]$, (ii) $D = (1, 8)$, and $(-\infty, \infty)$

8.2 The Derivative

8.2.1 Overview

Given a function, the derivative at a point allows us to measure the instantaneous rate of change at that point. This rate of change is defined as the limiting value of the average rate of change as the space between the two points used approaches zero. It would be quite tedious to compute the derivative at each point using this process. Fortunately, we can create a function known as the derivative that does this for us.

In this section, we introduce the derivative as a function rather than an isolated calculation. The definition of the derivative is still in terms of a limit, but with the point in question represented by a variable. The domain of the derivative consists of all points where the limit exists, and corresponds to the set of points where a tangent line can be defined as a function. For algebraically defined functions, we can use the limit and algebra to find an algebraic formula for the derivative. Several examples illustrate this process. The derivative function can then be used to calculate the instantaneous rate at any desired point.

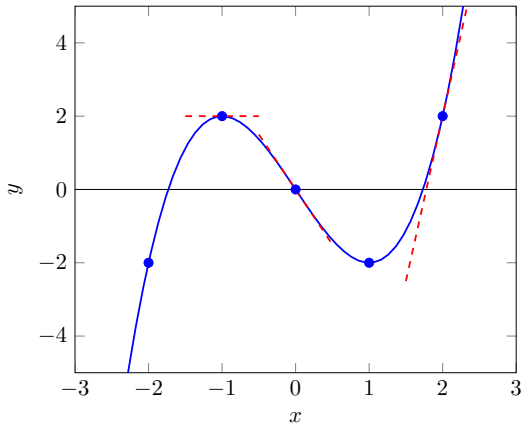
8.2.2 Introducing the Derivative

A function is a rule that associates a unique output with each input value. When we look at the graph of a function, the points on the graph are placed so that the input value is the first coordinate (e.g., x) and the output value is the second coordinate (e.g., y). Using the graph, we can find the value of the function for a given input by looking for the input along the horizontal (x) axis and then finding the point on the graph intersecting the corresponding vertical line. The height of that point gives the output value of the function.

If that point of the graph has a well-defined tangent line, then we could define another function that has as its output the slope of the tangent line at that point. This function is called the derivative function. The following website provides an interactive illustration of this concept: http://www.intmath.com/differentiation/derivative-graphs.php

Consider the figure illustrated below. The graph of a function $y = f(x)$is given and short segments of the tangent lines at various points have also been included. The point at $x = -1$ has a y-value of 2 and a slope of 0 (horizontal tangent line). So $f(-1) = 2$ while $\frac{df}{dx}(-1) = 0$. The point at $x = 0$ has a y-value of 0 and a slope of -3, so $f(0) = 0$ and $\frac{df}{dx}(0) = -3$. The third point, at $x = 2$ has a y-value of 2 and a slope of 9, corresponding to $f(2) = 2$ and $\frac{df}{dx}(2) = 9$. The equations of these tangent lines, listed in the same order as described above, and written in point–slope form, are given by

$$y = 2,$$
$$y = -3x,$$
$$y = 9(x - 2) + 2.$$



Definition 8.2.1 The Derivative. Suppose f is a function relating two variables $f : x \mapsto y$. The derivative of y is a new dependent variable $\frac{dy}{dx}$ that for every value of the independent variable, $x = a$, has a value equal to the instantaneous rate of change, $\frac{dy}{dx} = \frac{dy}{dx}\Big|_a$. The function $f' : x \mapsto \frac{dy}{dx}$ is called the **derivative** of f. That is,

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

The domain of f' is the set of all values x where the limit exists. ◊

Our previous use of the notation f' to represent the rate of accumulation of an accumulation function was intentional. It looked ahead to the Fundamental Theorem of Calculus that connects the ideas of derivatives and integrals. In this section, we will use the names f' and $\frac{df}{dx}$ interchangeably. The Fundamental Theorem of Calculus will justify this equivalence later.

8.2.3 Examples of Calculation

Recall that the definition of the instantaneous rate of change (which is what the derivative measures) is the limiting value of an average rate of change of the function between two points as the second point approaches the first. When computing the derivative, we will use two points at symbolic values x (the point of interest) and $x+h$ (the second point), where h is the spacing between the two points. The second point approaches the first when $h \to 0$. The basic process is outlined in the following steps:

1. Compute $f(x+h)$ using the rule for $f(x)$. (Find the output for the second point.)
2. Compute $f(x+h) - f(x)$ and simplify. (Find the change in output.)
3. Simplify $\dfrac{\Delta f}{\Delta x} = \dfrac{f(x+h) - f(x)}{h}$. (Determine a simplified formula for the average rate of change.)
4. Determine $\dfrac{df}{dx} = \lim\limits_{h \to 0} \dfrac{f(x+h) - f(x)}{h}$. (Evaluate the limiting value as the second point approaches the first.)

Example 8.2.2 Use the definition of the derivative to find $\frac{df}{dx}$ where $f(x) = x^2 - 3x$.

Solution.

- Find the output at the second point:

$$f(x+h) = (x+h)^2 - 3(x+h) = x^2 + 2xh + h^2 - 3x - 3h.$$

- Find the change in output between the two points:

$$\begin{aligned} f(x+h) - f(x) &= (x^2 + 2xh + h^2 - 3x - 3h) - (x^2 - 3x) \\ &= x^2 + 2xh + h^2 - 3x - 3h - x^2 + 3x \\ &= 2xh + h^2 - 3h. \end{aligned}$$

- Simplify the average rate of change between the two points:

$$\begin{aligned} \frac{f(x+h) - f(x)}{h} &= \frac{2xh + h^2 - 3h}{h} \\ &= \frac{h(2x + h - 3)}{h} \\ &= 2x + h - 3. \end{aligned}$$

- The derivative is the limit of the average rate of change:

$$\begin{aligned} \frac{df}{dx} &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h \to 0} 2x + h - 3 \\ &= 2x + 0 - 3 = 2x - 3. \end{aligned}$$

  So we have found the derivative function, $\frac{df}{dx}(x) = 2x - 3$. We can see this exactly agrees with the rate of accumulation $f'(x) = 2x - 3$ that we earlier learned to find in terms of the elementary accumulation formulas.

□

Often it is more convenient to combine some of these steps together. However, just be careful that you create valid equations. Always have an equation that says what you are computing, and do not write that two things are equal when they are not the same. In the previous example, note how each time I started to compute a new expression, I created a new system of equations.

In this next example, we are reminded of the need to find a common denominator when a fraction is involved. Also, it is useful to recall that division by a number $h$is the same as multiplication by its inverse $1/h$.

Example 8.2.3 Use the definition of the derivative to find $f'(x)$ where $f(x) = \dfrac{1}{2x+3}$.

Solution.

- Find the output at the second point:

$$f(x+h) = \frac{1}{2(x+h)+3} = \frac{1}{2x + 2h + 3}.$$

- Find the change in output between the two points:

$$f(x+h) - f(x) = \frac{1}{2x + 2h + 3} - \frac{1}{2x + 3}.$$

  Here is where we will need to use a common denominator. Recall from ordinary fractions that a common denominator is formed by multiplying

the top and bottom by a missing factor.

$$\begin{aligned} f(x+h)-f(x) &= \frac{(2x+3)}{(2x+3)(2x+2h+3)} - \frac{(2x+2h+3)}{(2x+3)(2x+2h+3)} \\ &= \frac{(2x+3)-(2x+2h+3)}{(2x+3)(2x+2h+3)} \\ &= \frac{-2h}{(2x+3)(2x+2h+3)} \end{aligned}$$

- Simplify the average rate of change between the two points. However, it is dangerous to write a fraction divided by something (division is not associative), so we will write division by h as multiplication by $1/h$ and simplify the resulting expression:

$$\begin{aligned} \frac{f(x+h)-f(x)}{h} &= \frac{-2h}{(2x+3)(2x+2h+3)} \cdot \frac{1}{h} \\ &= \frac{-2}{(2x+3)(2x+2h+3)} \end{aligned}$$

  Your goal at the simplification step should always be to make the h cancel as a common factor.

- The derivative is the limit of the average rate of change:

$$\begin{aligned} \frac{df}{dx} &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \\ &= \lim_{h\to 0} \frac{-2}{(2x+3)(2x+2h+3)} \\ &= \frac{-2}{(2x+3)(2x+0+3)} \\ &= \frac{-2}{(2x+3)^2}. \end{aligned}$$

  This gives us the derivative function,

$$\frac{df}{dx} = \frac{-2}{(2x+3)^2}.$$

□

For our last examples, we consider finding the derivative using the definition when the function involves the square root. We will find it necessary to use a trick from algebra involving conjugate pairs. Recall that $(a+b)(a-b) = a^2-b^2$. If a or b is a square root of some value, then the product of these conjugate pairs will have the square of the square root, thereby no longer involving the square root. For example,

$$(\sqrt{x}-2)(\sqrt{x}+2) = (\sqrt{x})^2 - (2)^2 = x-4.$$

Example 8.2.4 Use the definition of the derivative to find $f'(x)$ where $f(x) = \sqrt{x}$.

Solution.

- Find the output at the second point:

$$f(x+h) = \sqrt{x+h}$$

- Find the change in output between the two points:

$$f(x+h) - f(x) = \sqrt{x+h} - \sqrt{x}.$$

- Simplify the average rate of change between the two points. This will require multiplying the numerator and denominator by the conjugate pair:

$$\begin{aligned}\frac{f(x+h)-f(x)}{h} &= \frac{\sqrt{x+h}-\sqrt{x}}{h} \\ &= \frac{(\sqrt{x+h}-\sqrt{x})(\sqrt{x+h}+\sqrt{x})}{h(\sqrt{x+h}+\sqrt{x})} \\ &= \frac{(\sqrt{x+h})^2-(\sqrt{x})^2}{h(\sqrt{x+h}+\sqrt{x})} = \frac{x+h-x}{h(\sqrt{x+h}+\sqrt{x})} \\ &= \frac{h}{h(\sqrt{x+h}+\sqrt{x})} = \frac{1}{\sqrt{x+h}+\sqrt{x}}.\end{aligned}$$

- The derivative is the limit of the average rate of change:

$$\begin{aligned}f'(x) &= \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} \\ &= \lim_{h\to 0}\frac{1}{\sqrt{x+h}+\sqrt{x}} \\ &= \frac{1}{\sqrt{x+0}+\sqrt{x}} \\ &= \frac{1}{2\sqrt{x}}\end{aligned}$$

  This gives us the derivative function,

$$f'(x) = \frac{1}{2\sqrt{x}}.$$

□

Example 8.2.5 Use the definition of the derivative to find $f'(x)$ where $f(x) = \sqrt{2x-5}$.

Solution.

- Find the output at the second point:

$$f(x+h) = \sqrt{2(x+h)-5} = \sqrt{2x+2h-5}$$

- Find the change in output between the two points:

$$f(x+h) - f(x) = \sqrt{2x+2h-5} - \sqrt{2x-5}.$$

- Simplify the average rate of change between the two points. This will require multiplying the numerator and denominator by the conjugate pair:

$$\begin{aligned}\frac{f(x+h)-f(x)}{h} &= \frac{\sqrt{2x+2h-5}-\sqrt{2x-5}}{h} \\ &= \frac{(\sqrt{2x+2h-5}-\sqrt{2x-5})(\sqrt{2x+2h-5}+\sqrt{2x-5})}{h(\sqrt{2x+2h-5}+\sqrt{2x-5})}\end{aligned}$$

$$= \frac{(\sqrt{2x+2h-5})^2 - (\sqrt{2x-5})^2}{h(\sqrt{2x+2h-5}+\sqrt{2x-5})} = \frac{(2x+2h-5)-(2x-5)}{h(\sqrt{2x+2h-5}+\sqrt{2x-5})}$$
$$= \frac{2h}{h(\sqrt{2x+2h-5}+\sqrt{2x-5})} = \frac{2}{\sqrt{2x+2h-5}+\sqrt{2x-5}}.$$

- The derivative is the limit of the average rate of change:

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \\ &= \lim_{h\to 0} \frac{2}{\sqrt{2x+2h-5}+\sqrt{2x-5}} \\ &= \frac{2}{\sqrt{2x+0-5}+\sqrt{2x-5}} \\ &= \frac{1}{\sqrt{2x-5}} \end{aligned}$$

This gives us the derivative function,

$$f'(x) = \frac{1}{\sqrt{2x-5}}.$$

□

8.2.4 Differentiability

A function is **differentiable** at points where the derivative is defined. Alternatively, because the derivative at a point represents the slope of the tangent line, we say the function is differentiable at a point wherever the function has a well-defined tangent line.

Definition 8.2.6 Differentiability. A function f is **differentiable** at a if $f'(a)$ exists, or more precisely the limit

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

exists. ◊

A function f is not differentiable at $x=a$ if the limit defining $f'(a)$ does not exist. There are several reasons this might occur. The first reason is if the function is not continuous.

Theorem 8.2.7 Differentiable Implies Continuous. *If f is differentiable at a, then f must be continuous at a. Equivalently, if f is not continuous, then f must not be differentiable.*

Proof. Suppose that f is differentiable at a. This means that $f'(a)$ is a value defined by

$$\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = f'(a).$$

We also know that

$$\lim_{x\to a} [x-a] = a-a = 0.$$

Using the product rule for limits (LC:Product), this implies

$$\lim_{x\to a} [f(x)-f(a)] = \lim_{x\to a} \frac{f(x)-f(a)}{x-a} \cdot (x-a) = f'(a)\cdot 0 = 0.$$

Because $\lim_{x\to a} f(a) = f(a)$ (LE:Constant), we know that

$$\lim_{x\to a} f(x) = \lim_{x\to a}[f(x) - f(a) + f(a)] = 0 + f(a) = f(a)$$

using the sum rule (LC:Sum). Therefore, f is continuous at a. ■

Example 8.2.8 Show that $f(x)$ defined below is not continuous and not differentiable at $x = 1$.

$$f(x) = \begin{cases} 5 - 2x, & x < 1 \\ 3x + 1, & x \geq 1 \end{cases}$$

Solution. To check continuity, we evaluate limits from the left and right and compare it to the value of the function $f(1) = 3(1) + 1 = 4$.

$$\begin{aligned} \lim_{x\to 1^-} f(x) &= \lim_{x\to 1^-} 5 - 2x \\ &= 5 - 2(1) = 3 \\ \lim_{x\to 1^+} f(x) &= \lim_{x\to 1^+} 3x + 1 \\ &= 3(1) + 1 = 4 \end{aligned}$$

The function has a jump discontinuity at $x = 1$. The value on the function does agree with the limit on the right, but not the limit on the left.

Theorem 8.2.7 guarantees that we will find f is not differentiable at $x = 1$. We can verify this by computing the actual limits defining the derivative. When $h < 0$,

$$f(1 + h) = 5 - 2(1 + h) = 3 - 2h.$$

Because f is not continuous from the left, computing the derivative from the left will result in an infinite limit.

$$\begin{aligned} \left.\frac{df}{dx}\right|_{1^-} &= \lim_{h\to 0^-} \frac{f(1+h) - f(1)}{h} \\ &= \lim_{h\to 0^-} \frac{(3 - 2h) - 4}{h} \\ &= \lim_{h\to 0^-} \frac{-2h - 1}{h} \\ &\to \frac{-1}{0^-} = +\infty \end{aligned}$$

On the other side, when $h > 0$,

$$f(1 + h) = 3(1 + h) + 1 = 4 + 3h$$

so that

$$\begin{aligned} \left.\frac{df}{dx}\right|_{1^+} &= \lim_{h\to 0^+} \frac{f(1+h) - f(1)}{h} \\ &= \lim_{h\to 0^+} \frac{(4 + 3h) - 4}{h} \\ &= \lim_{h\to 0^+} \frac{3h}{h} \\ &= 3 \end{aligned}$$

With the derivative from the left being infinite, $f'(1)$ is undefined and f is not differentiable at $x = 1$. □

Another way that a function might not have be differentiable is where it is continuous but has a corner. This means that the slope at the point looks different from either of the two sides. Mathematically, if we computed the one-sided limits of the difference quotient defining the derivative, we would get two different values. The difference quotient has a jump discontinuity at $h = 0$.

Example 8.2.9 Consider the piecewise function defined by

$$f(x) = \begin{cases} x^2, & x \leq 1, \\ x, & x > 1. \end{cases}$$

Determine if f is differentiable at $x = 1$.

Solution. This function is continuous because the limit on the left and the limit on the right are equal to the value of the function at $x = 1$, as follows:

$$\begin{aligned} \lim_{x \to 1^-} f(x) &= \lim_{x \to 1} x^2 = 1^2 = 1, \\ \lim_{x \to 1^+} f(x) &= \lim_{x \to 1} x = 1 = 1, \\ f(1) &= 1^2 = 1. \end{aligned}$$

Now that we know the function is continuous, we can can compute the derivative using limits from the left and from the right. Using the function for $x < 1$, we have $f(1 + h) = (1 + h)^2$ for $h < 0$. Consequently, the slope computed from the left would be

$$\begin{aligned} \frac{df}{dx}\Big|_{1^-} &= \lim_{h \to 0^-} \frac{f(1+h) - f(1)}{h} \\ &= \lim_{h \to 0^-} \frac{(1+h)^2 - 1}{h} \\ &= \lim_{h \to 0^-} \frac{1 + 2h + h^2 - 1}{h} \\ &= \lim_{h \to 0^-} \frac{2h + h^2}{h} \\ &= \lim_{h \to 0^-} 2 + h \\ &= 2 + 0 = 2. \end{aligned}$$

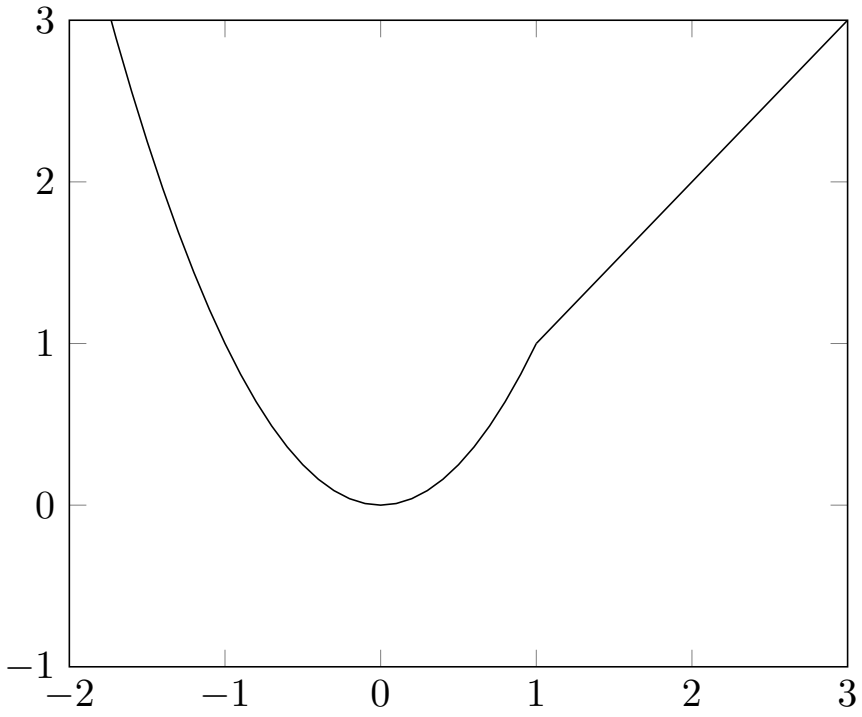
Next, using the function for $x > 1$, we have $f(1 + h) = 1 + h$ when $h > 0$. The slope computed from the right would be

$$\begin{aligned} \frac{df}{dx}\Big|_{1^+} &= \lim_{h \to 0^+} \frac{f(1+h) - f(1)}{h} \\ &= \lim_{h \to 0^+} \frac{(1+h) - 1}{h} \\ &= \lim_{h \to 0^+} \frac{h}{h} \\ &= \lim_{h \to 0^+} 1 \\ &= 1. \end{aligned}$$

With different limits from the left and right,

$$\frac{df}{dx}(1) = \lim_{h \to 0} \frac{f(1+h) - f(1)}{h}$$

does not exist and f is not differentiable at $x = 1$. The figure below illustrates the graph of this function, showing that there is a corner at $x = 1$.



□

Example 8.2.10 Consider the piecewise function defined by

$$f(x) = \begin{cases} x^2 - 3x + 8 & x < 2, \\ 5x - x^2, & x \geq 2. \end{cases}$$

Determine if f is differentiable at $x = 2$.

Solution. This function is continuous because the limit on the left and the limit on the right are equal to the value of the function at $x = 2$, as follows:

$$\begin{aligned} \lim_{x \to 2^-} f(x) &= \lim_{x \to 2} (x^2 - 3x + 8) = 2^2 - 3(2) + 8 = 6, \\ \lim_{x \to 2^+} f(x) &= \lim_{x \to 2} (5x - x^2) = 5(2) - 2^2 = 6, \\ f(2) &= 5(2) - 2^2 = 6. \end{aligned}$$

Now that we know the function is continuous, we can compute the derivative using left- and right-limits. Because $f(x) = x^2 - 3x + 8$ for $x < 2$, we have

$$f(2 + h) = (2 + h)^2 - 3(2 + h) + 8 = 6 + h + h^2$$

when $h < 0$. Consequently, the derivative from the left is

$$\begin{aligned} \frac{df}{dx}\Big|_{2^-} &= \lim_{h \to 0^-} \frac{f(2 + h) - f(2)}{h} \\ &= \lim_{h \to 0^-} \frac{(6 + h + h^2) - 6}{h} \\ &= \lim_{h \to 0^-} \frac{h(1 + h)}{h} \end{aligned}$$

$$\begin{aligned} &= \lim_{h\to 0^-} 1+h \\ &= 1+0=1. \end{aligned}$$

When $h > 0$, we have

$$f(2+h) = 5(2+h) - (2+h)^2 = 6 + h - h^2.$$

The derivative from the right is

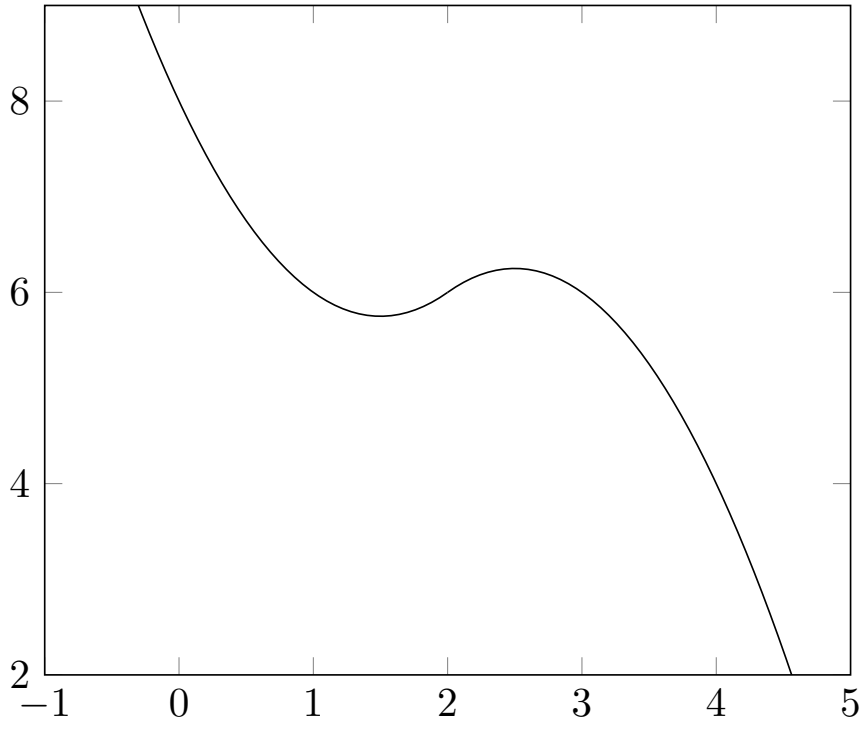
$$\begin{aligned} \frac{df}{dx}\Big|_{2^+} &= \lim_{h\to 0^+} \frac{f(2+h)-f(2)}{h} \\ &= \lim_{h\to 0^+} \frac{(6+h-h^2)-6}{h} \\ &= \lim_{h\to 0^+} \frac{h(1-h)}{h} \\ &= \lim_{h\to 0^+} 1-h \\ &= 1-0=1. \end{aligned}$$

Since the left and right limits computing the left and right derivatives are the same, we conclude that

$$\frac{df}{dx}(2) = \lim_{h\to 0} \frac{f(2+h)-f(2)}{h} = 1$$

So f is differentiable at $x = 2$.

The function consists of two parabolas joined together at $x = 2$. When the left and right derivatives agree, the function transitions smoothly with no corner at $x = 2$.



□

8.2.5 Summary

- The derivative of a function f is the function

$$\frac{df}{dx} : x = a \mapsto \frac{df}{dx}\Big|_a$$

defined by the limit

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

whenever the limit exists. Alternatively,

$$\frac{df}{dx}(x) = \lim_{t \to x} \frac{f(t) - f(x)}{t - x}.$$

- The derivative function $\frac{df}{dx}$ is also written $f'(x)$.
- The value $f'(a)$ is the slope of the tangent line of $y = f(x)$ at $x = a$.
- Saying f is **differentiable** at $x = a$ means $f'(a)$ exists. A function is not differentiable at any point where there is not a well-defined tangent line.
- If a function is differentiable, it must be continuous. If a function has a discontinuity, it will not be differentiable at that point.
- A function can be continuous without being differentiable. For example, a piecewise function that is continuous but that has mismatching slopes at a point (the graph shows a corner) will not be differentiable.

8.2.6 Exercises

For each function, compute the derivative using the definition of the derivative. Use the result to find the equation of the tangent line at $x = 5$.

1. $f(x) = 4x - 7$

2. $g(x) = x^2 + 4x$

3. $k(x) = x^2 - 5x + 2$

4. $p(t) = 3t - t^2$

5. $q(x) = x^3$

6. $V(p) = \dfrac{3}{p+2}$

7. $F(x) = \dfrac{1}{2x-3}$

8. $G(x) = \sqrt{x+4}$

9. $H(x) = \sqrt{3x+1}$

For each piecewise function, determine if it is continuous and differentiable at the break point by evaluating the relevant limits.

10. $f(x) = \begin{cases} 3x - 2, & x \leq 2, \\ x^2 - x, & x > 2. \end{cases}$

11. $g(x) = \begin{cases} 2x - 3, & x \leq 1, \\ x^2 - 2, & x > 1. \end{cases}$

12. $k(x) = \begin{cases} x^2 - 2x, & x < 2, \\ -x^2 + 6x - 8, & x \geq 2. \end{cases}$

Chapter 9

Rules of Differentiation

9.1 Derivative Rules

We have learned that the derivative is defined by the limit of an average rate of change as the gap between the two points goes to zero. For functions already defined as an accumulation function with a known, continuous rate of accumulation, the Fundamental Theorem of calculus guarantees that the derivative equals the rate of accumulation. Every time we need a derivative of any other function, we must use the definition and compute the limit

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

and then go through the algebra and simplification to find the resulting formula.

It would be much nicer if we could look at the formula of $f(x)$ and know what the formula of the derivative $f'(x)$ should be. Computing the formula for $f'(x)$ based on the structure of the formula for $f(x)$ is a process called **differentiation**. Rules for derivatives will provide us with a methodical way to differentiate algebraic formulas.

Differentiation is a process of taking a function and using it to determine another function. That is, differentiation defines a map between functions, $f \mapsto f'$. The entire function $f(x)$, not just a value, is the input to the map and an entirely different function $f'(x)$ is the output. Maps that take numbers as inputs and give numbers as outputs are called **functions**; a map that takes an entire function as an input is called an **operator**. When the independent variable is x, the symbol for the differential operator is $\frac{d}{dx}$. The x in the symbol is replaced by the appropriate independent variable for the function of interest.

Definition 9.1.1 The differential operator $\frac{d}{dx}$ takes a function as its input and provides the derivative function as its output,

$$\frac{d}{dx}[f(x)] = \frac{df}{dx}(x) = f'(x).$$

If y is a dependent variable defined by a function $y = f(x)$, then we can also write

$$\frac{dy}{dx} = \frac{d}{dx}[y] = f'(x).$$

$\Diamond$

In this section, we establish some elementary rules of differentiation. The rules of differentiation begin with linearity, matching the corresponding properties of definite integrals. The similarity ends here. Differentiation also has rules for multiplication and division, where the definite integral has no such rules. Differentiation goes even further and has a rule for function composition, called the chain rule. Each rule is justified by returning to the definition of the derivative using a limit of the difference quotient that represents an average rate of change.

9.1.1 Derivative Building Blocks

In order to differentiate algebraic formulas, we need to know the derivatives of elementary functions that will be our building blocks. Because we have the Fundamental Theorem of Calculus, any rate of accumulation that we know for an accumulation function is automatically going to be a derivative. However, it is also useful to show derivatives of elementary functions directly.

Theorem 9.1.2 Derivative of a Constant. *For a constant k,*

$$\frac{d}{dx}[k] = 0.$$

Proof. The function of interest is $x \mapsto k$, or $f(x) = k$. Using the definition of the derivative,

$$\begin{aligned}\frac{d}{dx}[k] &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h\to 0} \frac{k-k}{h} \\ &= \lim_{h\to 0} 0 \\ &= 0.\end{aligned}$$

The last step in this sequence of equations is a consequence of the Limit Rule for a Constant. ■

Theorem 9.1.3 Derivative of the Identity. *For the identity function $f(x) = x$,*

$$\frac{d}{dx}[x] = 1.$$

Proof. Using the definition of the derivative,

$$\begin{aligned}\frac{d}{dx}[x] &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h\to 0} \frac{x+h-x}{h} \\ &= \lim_{h\to 0} \frac{h}{h} \\ &= \lim_{h\to 0} 1 \\ &= 1.\end{aligned}$$

Again, the limit calculation at the last step uses the Limit Rule for a Constant. ■

The identity function and constant functions are special cases of linear functions.

Theorem 9.1.4 Derivative of Linear Functions. *For a linear function $f(x) = mx + b$,*

$$\frac{d}{dx}[mx+b] = m.$$

Proof. Using the definition of the derivative,

$$\begin{aligned}\frac{d}{dx}[mx+b] &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h\to 0} \frac{(m(x+h)+b) - (mx+b)}{h} \\ &= \lim_{h\to 0} \frac{mx+mh+b-mx-b}{h} \\ &= \lim_{h\to 0} \frac{mh}{h} \\ &= \lim_{h\to 0} m \\ &= m.\end{aligned}$$

■

To establish additional derivatives of elementary formulas, we need now to develop rules that are based on the algebra of combining other formulas.

9.1.2 Overview of the Derivative Rules

Derivative rules are theorems that take as a hypothesis that one or two functions have known derivatives and the conclusion tells how to find the derivative of some combination of those functions. We start by stating the basic rules together for convenience in finding them.

Theorem 9.1.5 Differentiation Rules. *This theorem is a collection of multiple theorems. The hypothesis for any statement that involves $f(x)$ or $g(x)$ is that $\frac{d}{dx}[f(x)] = f'(x)$ or that $\frac{d}{dx}[g(x)] = g'(x)$. Further, k is assumed to be a constant.*

Table 9.1.6 Summary of the Differentiation Rules

$\frac{d}{dx}[k \cdot f(x)] = k \cdot f'(x)$	*Constant Multiple Rule*
$\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$	*Sum Rule*
$\frac{d}{dx}[f(x) - g(x)] = f'(x) - g'(x)$	*Difference Rule*
$\frac{d}{dx}[\frac{1}{g(x)}] = \frac{-g'(x)}{(g(x))^2}$	*Reciprocal Rule*
$\frac{d}{dx}[f(x) \cdot g(x)] = f'(x)g(x) + f(x)g'(x)$	*Product Rule*
$\frac{d}{dx}[\frac{f(x)}{g(x)}] = \frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2}$	*Quotient Rule*
$\frac{d}{dx}[f\Big(g(x)\Big)] = f'\Big(g(x)\Big) \cdot g'(x)$	*Chain Rule*

One of these differentiation rules, the chain rule, will require its own section. That rule is focused on how to differentiate compositions of functions. The other rules focus on arithmetic combinations of functions and are the primary focus of this section. The chain rule was included for completeness in the listing of differentiation rules.

The proofs for these differentiation rules are based on applying the definition of a derivative to the formula in question while knowing that the limits that define the derivatives in the hypothesis are valid. To illustrate, we will look at four of the differentiation rules in detail.

9.1.3 Proofs of Algebraic Differentiation Rules

In each of our proofs for derivative rules, we are going to use the definition of the derivative for the function defined by the conclusion of the rule. We then use algebra to rewrite the difference quotient in a way that the divison by h (and recall $h \to 0$) only appears in expressions where a ratio converges to a known derivative.

In the proofs, in order to keep the algebra a little cleaner, we will use the notation Δf to represent

$$\Delta f = f(x + h) - f(x)$$

and similarly define

$$\Delta g = g(x + h) - g(x).$$

Then, because $f'(x)$ and $g'(x)$ are the derivatives of $f(x)$ and $g(x)$, respectively, we can substitute the following limits

$$\lim_{h \to 0} \frac{\Delta f}{h} = f'(x) \quad \text{and} \quad \lim_{h \to 0} \frac{\Delta g}{h} = g'(x).$$

The first rule we consider is the constant multiple rule. This rules states that if we know how to differentiation a function, then we can compute any

constant multiple of that function by multiplying the derivative by the same constant.

Theorem 9.1.7 Constant Multiple Rule for Derivatives. *If* $\frac{d}{dx}[f(x)] = f'(x)$ *and* k *is a constant, then* $\frac{d}{dx}[kf(x)] = kf'(x)$.

Proof. The rule is interested in finding the rate of change of a new function $k \cdot f(x)$ knowing that $\frac{d}{dx}[f(x)] = f'(x)$. We begin by stating the definition of the derivative of the function $x \mapsto k \cdot f(x)$, and then we use algebra to factor k out as a common factor in the numerator:

$$\begin{aligned}\frac{d}{dx}[k \cdot f(x)] &= \lim_{h\to 0} \frac{k \cdot f(x+h) - k \cdot f(x)}{h} \\ &= \lim_{h\to 0} \frac{k(f(x+h) - f(x))}{h}. \\ &= \lim_{h\to 0} k \cdot \frac{\Delta f}{h}.\end{aligned}$$

Now notice that the formula is a product of the constant k and the average rate of change of f. The rules for limits act with the ordinary rules of arithmetic. In particular, the constant multiple rule for limits states that the limit of a constant times a function equals that constant times the limit of the function. Consequently, we have

$$\begin{aligned}\frac{d}{dx}[k \cdot f(x)] &= \lim_{h\to 0} k\frac{\Delta f}{h} \\ &= k \cdot \lim_{h\to 0} \frac{\Delta f}{h} \\ &= k \cdot f'(x).\end{aligned}$$

■

For our second example of proving a differentiation rule, we consider the reciprocal of a function. We know that $\frac{d}{dx}[x^2 + 3] = 2x$. This reciprocal rule will tell us how to compute the derivative $\frac{d}{dx}[\frac{1}{x^2+3}]$. We might be tempted to think the answer would be $\frac{1}{2x}$, but this is not correct. Derivatives do not follow simple rules for either division or multiplication.

Theorem 9.1.8 Reciprocal Rule for Derivatives. *If* $\frac{d}{dx}[g(x)] = g'(x)$, *then* $\frac{d}{dx}[\frac{1}{g(x)}] = \frac{-g'(x)}{(g(x))^2}$.

Proof. By hypothesis, $\frac{d}{dx}[g(x)] = g'(x)$. This means that $g'(x)$ is defined by its limit

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

The rule is interested in finding the rate of change of a new function $x \mapsto \frac{1}{g(x)}$. We will use that function to compute the derivative using the definition, which will require finding a common denominator.

$$\begin{aligned}\frac{d}{dx}[\frac{1}{g(x)}] &= \lim_{h\to 0} \frac{\frac{1}{g(x+h)} - \frac{1}{g(x)}}{h} \\ &= \lim_{h\to 0} \frac{\frac{g(x)}{g(x)g(x+h)} - \frac{g(x+h)}{g(x)g(x+h)}}{h}\end{aligned}$$

$$= \lim_{h\to 0} \frac{g(x) - g(x+h)}{g(x)g(x+h)} \cdot \frac{1}{h}$$
$$= \lim_{h\to 0} \frac{-\Delta g}{g(x)g(x+h)} \cdot \frac{1}{h}$$
$$= \lim_{h\to 0} \frac{-1}{g(x)g(x+h)} \cdot \frac{\Delta g}{h}$$

Since the limit involves $h \to 0$, $g(x)$ is a constant relative to the limit. In addition, because a differentiable function is continuous, we have $g(x+h) \to g(x)$ as $x \to h$. Consequently, the limit rules for reciprocals and constant multiples imply

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

Using the limit rule for a product, we have

$$\frac{d}{dx}[\frac{1}{g(x)}] = \lim_{h\to 0} \frac{-1}{g(x)g(x+h)} \cdot \frac{g(x+h) - g(x)}{h}$$
$$= \frac{-1}{(g(x))^2} \cdot g'(x) = \frac{-g'(x)}{(g(x))^2}.$$

■

Example 9.1.9 Find $\frac{d}{dx}[\frac{1}{x^2+3}]$.

Solution. We start by recognizing the formula $\frac{1}{x^2+3}$ as the reciprocal of $g(x) = x^2 + 3$. We know $g'(x) = 2x$, so the Theorem 9.1.8 gives

$$\frac{d}{dx}[\frac{1}{x^2+3}] = \frac{-g'(x)}{(g(x))^2}$$
$$= \frac{-2x}{(x^2+3)^2}.$$

That is, if $f(x) = \frac{1}{x^2+3}$, the derivative is $f'(x) = \frac{-2x}{(x^2+3)^2}$. □

In the solution to the previous example, we introduced a name for a function for the sole reason of being able to refer to its derivative. This is one of the primary reasons for introducing the differentiation operator $\frac{d}{dx}$. It allows us to refer to derivatives using the operator with the original function as input. The work in the example could be rewritten

$$\frac{d}{dx}[\frac{1}{x^2+3}] = \frac{-\frac{d}{dx}[x^2+3]}{(x^2+3)^2}$$
$$= \frac{-2x}{(x^2+3)^2}.$$

The two rules of differentiation proved thus far involve operations on a single function. We now turn our attention to rules that combine multiple functions. The first rule we consider is the sum rule, which states that the derivative of a function formed by adding two functions will be the sum of those functions' derivatives.

Theorem 9.1.10 Sum Rule for Derivatives. *If* $\frac{d}{dx}[f(x)] = f'(x)$ *and* $\frac{d}{dx}[g(x)] = g'(x)$, *then* $\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$.

Proof. By hypothesis, $\frac{d}{dx}[f(x)] = f'(x)$ and $\frac{d}{dx}[g(x)] = g'(x)$. This means

that

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

The sum rule is interested in finding the rate of change of a new function $x\mapsto f(x)+g(x)$. Because $\Delta f=f(x+h)-f(x)$, we can rewrite $f(x+h)=f(x)+\Delta f$. Similarly, we can rewrite $g(x+h)=g(x)+\Delta g$. When we use the definition of the derivative, we find

$$\begin{aligned}\frac{d}{dx}[f(x)+g(x)]&=\lim_{h\to 0}\frac{[f(x+h)+g(x+h)]-[f(x)+g(x)]}{h}\\&=\lim_{h\to 0}\frac{f(x)+\Delta f+g(x)+\Delta g-f(x)-g(x)}{h}\\&=\lim_{h\to 0}\frac{\Delta f+\Delta g}{h}\\&=\lim_{h\to 0}\left[\frac{\Delta f}{h}+\frac{\Delta g}{h}\right].\end{aligned}$$

Again, because limit rules satisfy the ordinary rules of arithmetic, the limit rule for sums implies

$$\begin{aligned}\frac{d}{dx}[f(x)+g(x)]&=\lim_{h\to 0}\left[\frac{\Delta f}{h}+\frac{\Delta g}{h}\right]\\&=\lim_{h\to 0}\left[\frac{\Delta f}{h}\right]+\lim_{h\to 0}\left[\frac{\Delta g}{h}\right]\\&=f'(x)+g'(x).\end{aligned}$$

■

For our final example of a proof of a differentiation rule, we consider the derivative of a product. Consider a function like $f(x)=(x^2-3)\cdot(2x+1)$. It is tempting to take derivatives of each formula in place and assume that $f'(x)$ would be $(2x)\cdot(2)=4x$. We can see that this is incorrect if we rewrote $f(x)$ after expanding the product,

$$f(x)=2x^3+x^2-6x-3.$$

Once written as a simple polynomial, our experience with accumulation functions and the Fundamental Theorem of Calculus allows us to recognize

$$f'(x)=6x^2+2x-6.$$

Proper differentiation rules will be consistent regardless of how a function is represented. For a function that is represented as a product of two other functions, the product rule shows that the derivative is a sum of contributions resulting from the rate of change of each factor.

Theorem 9.1.11 Product Rule for Derivatives. *If* $\frac{d}{dx}[f(x)]=f'(x)$ *and* $\frac{d}{dx}[g(x)]=g'(x)$, *then* $\frac{d}{dx}[f(x)\cdot g(x)]=f'(x)\cdot g(x)+f(x)\cdot g'(x)$.

Proof. By hypothesis, $\dfrac{d}{dx}[f(x)]=f'(x)$ and $\dfrac{d}{dx}[g(x)]=g'(x)$. This means that

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

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

The product rule is interested in finding the rate of change of a new function $x\mapsto f(x)g(x)$. As we did in the sum rule, we will take advantage of rewriting $f(x+h)=f(x)+\Delta f$ and $g(x+h)=g(x)+\Delta g$. When $f'(x)$ and $g'(x)$ both exist, f and g are both continuous so that

$$\lim_{h\to 0}\Delta f=\lim_{h\to 0}f(x+h)-f(x)=0,$$
$$\lim_{h\to 0}\Delta g=\lim_{h\to 0}g(x+h)-g(x)=0.$$

The derivative in question is defined by

$$\begin{aligned}
\frac{d}{dx}[f(x)g(x)] &= \lim_{h\to 0}\frac{[f(x+h)g(x+h)]-[f(x)g(x)]}{h}\\
&= \lim_{h\to 0}\frac{(f(x)+\Delta f)(g(x)+\Delta g)-f(x)g(x)}{h}\\
&= \lim_{h\to 0}\frac{f(x)g(x)+\Delta f g(x)+f(x)\Delta g+\Delta f\Delta g-f(x)g(x)}{h}\\
&= \lim_{h\to 0}\frac{\Delta f g(x)+f(x)\Delta g+\Delta f\Delta g}{h}\\
&= \lim_{h\to 0}\left[\frac{\Delta f g(x)}{h}+\frac{f(x)\Delta g}{h}+\frac{\Delta f\Delta g}{h}\right]\\
&= \lim_{h\to 0}\left[\frac{\Delta f}{h}\cdot g(x)+f(x)\cdot\frac{\Delta g}{h}+\Delta f\cdot\frac{\Delta g}{h}\right].\\
&= f'(x)\cdot g(x)+f(x)\cdot g'(x)+0\cdot g'(x)\\
&= f'(x)\cdot g(x)+f(x)\cdot g'(x),
\end{aligned}$$

using the Limit Rule of a Sum and the Limit Rule of a Product. ■

Example 9.1.12 Show that the derivative using the product rule for $\frac{d}{dx}[(x^2-3)\cdot(2x+1)]$ is consistent with first expanding and then differentiating the polynomial.

Solution. The function $f(x)=(x^2-3)\cdot(2x+1)$ is a product of $u=x^2-3$ and $v=2x+1$. The product rule informs us that $\frac{d}{dx}[uv]=\frac{du}{dx}v+u\frac{dv}{dx}$:

$$\begin{aligned}
\frac{d}{dx}[(x^2-3)\cdot(2x+1)] &= \frac{d}{dx}[x^2-3]\cdot(2x+1)+(x^2-3)\cdot\frac{d}{dx}[2x+1]\\
&= (2x)\cdot(2x+1)+(x^2-3)\cdot(2)\\
&= 4x^2+2x+2x^2-6\\
&= 6x^2+2x-6
\end{aligned}$$

We saw prior to the theorem that $f(x)=(x^2-3)(2x+1)=2x^3+x^2-6x-3$ has a derivative $f'(x)=6x^2+2x-6$, which is consistent with the result we obtained using the product rule. □

The quotient rule is a combination of the product rule and the reciprocal rule.

Theorem 9.1.13 Quotient Rule for Derivatives. *If* $\frac{d}{dx}[f(x)]=f'(x)$ *and* $\frac{d}{dx}[g(x)]=g'(x)$, *then* $\dfrac{d}{dx}\left[\dfrac{f(x)}{g(x)}\right]=\dfrac{f'(x)g(x)-f(x)g'(x)}{(g(x))^2}$.

9.1.4 Applying the Rules to Formulas

In this section, we have established rules of differentiation for elementary formulas and for algebraic combinations of functions with known derivatives. We now consider how we can apply these rules together to compute derivatives from more complex formulas.

To illustrate how functions are formed as combinations of elementary functions, let us revisit the derivative of a linear function, $f(x) = mx + b$. This function is algebraically the *sum* of the expressions mx and b. Consequently, the derivative will be the sum of the derivatives of those expressions. The first, mx is a constant multiple of the identity function, so

$$\frac{d}{dx}[mx] = m\frac{d}{dx}[x] = m \cdot 1 = m.$$

The second, b is a constant function, so

$$\frac{d}{dx}[b] = 0.$$

Adding these together, we find

$$\frac{d}{dx}[mx + b] = \frac{d}{dx}[mx] + \frac{d}{dx}[b] = m + 0 = m,$$

exactly the same as we found applying the definition of the derivative in Theorem 9.1.4.

Example 9.1.14 Use the differentiation rules to show

$$\frac{d}{dx}[x^2] = 2x.$$

Solution. The function that is the input to the differentiation operator is x^2. The elementary building blocks so far only include constant functions, the identity function, and other linear functions. We need to see how x^2 is a combination of these elementary functions. To do this, we need to recognize that the square corresponds to multiplication,

$$x^2 = x \cdot x.$$

Once we recognize that our function is a product, we use the product rule with $f(x) = x$ and $g(x) = x$. The product rule says

$$\frac{d}{dx}[f(x)g(x)] = f'(x)g(x) + f(x)g'(x),$$

for which we use $f'(x) = 1$ and $g'(x) = 1$. Consequently,

$$\begin{aligned}\frac{d}{dx}[x^2] &= \frac{d}{dx}[x \cdot x] \\ &= 1 \cdot x + x \cdot 1 \\ &= x + x = 2x.\end{aligned}$$

□

We can repeat this process to find the derivative of x^3 and then x^4. The pattern generalizes to a rule that we call the power rule.

Example 9.1.15 Continue to use the product rule of derivatives to show that

$$\frac{d}{dx}[x^3] = 3x^2,$$
$$\frac{d}{dx}[x^4] = 4x^3.$$

Solution. We rewrite the power as products:

$$x^3 = x \cdot x^2, \quad x^4 = x \cdot x^3.$$

We already know

$$\frac{d}{dx}[x] = 1,$$
$$\frac{d}{dx}[x^2] = 2x.$$

It is useful to remember the product rule using dependent variables instead of functions. That is, if $u = f(x)$ and $v = g(x)$, then the product rule becomes

$$\frac{d}{dx}[u \cdot v] = \frac{du}{dx} \cdot v + u \cdot \frac{dv}{dx}.$$

because this will guide our use of the differentiation operator.

The derivative of $x^3 = x \cdot x^2$ will use $u = x$ and $v = x^2$:

$$\begin{aligned}\frac{d}{dx}[x^3] &= \frac{d}{dx}[x \cdot x^2] \\ &= \frac{d}{dx}[x] \cdot x^2 + x \cdot \frac{d}{dx}[x^2] \\ &= 1 \cdot x^2 + x \cdot (2x) \\ &= 3x^2\end{aligned}$$

The derivative of $x^4 = x \cdot x^3$ uses $u = x$ and $v = x^3$, whose derivative we learned just above.

$$\begin{aligned}\frac{d}{dx}[x^4] &= \frac{d}{dx}[x \cdot x^3] \\ &= \frac{d}{dx}[x] \cdot x^3 + x \cdot \frac{d}{dx}[x^3] \\ &= 1 \cdot x^3 + x \cdot (3x^2) \\ &= 4x^3\end{aligned}$$

Notice that the derivative rules are self consistent. We could have written $x^4 = x^2 \cdot x^2$, and the product rule would still have given the same answer.

$$\begin{aligned}\frac{d}{dx}[x^4] &= \frac{d}{dx}[x^2 \cdot x^2] \\ &= \frac{d}{dx}[x^2] \cdot x^2 + x^2 \cdot \frac{d}{dx}[x^2] \\ &= 2x \cdot x^2 + x^2 \cdot 2x \\ &= 4x^3.\end{aligned}$$

□

We can continue to find more derivatives using these results. In particular,

all of the polynomials that we learned earlier in terms of the rates of accumulation for accumulation functions, we now can justify as derivatives using the derivative rules.

Example 9.1.16 Find $\frac{d}{dx}[x^3 + 5x^2 - 8x + 3]$.

Solution. It is helpful to give the original function a name, so we define $f(x) = x^3 + 5x^2 - 8x + 3$. We start by using the sum rule of derivatives. That rule was formulated with adding two formulas together. Consequently, we need to repeat our use of the rule, breaking up the sum into two parts at a time.

$$\begin{aligned}\frac{df}{dx} &= \frac{d}{dx}[x^3 + 5x^2 - 8x + 3] \\ &= \frac{d}{dx}[x^3] + \frac{d}{dx}[5x^2 - 8x + 3] \\ &= \frac{d}{dx}[x^3] + \frac{d}{dx}[5x^2] + \frac{d}{dx}[-8x + 3]\end{aligned}$$

We can stop at this point with the sum rule because $-8x + 3$ is a linear function, and we have a derivative rule for any linear function. We can use the constant multiple rule to factor the 5 from the derivatives of x^2, and then use the derivatives that we know.

$$\begin{aligned}f'(x) &= \frac{d}{dx}[x^3] + 5\frac{d}{dx}[x^2] + \frac{d}{dx}[-8x + 3] \\ &= 3x^2 + 5(2x) + -8 \\ &= 3x^2 + 10x - 8.\end{aligned}$$

We have shown $\frac{d}{dx}[x^3 + 5x^2 - 8x + 3] = 3x^2 + 10x - 8$. □

The previous example was one that we learned previously using accumulation functions. Because definite integrals also have constant multiple and sum rules, the process we used earlier is essentially the same. The new differentiation rules really show their value in finding derivatives of functions we that are not written as a sum.

Example 9.1.17 Find $\frac{d}{dx}[\frac{1}{x^3}]$.

Solution. We use the reciprocal rule of derivatives,

$$\frac{d}{dx}[\frac{1}{u}] = \frac{-\frac{du}{dx}}{u^2}.$$

So our derivative is

$$\begin{aligned}\frac{d}{dx}[\frac{1}{x^3}] &= \frac{-\frac{d}{dx}[x^3]}{(x^3)^2} \\ &= \frac{-(3x^2)}{x^3 \cdot x^3} = \frac{-3x^2}{x^6} \\ &= \frac{-3}{x^4}.\end{aligned}$$

□

We have found derivative formulas for quite a few different elementary powers now. One of the themes in mathematics is to look for patterns and then determine whether that pattern would always hold. Consider the following

sequence of statements that we have proved:

$$\frac{d}{dx}[x^2] = 2x,$$
$$\frac{d}{dx}[x^3] = 3x^2,$$
$$\frac{d}{dx}[x^4] = 4x^3.$$

There appears to be a pattern that the derivative of the power includes as a constant multiple the value of the power and another power of the variable that is one lower than the original function.

How does this relate to more other expressions? The identity function can also be thought of as a power, $x = x^1$, and we can rewrite the derivative rule for the identity as

$$\frac{d}{dx}[x^1] = 1x^0 = 1.$$

We can think of reciprocals of powers as equivalent negative powers. For example, the reciprocal rule guarantees

$$\frac{d}{dx}[\frac{1}{x}] = \frac{-1}{x^2}$$

and we just showed

$$\frac{d}{dx}[\frac{1}{x^3}] = \frac{-3}{x^4}.$$

If we rewrote these derivatives in the form of simple negative powers, we discover the pattern continues:

$$\frac{d}{dx}[x^{-1}] = -1x^{-2},$$
$$\frac{d}{dx}[x^{-3}] = -3x^{-4}.$$

When learning about the definition of the derivative, we found the derivative of the square root function,

$$\frac{d}{dx}[\sqrt{x}] = \frac{1}{2\sqrt{x}}.$$

By rewriting the square root as a fractional power, we discover that even this rule is following the same pattern.

$$\frac{d}{dx}[x^{1/2}] = \frac{1}{2}x^{-1/2}.$$

We have seen seven examples that appear to follow the same pattern. Inductive reasoning is the process of using examples to develop a generalization that we belief might be true. For this example, inductive reasoning would lead us to a **conjecture** that for *any* power,

$$\frac{d}{dx}[x^p] = px^{p-1}.$$

Deductive reasoning is the process of establishing the truth of such a statement built on a logical argument, or a proof, that applies the definitions and other proved conclusions to show whether or not that conjecture is true.

In this case, the claim can be proved true. We will prove the result in stages. First, we will generalize to all positive integer powers. Next, we will show that the result for positive integer powers implies a similar result for negative integer powers. As we continue to develop calculus, we will show our result is true for rational powers and ultimately for any real number.

Theorem 9.1.18 Power Rule for Derivatives. *For any real number p,*

$$\frac{d}{dx}[x^p] = px^{p-1}.$$

Proof. As indicated, we currently are only ready to prove this theorem for the integer powers. We start with positive integers. We have already proved the result for $p = 1, 2, 3, 4$ as part of our discovering the pattern. As we developed that pattern, we discovered that we were using a recursive argument each time. The powers $p = 3$ and $p = 4$ were based on knowing the results for $p = 2$ and $p = 3$, respectively. Our proof builds on this recursive argument to create a general statement.

Suppose that we know for $p = n$,

$$\frac{d}{dx}[x^n] = nx^{n-1}.$$

Next consider the power $p = n + 1$, and rewrite it $x^{n+1} = x \cdot x^n$. The product rule guarantees

$$\begin{aligned}\frac{d}{dx}[x^{n+1}] &= \frac{d}{dx}[x \cdot x^n] \\ &= \frac{d}{dx}[x] \cdot x^n + x \cdot \frac{d}{dx}[x^n] \\ &= 1 \cdot x^n + x \cdot nx^{n-1} \\ &= x^n + nx^n \\ &= (n+1)x^n.\end{aligned}$$

This general recursive argument shows that if the rule is satisfied for an initial value of $p = n$, the statement will be immediately known to be true for the sequence of values $p \in (n, n+1, n+2, \ldots)$. Having earlier shown the rule was true for $p = 1$, the recursive argument shows it will be true for all positive integers. In addition, because we know the rule is true for $p = \frac{1}{2}$, the same argument shows that the rule is true for all $p \in (\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots)$.

The reciprocal rule for derivatives allows us to use the rule for positive integers to show the pattern holds for negative integers. For a positive integer n, consider the corresponding negative power $p = -n$ as a reciprocal.

$$\begin{aligned}\frac{d}{dx}[x^p] = \frac{d}{dx}[x^{-n}] &= \frac{d}{dx}[\frac{1}{x^n}] \\ &= \frac{-\frac{d}{dx}[x^n]}{(x^n)^2} \\ &= \frac{-nx^{n-1}}{x^{2n}} \\ &= -nx^{n-1} \cdot x^{-2n} = -nx^{n-1-2n} \\ &= -nx^{-n-1} = px^{p-1}.\end{aligned}$$

The argument did not depend so much on n being an integer as it required that we knew the power rule was true for $p = n$. Consequently, the same argument proves that the power rule is also true for $p \in (-\frac{1}{2}, -\frac{3}{2}, \ldots)$. ■

9.1.5 Exercises

Each function is an algebraic combination of more elementary expressions. Identify the last operation in the expression and the component expressions

that operation combines. Repeat the process for each of the component expressions.

For example, $3x^2 - 8x$ involves a final operation of subtraction involving the terms $3x^2$ and $8x$; $3x^2$ is a constant multiple of the expression x^2 with the constant 3; and $8x$ is a constant multiple of the identity x with constant 8.

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

2. $f(x) = (x^2 + 5x)(x^3 - 7)$

3. $g(x) = \dfrac{2x^3}{4x + 1}$

Show that the derivative of each product is the same whether the function is expanded into a sum before differentiation or the product rule is used on the original formula.

4. $P(x) = (2x + 5)(3x - 4)$

5. $Q(t) = (3t - 1)(t^2 + 4t - 5)$

6. $R(y) = (y^2 - 2)(y^2 + 2)$

Compute the derivatives.

7. $\dfrac{d}{dx}[4\sqrt{x}]$

8. $\dfrac{d}{dx}[2x^{5/2} - 5x^{3/2}]$

9. $\dfrac{d}{dt}[\dfrac{1}{t^2 + 4t}]$

10. $\dfrac{d}{dt}[\dfrac{5}{3t - 1}]$

11. $\dfrac{d}{dr}[\dfrac{r}{r^2 + 4}]$

Applications

12. Find the tangent line for $y = \frac{x-1}{x+1}$ at $x = 2$.

13. Find the tangent line for $y = \sqrt{x}$ at $x = 100$.

14. Find all points on the graph $y = \dfrac{1}{x}$ such that the slope of the tangent line has slope -4.

15. The height y (feet) from the ground of an object tossed from a tower is a quadratic function of time t (seconds) given by

$$y = 50 + 40t - 16t^2.$$

(a) Determine the velocity at which the object is thrown. (Velocity is the instantaneous rate of change of height.)

(b) Find the time when the object is traveling at the same speed but opposite direction as when it was thrown.

(c) Find the time such that the velocity is equal to the average velocity over the first two seconds of flight.

9.2 Differentiation and Related Rates

The rules of differentiation provide directions for how a desire rate of change is computed relative to the rates of change of its components. We often think of these rules in terms of differentiating formulas. However, because a derivative is a function that gives the instantaneous rate of change, the rules also apply to any instantaneous rate of change of a dependent variable that is made from other variables.

In this section, we will develop our understanding of the differentiation rules. First, we focus on how the rules apply to formulas. That is, given the explicit formula for a function, we can compute the explicit formula for its derivative. Then we study related rates. In that setting, we do not have explicit formulas for the dependent variable of interest. Instead, we know how the variable relates to other dependent variables. If we know the instantaneous rates of change of the related variables, then the differentiation rules will allow us to compute the instantaneous rate of change of our variable of interest.

9.2.1 Derivatives Take Practice

I want to recommend that you practice as much as possible. You might find it useful to do some of this practice using the following web-based app that will also work on smart phones or tablets: Derivative Practice on Algebraic Formulas. Work your way until you can do all of the types of calculations without hesitation.

Start by knowing basic derivatives of power functions using the power rule

$$\frac{d}{dx}[x^p] = px^{p-1}.$$

We looked at why this rule is true when p is a positive integer, but the rule is true for any power function. Combining this with the constant multiple rule, you can find the derivative

$$\frac{d}{dx}[Ax^p] = Apx^{p-1}.$$

Example 9.2.1 Compute the following derivatives:

1. $\frac{d}{dx}[5x^3]$
2. $\frac{d}{dx}[\frac{x^4}{7}]$
3. $\frac{d}{dx}[\frac{2}{7x^2}]$

Solution.

1. To compute $\frac{d}{dx}[5x^3]$, we recognize the elementary power x^3 which has power $p = 3$ so that its derivative is $\frac{d}{dx}[x^3] = 3x^2$. Use the constant multiple rule to get the final derivative.

$$\frac{d}{dx}[5x^3] = 5(3x^2) = 15x^2.$$

2. To compute $\frac{d}{dx}[\frac{x^4}{7}]$, we recognize the elementary power x^4 which has power $p = 4$ so that its derivative is $\frac{d}{dx}[x^4] = 4x^3$. The fraction is a

constant multiple in disguise with constant $\frac{1}{7}$.

$$\frac{d}{dx}[\frac{x^4}{7}] = \frac{d}{dx}[\frac{1}{7}x^4] = \frac{1}{7}(4x^3) = \frac{4}{7}x^3.$$

3. To compute $\frac{d}{dx}[\frac{2}{7x^2}]$, we use the properties of powers to rewrite division by a power as a negative power,

$$\frac{2}{7x^2} = \frac{2}{7}x^{-2}.$$

The basic power $p = -2$ has a derivative with new power $p - 1 = -3$, so

$$\frac{d}{dx}[\frac{2}{7x^2}] = \frac{d}{dx}[\frac{2}{7}x^{-2}] = \frac{2}{7}(-2x^{-3}) = \frac{-4}{7x^3}.$$

□

Once you have mastered these elementary building blocks with the constant multiple rule, you can move to sums of these building blocks. Derivatives behave nicely with sums, since the derivative of a sum is the sum of the derivatives,

$$\frac{d}{dx}[f(x) + g(x)] = \frac{d}{dx}[f(x)] + \frac{d}{dx}[g(x)] = f'(x) + g'(x).$$

In practice, this means that as soon as you recognize a function is combined as a sum of elementary parts, you can just compute the derivatives of each part separately and add the results. (Subtraction is just addition with an inverse, so both are done at the same time.)

Example 9.2.2 Compute the following derivatives:

1. $\frac{d}{dx}[2x^5 - 3x^3 + 5x^2 + 7]$
2. $\frac{d}{dx}[x^2 + 3x - \frac{1}{2x^3}]$

Solution. For each problem, pay attention to how the differentiation operator is applied, starting from the entire formula to individual components until the ultimate answer is found.

1.

$$\begin{aligned}\frac{d}{dx}[2x^5 - 3x^3 + 5x^2 + 7] &= \frac{d}{dx}[2x^5] + \frac{d}{dx}[-3x^3] + \frac{d}{dx}[5x^2] + \frac{d}{dx}[7] \\ &= 2(5x^4) + -3(3x^2) + 5(2x) + 0 \\ &= 10x^4 - 9x^2 + 10x\end{aligned}$$

2.

$$\begin{aligned}\frac{d}{dx}[x^2 + 3x - \frac{1}{2x^3}] &= \frac{d}{dx}[x^2] + \frac{d}{dx}[3x] + \frac{d}{dx}[-\frac{1}{2}x^{-3}] \\ &= 2x + 3 + \frac{-1}{2}(-3x^{-4}) \\ &= 2x + 3 + \frac{3}{2x^4}\end{aligned}$$

□

Now that you can compute derivatives of sums of elementary terms, you should practice computing derivatives of products. The product rule for derivatives do not follow the same simple rule as sums. A little memorization jingle that might help is, "The derivative of u times v is U dee-V plus V dee-U," which as formula is

$$\frac{d}{dx}[u \cdot v] = u\frac{dv}{dx} + v\frac{du}{dx}.$$

Others like to say, "First D-Last plus Last D-First." Alternatively, I personally use a tactile approach where I touch each factor one at a time and write down a new product where I replace the factor I am touching with its derivative and leave all other factors alone, adding the results. For a product of u and v, I would write

$$\frac{d}{dx}[u \cdot v] = \frac{du}{dx} \cdot v + u \cdot \frac{dv}{dx},$$

and for a product of three terms, u, v, w, I would write

$$\frac{d}{dx}[u \cdot v \cdot w] = \frac{du}{dx} \cdot v \cdot w + u \cdot \frac{dv}{dx} \cdot w + u \cdot v \cdot \frac{dw}{dx}.$$

Example 9.2.3 Compute the following derivatives:

1. $\frac{d}{dx}[(2x+5)(3x-7)]$
2. $\frac{d}{dx}[x^2(x^3+5)]$
3. $\frac{d}{dx}[4x^2(3x-1)(4x+5)]$

Solution. For each problem, continue to watch how the differentiation operator is applied, starting from the entire formula to individual components until the ultimate answer is found.

1.

$$\begin{aligned}\frac{d}{dx}[(2x+5)(3x-7)] &= \frac{d}{dx}[2x+5] \cdot (3x-7) + (2x+5) \cdot [3x-7] \\ &= 2(3x-7) + (2x+5)(3) \\ &= 6x - 14 + 6x + 15 \\ &= 12x + 1\end{aligned}$$

2.

$$\begin{aligned}\frac{d}{dx}[x^2(x^3+5)] &= \frac{d}{dx}[x^2] \cdot (x^3+5) + x^2 \cdot \frac{d}{dx}[x^3+5] \\ &= (2x)(x^3+5) + x^2(3x^2+0) \\ &= 2x^4 + 10x + 3x^4 \\ &= 5x^4 + 10x\end{aligned}$$

3.

$$\begin{aligned}\frac{d}{dx}[4x^2(3x-1)(4x+5)] &= \frac{d}{dx}[4x^2](3x-1)(4x+5) \\ &\quad + (4x^2)\frac{d}{dx}[3x-1](4x+5) \\ &\quad + (4x^2)(3x-1)\frac{d}{dx}[4x+5]\end{aligned}$$

$$\begin{aligned}
&= (8x)(3x-1)(4x+5) + (4x^2)(3)(4x+5) + (4x^2)(3x-1)(4)\\
&= 8x(12x^2+15x-4x-5) + 12x^2(4x+5) + 16x^2(3x-1)\\
&= 96x^3 + 88x^2 - 40x + 48x^3 + 60x^2 + 48x^3 - 16x^2\\
&= 192x^3 + 132x^2 - 40x
\end{aligned}$$

In each of these examples, it would also be possible to multiply out the formulas before taking a derivative. This is often easier because then you only need to use the sum rule rather than the product rule.

1.
$$\begin{aligned}
\frac{d}{dx}[(2x+5)(3x-7)] &= \frac{d}{dx}[6x^2 - 14x + 15x - 35]\\
&= \frac{d}{dx}[6x^2 + x - 35]\\
&= 12x + 1
\end{aligned}$$

2.
$$\begin{aligned}
\frac{d}{dx}[x^2(x^3+5)] &= \frac{d}{dx}[x^5 + 5x^2]\\
&= 5x^4 + 10x
\end{aligned}$$

3.
$$\begin{aligned}
\frac{d}{dx}[4x^2(3x-1)(4x+5)] &= \frac{d}{dx}[4x^2(12x^2+15x-4x-5)]\\
&= \frac{d}{dx}[48x^4 + 44x^3 - 20x^2]\\
&= 48(4x^3) + 44(3x^2) - 20(2x)\\
&= 192x^3 + 132x^2 - 40x
\end{aligned}$$

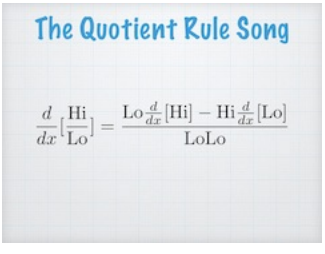
However, it is good to practice the product rule for those cases later where it is not possible to expand a formula so that the product rule isn't necessary. □

After the product rule, you should master the quotient rule,

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2}.$$

I like to sing it as a song. In symbols, this rhyme would be written

$$\frac{d}{dx}\left[\frac{\text{Hi}}{\text{Lo}}\right] = \frac{\text{Lo}\frac{d}{dx}[\text{Hi}] - \text{Hi}\frac{d}{dx}[\text{Lo}]}{\text{LoLo}}.$$



YouTube: https://www.youtube.com/watch?v=P7sKe46F8kY

Figure 9.2.4 Lo D Hi minus Hi D Lo over Lo Lo.

Example 9.2.5 Compute the following derivatives:

1. $\dfrac{d}{dx}\left[\dfrac{2x+5}{3x-7}\right]$
2. $\dfrac{d}{dx}\left[\dfrac{x^2}{x^3+5}\right]$

Solution. Applying the quotient rule for derivatives leads to each answer. You do not need to expand the square of the denominator, but you should simplify the numerator.

1.
$$\begin{aligned}\frac{d}{dx}\left[\frac{2x+5}{3x-7}\right] &= \frac{(3x-7)\frac{d}{dx}[2x+5]-(2x+5)\frac{d}{dx}[3x-7]}{(3x-7)^2}\\ &= \frac{(3x-7)(2)-(2x+5)(3)}{(3x-7)^2}\\ &= \frac{6x-14-6x-15}{(3x-7)^2}\\ &= \frac{-29}{(3x-7)^2}\end{aligned}$$

2.
$$\begin{aligned}\frac{d}{dx}\left[\frac{x^2}{x^3+5}\right] &= \frac{(x^3+5)\frac{d}{dx}[x^2]-x^2\frac{d}{dx}[x^3+5]}{(x^3+5)^2}\\ &= \frac{(x^3+5)(2x)-x^2(3x^2)}{(x^3+5)^2}\\ &= \frac{2x^4+10x-3x^4}{(x^3+5)^2}\\ &= \frac{-x^4+10x}{(x^3+5)^2}\end{aligned}$$

□

9.2.2 Related Rates

We often fall into a trap thinking that the rules of differentiation apply only to formulas. Some times, two or more quantities are added together to form a new quantity representing their sum. Other times, a quantity of interest is determined by multiplying the values of two measurements. The rules of differentiation apply to any setting where we are interested in how the rate of change a quantity relates to the rates of change of quantities with which it is related. If we know the instantaneous values and rates of change of these quantities, we can find the instantaneous rate of change of the new variable even when we do not know any formulas for our underlying variables.

Example 9.2.6 Suppose $y = t^2 f(t) - 3g(t)$ and $z = \dfrac{2f(t)}{g(t)+1}$, where the functions f and g are not known explicitly. However, we do know the following values at specific times, as shown in a table. Find $\left.\frac{dy}{dt}\right|_2$ and $\left.\frac{dz}{dt}\right|_4$.

Table 9.2.7 Values of the functions f and g and their derivatives at specified points.

t	0	1	2	3	4	5
$f(t)$	3	1	5	2	-1	-3
$f'(t)$	-2	4	-3	-3	2	6
$g(t)$	1	3	5	6	4	2
$g'(t)$	2	3	1	-1	-4	-2

Solution. We start with what we know, the equation $y = t^2 f(t) - 3g(t)$. We may not know the explicit formulas for $f(t)$ or $g(t)$, but we do know the algebraic operations that put the formulas together. In this problem, t is the independent variable. The differentiation operator will therefore be $\frac{d}{dt}$.

The last operation used to form y is addition (i.e. subtraction) of $t^2 f(t)$ and $-3g(t)$. The sum rule of derivatives 9.1.10 then allows us to write

$$\frac{d}{dt}[y] = \frac{d}{dt}[t^2 f(t)] + \frac{d}{dt}[-3g(t)].$$

The expression $t^2 f(t)$ is a product of t^2 and $f(t)$, so the product rule 9.1.11 tells us

$$\begin{aligned}\frac{d}{dt}[t^2 f(t)] &= \frac{d}{dt}[t^2] \cdot f(t) + t^2 \cdot \frac{d}{dt}[f(t)] \\ &= 2tf(t) + t^2 f'(t).\end{aligned}$$

Meanwhile, $-3g(t)$ is a constant multiple 9.1.7 of $g(t)$ so that

$$\frac{d}{dt}[-3g(t)] = -3\frac{d}{dt}[g(t)] = -3g'(t).$$

Putting all of these together in a single statement, we obtain

$$\frac{dy}{dt} = 2tf(t) + t^2 f'(t) - 3g'(t).$$

Now that we have the equation relating the different rates expressed as derivatives, we can use our data from the table to find actual instantaneous rates of change. When $t = 2$, we find

$$\begin{aligned}\left.\frac{dy}{dt}\right|_2 &= 2(2)f(2) + 2^2 f'(2) - 3g'(2) \\ &= 4(5) + 4(-3) - 3(1) = 5.\end{aligned}$$

In a similar way, we know $z = \dfrac{2f(t)}{g(t)+1}$ is a quotient. The quotient rule tells us

$$\begin{aligned}\frac{dz}{dt} &= \frac{(g(t)+1)\frac{d}{dt}[2f(t)] - (2f(t))\frac{d}{dt}[g(t)+1]}{(g(t)+1)^2} \\ &= \frac{2(g(t)+1)f'(t) - 2f(t)g'(t)}{(g(t)+1)^2}.\end{aligned}$$

When $t = 4$, we find

$$\begin{aligned}\left.\frac{dz}{dt}\right|_4 &= \frac{2(g(4)+1)f'(4) - 2f(4)g'(4)}{(g(4)+1)^2} \\ &= \frac{2(4+1)(-1) - 2(-1)(-4))}{(4+1)^2} \\ &= \frac{-10-8}{25} = \frac{-18}{25}.\end{aligned}$$

We now consider examples of using related rates of change to find the instantaneous rate of change of a quantity that depends on other related variables. In each example, we will first recognize how related dependent variables are algebraically combined. Then we can use rules of differentiation to identify

a new equation that relates their rates of change. This equation, in turn, allows us to solve for the unknown rate.

9.2.2.1 A Physical Example of the Sum Rule

The sum rule for derivatives tells us that the derivative of a sum of two functions equals the sum of the individual derivatives. In the context of rates of change, this means that when a dependent variable is equal to the sum of two other dependent variables, then the rate of change of the new variable must equal the sum of the rates of change of the dependent variables being added.

Example 9.2.8 A tank is being filled with water two supply hoses. At a particular instant, if the first hose is pumping water at a rate of 20 gal/min and the second hose is pumping water at a rate of 30 gal/min, at what rate is volume of water in the tank changing?

Solution. We know the intuitive solution to the problem is 50 gal/min. This is actually a consequence of the sum rule of derivatives. We can think of the water in the tank as having two components: W_1, the volume of water (gal) that was pumped by hose 1, and W_2, the volume of water (gal) that was pumped by hose 2. These two variables are functions of time t (min), although we do not know any formulas for these functions (and don't need to).

The rates of water flowing from the hoses correspond to derivatives:

$$\frac{dW_1}{dt} = 20, \quad \frac{dW_2}{dt} = 30.$$

The total volume of water in the tank at a given time t is the sum $W(t) = W_1(t) + W_2(t)$. By the sum rule of derivatives,

$$\frac{dW}{dt} = \frac{dW_1}{dt} + \frac{dW_2}{dt} = 20 + 30 = 50.$$

Technically, we should have a constant added to W that represents the initial amount of water in the tank and didn't come from either hose. Because the derivative of a constant is zero, this will not change the result. □

9.2.2.2 A Physical Example of the Product Rule

The sum rule for derivatives feels very intuitive. If a quantity is the sum of parts, then the total rate of change for the quantity is the sum of the rates of change for each of the parts. The product rule is less intuitive because we don't get to multiply rates of change when a quantity is a product. To illustrate this example, we focus on a geometric example on the area of a rectangle when the lengths of the sides are changing.

Example 9.2.9 A city is in the shape of a rectangle with sides aligned with North-South and East-West lines. Suppose that the city is currently 5 miles east-to-west and 3 miles north-to-south and plans to expand to a size 8 miles east-to-west by 5 miles north-to-south over the next 10 years. What is the average rate of change of the total area in the city over the 10 years? If the borders were to move at a constant rate over those 10 years, what is the instantaneous rate of change of the total area of the city at the beginning and at the end of the 10 years?

Solution. The average rate of change of total area is calculated according to the usual formula. It does not follow the differentiation rules, which are about instantaneous rates of change. We let A represent the area of the city and t

the time in years from now. The city currently has a total area of

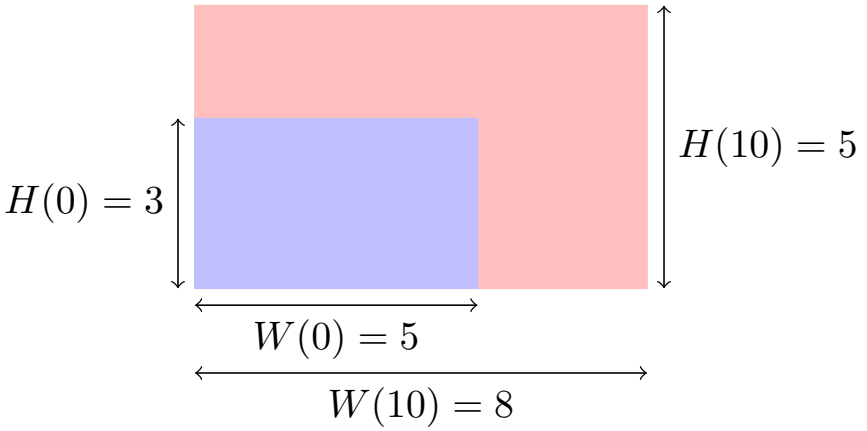
$$A(0) = 5 \times 3 = 15 \text{ mi}^2.$$

After 10 years, the city will have a total area of

$$A(10) = 8 \times 5 = 40 \text{ mi}^2.$$

The change in area is $\Delta A = A(10) - A(0) = 25 \text{ mi}^2$ and the change in time is $\Delta t = 10$ years. Consequently, the average rate of change of area is

$$\left.\frac{\Delta A}{\Delta t}\right|_{0,10} = \frac{25}{10} = 2.5 \text{ mi}^2/\text{yr}.$$



To connect our intuition with functions and to prepare for the next calculations, let us introduce variables in addition to time t and total area A. The state of the city can be characterized more precisely with two more variables: the distance east-to-west, which we'll call the width W (mi), and the distance north-to-south, which we'll call the height H (mi). We think of W, H and A as being dependent variables as they are each a function of time t. They are related variables because the area always equals the product of W and H:

$$A(t) = W(t) \cdot H(t).$$

To find the instantaneous rates of change, we need to know how fast the width and height measurements are changing in time. Because the problem stated that these changed at a constant rate, we can use the average rates of change to compute the instantaneous rates:

$$\frac{dW}{dt} = \left.\frac{\Delta W}{\Delta t}\right|_{[0,10]} = \frac{W(10) - W(0)}{10 - 0} = \frac{8 - 5}{10} = 0.3,$$
$$\frac{dH}{dt} = \left.\frac{\Delta H}{\Delta t}\right|_{[0,10]} = \frac{H(10) - H(0)}{10 - 0} = \frac{5 - 3}{10} = 0.2.$$

Since the area A is the product of W and H, the product rule for derivatives will provide the instantaneous rate of change for area:

$$\frac{dA}{dt} = \frac{d}{dt}[W \cdot H] = \frac{dW}{dt} \cdot H + W \cdot \frac{dH}{dt}.$$

This equation is the related rates equation.

When $t = 0$ we have $W(0) = 5$ and $H(0) = 3$ so that

$$\left.\frac{dA}{dt}\right|_0 = \left.\frac{dW}{dt}\right|_0 \cdot H(0) + W(0) \cdot \left.\frac{dH}{dt}\right|_0$$

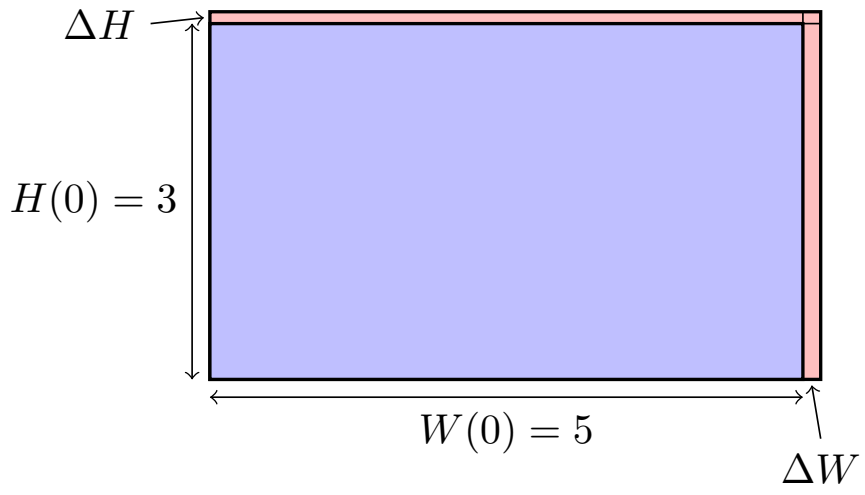
$$= 0.3(3) + 5(0.2) = 1.9.$$

That is, at the beginning, the city is expanding at a rate of $1.9\ \text{mi}^2/\text{yr}$. After 10 years, $t = 10$, we have $W(10) = 8$ and $H(10) = 5$ so that

$$\left.\frac{dA}{dt}\right|_{10} = \left.\frac{dW}{dt}\right|_{10} \cdot H(10) + W(10) \cdot \left.\frac{dH}{dt}\right|_{10}$$
$$= 0.3(5) + 8(0.2) = 3.1.$$

At the end of the 10 years, the city is expanding at a rate of $3.1\ \text{mi}^2/\text{yr}$.

The picture of expanding area helps provide some intuition for why the product rule is the appropriate technique. If we consider the city after 6 months ($t = 0.5$), both the width and the height have changed by a small amount, as shown in the figure below. The total change in area has two primary contributions, corresponding to long, skinny rectangles with areas $W(0) \cdot \Delta H$ and $\Delta W \cdot H(0)$, and a very small rectangle with area $\Delta W \cdot \Delta H$. The product rule corresponds to the rate of change coming from the two primary contributions while the small rectangle leads to a term that has a limit of zero in the calculation of the derivative.



□

9.2.2.3 A Physical Example of the Quotient Rule

Quotients often appear when working with densities, concentrations, or other ratios.

Example 9.2.10 A salt-water solution is being formulated. At a particular instant, the solution consists of 10 L of water with 5 kg of salt. At that instant, water is being added at a rate of 0.5 L/s while salt is being added at a rate of 0.2 kg/s. What is the instantaneous rate of change of the concentration?

Solution. We start by identifying the variables that define the state of our system. The variables include the time t, measured in seconds (s), the total volume of water V, measured in liters (L), the total amount of salt in the water S, measured in kilograms (kg), and the concentration of salt water C, measured in kilograms per liter (kg/L). The variables V, S and C are functions of time t with an equation relating them by

$$C(t) = \frac{S(t)}{V(t)} \quad \Leftrightarrow \quad C = \frac{S}{V}.$$

The instantaneous rate of change is computed using the quotient rule for derivatives, giving us a related rates equation

$$\frac{dC}{dt} = \frac{V\frac{dS}{dt} - S\frac{dV}{dt}}{V^2}.$$

The values at the instant in question are given by

$$V = 10, \quad \frac{dV}{dt} = 0.5,$$
$$S = 5, \quad \frac{dS}{dt} = 0.2.$$

Using these values in the quotient rule for derivatives, we have

$$\frac{dC}{dt} = \frac{10(0.2) - 5(0.5)}{10^2} = \frac{2 - 2.5}{100} = -0.005.$$

That is, the concentration is changing at a rate of -0.005 kg salt per liter water per second. Alternatively, we could say that the concentration is decreasing at a rate of 0.005 kg/L/s. □

9.2.3 Summary

1. When differentiating a formula, we must identify the *last* operation that acts on an expression. Last is determined according to the order of operations.
2. The linear rules of differentiation include the constant multiple and sum rules. These feel more intuitive because differentiation occurs in place.
3. The nonlinear rules of differentiation include the product and quotient rules. The derivative of a product consists of the sum of two terms, not just the product of the derivatives. THe derivative of a quotient involves subtraction of two terms and a denominator that is squared.
4. Practice applying the rules until they are mastered. For example, try Derivative Practice on Algebraic Formulas.
5. Rules of differentiation apply to any instantaneous rate of change, whether expressed as a function or not. Related rates are calculated by first expressing an equation that defines the relation between quantities. The rules of differentiation produce an equation relating the rates of those quantities.

9.2.4 Exercises

Use the values of $f(x)$ and $g(x)$ and their derivatives from the following table to calculate the indicated derivative.

x	0	1	2	3
$f(x)$	2	5	-3	4
$g(x)$	-2	1	6	5
$f'(x)$	3	1	2	-4
$g'(x)$	4	7	-1	2

1. If $h(x) = 3f(x) + 2g(x)$, find $h'(0)$.

2. If $H(x) = x^2 f(x) + 4$, find $H'(3)$.

3. If $p(x) = \dfrac{3}{g(x)}$, find $p'(1)$.

4. If $Q(x) = \dfrac{f(x)}{2g(x)}$, find $Q'(2)$.

Related Rates As you solve these related rates problems, practice clearly identifying the dependent variables and the independent variable. State the equation that relates these variables. Use the rules of differentiation to create an equation that relates the rates of change.

5. A candle is lit at both ends. One end is burning at a rate of 1 cm/ hour. The other end is burning at a rate of 2 cm/hour. What is the rate of change of the length of the candle?

6. A population of birds on an island changes due to births, deaths, and migration of individuals. If the population has births occuring at a rate of 800 per year, deaths occuring at a rate of 720 per year, immigration of 100 per year and emigration of 150 per year, what is the overall rate of change of the population?

7. A movie company's income is based on two money streams: direct online rental and DVD sales. Suppose that the company receives \$2.50 for each online rental and \$6.00 for each sold DVD. If the company rents movies at a rate of 2000 movies per month and sells DVDs at a rate of 1500 DVDs per month, what is the rate of income?

8. The concentration of an antibiotic drug in the bloodstream is affected by the rate of administration and by the rate of metabolism. Suppose that an individual has 5 liters of blood and the drug is being administered by injection at a constant rate of 0.3 grams per hour. In addition, the body removes the drug by metabolism at an instantaneous rate (grams per hour) that is proportional to the total amount (mass in grams) of the drug in the body at that instant, where the proportionality constant is 0.8.

Let M represent the total mass (grams) of the drug in the body, and let C represent the concentration (grams per liter) of the drug. Let t measure the time (hours) since the treatment began.

- State an equation relating $\frac{dM}{dt}$ and M based on the description of injection and metabolism.
- State an equation relating M and C. What is the corresponding related rates equation?
- What is $\frac{dC}{dt}$ at a particular moment when $C = 0.04$ grams per liter?
- An equilibrium has been reached if $\frac{dC}{dt} = 0$. What is the equilibrium concentration? That is, find C so that $\frac{dC}{dt} = 0$.

9. A city has a population of 40,000 and a total debt of \$54 million. If the city's population is growing at a rate of 1,000 per year and is borrowing additional money at a rate of \$2 million per year, what is the rate of change of the per capita debt (total debt divided by population)?

10. A company has 300 employees that earn an average annual salary of \$40,000 per employee. If the company's workforce is growing at a rate of 20 employees per year and the average annual salary is increases at

a rate of \$500 per employee per year. What is the rate of change of total salary costs for the company?

11. Potential energy P of an object raised above the ground is defined as the product of the mass m of the object times the height h above the ground times the gravitational constant $g = 9.8\ \frac{\text{m}}{\text{s}^2}$. A bag of sand weighing 4 kilograms is at a height of 2 meters. If the bag is losing 0.05 kg of sand per second and is being lifted at a rate of 2 cm per second, what is the rate of change of the potential energy? Note: 1 joule of energy is the same as $1\,\text{kg} \cdot \text{m} \cdot \text{s}^2$.

12. A city has a population of 40,000 and a total debt of \$54 million. If the city's population is growing at a rate of 1,000 per year and is borrowing additional money at a rate of \$2 million per year, what is the rate of change of the per capita debt (total debt divided by population)?

9.3 The Chain Rule

The derivative rules we have learned this far focus on the arithmetic operations that combine expressions into more complex operations—addition, subtraction, multiplication, and division. Another operation that combines expressions is composition. A function f represents a map from an independent variable to a dependent variable, say $f : x \mapsto y$. Composition occurs when the output from another function becomes the input. The chain rule provides the differentiation rule for composition.

In this section, we develop the chain rule. We begin by reviewing the idea of a chain of variables and the relation this has to function composition. The chain rule is based on the derivative being the limiting rate of change. By considering how an increment of change in the independent variable propogates through the chain, we will see that the rates of change at each step in the chain are multiplied together. After a few examples of using the chain rule for formulas, we then explore a few examples of the chain rule for related rates.

9.3.1 Review: Rate of Change and Composition

We start by reminding ourselves that a rate of change is a ratio of changes for two variables. If y is a function of x, say $x \mapsto y = f(x)$, then the rate of change $\left.\frac{dy}{dx}\right|_a = f'(a)$ is the rate of change of y with respect to x at the value $x = a$. This measures the instantaneous ratio of changes in y from $f(a)$ to changes in x from a. At any value x close to a, this means that

$$y - f(a) \approx \left.\frac{dy}{dx}\right|_a \cdot (x - a).$$

Changes in the value of y are approximately proportional to changes in x from a and the derivative $f'(a)$ is the proportionality constant.

Second, we remind ourselves that compositions correspond to chains of dependent variables. Suppose that u is a function of x, say $u = g(x)$, and y is subsequently a function of u, say $y = f(u)$. We would write this chain as

$$\begin{cases} u = g(x) \\ y = f(u) \end{cases}.$$

Using substitution, we could also just write that y is a function of x using composition.

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

Now, let us consider a particular value for x and ask how would we determine the rate of change of y with respect to x when it is defined with such a composition? A change in x from a, $\Delta x = x - a$, would lead to a change in u from $g(a)$ using the rate of change

$$\Delta u = u - g(a) \approx \left.\frac{du}{dx}\right|_a \cdot (x - a) = g'(a) \cdot \Delta x.$$

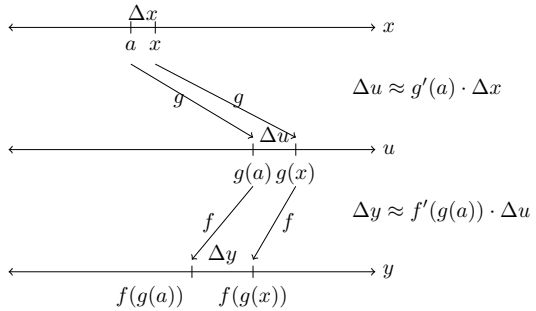
In a similar way, a change in u from its starting value $g(a)$ would lead to a change in y from $f(g(a))$ using the rate of change

$$\Delta y = y - f(g(a)) \approx \left.\frac{dy}{du}\right|_{g(a)} \cdot (u - g(a)) = f'(g(a)) \cdot \Delta u.$$

Putting these two results of the chain together, we find that

$$\Delta y \approx \left.\frac{dy}{du}\right|_{g(a)} \cdot \left.\frac{du}{dx}\right|_{a} \cdot \Delta x = f'(g(a)) \cdot g'(a) \cdot \Delta x.$$

Graphically, this is illustrated in the figure below. The inputs and outputs of the functions for g and f are illustrated as maps between number lines. The input $x = a$ to the function $g : x \mapsto u$ is mapped to the output $u = g(a)$. A nearby input x is mapped to an output $g(x)$ that is not too far from $g(a)$. The differences are the values $\Delta x = x - a$ and $\Delta u = g(x) - g(a)$. In composition, the outputs $g(a)$ and $g(x)$ act as inputs to f.



The derivative provides an approximate ratio in the changes of output values to the changes of input values. The smaller the input, the closer the approximation. This is why the derivative must be defined as a limit of the average rate of change. When functions are in composition, each function effectively amplifies the difference in output by the factor of the derivative. So the overall change in the output is a result of the product of the derivatives.

9.3.2 The Chain Rule for Derivatives

The chain rule formalizes the ideas in the previous paragraphs. It states that the derivative of a composition $f(g(x))$ has a derivative given by

$$\frac{d}{dx}[f(g(x))] = f'(g(x)) \cdot g'(x).$$

Pay close attention to the inputs of f' and g'. Compare those values to what we had to do in the previous paragraphs. The inputs are different because the functions $f : u \mapsto y$ and $g : x \mapsto u$ have different inputs in the composition.

Theorem 9.3.1 *If we have an explicit chain representation,*

$$\begin{cases} u = g(x) \\ y = f(u) \end{cases},$$

then the chain rule can be rewritten:

$$\begin{aligned} \frac{dy}{dx} &= \frac{d}{dx}[f(g(x))] \\ &= f'(g(x)) \cdot g'(x) \\ &= f'(u) \cdot \frac{du}{dx} \\ &= \left.\frac{dy}{du}\right|_{u=g(x)} \cdot \left.\frac{du}{dx}\right|_{x}. \end{aligned}$$

The chain rule is often abbreviated as

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}.$$

Notice that this form almost looks like algebra would cancel the symbol du on the right to give the formula $\frac{dy}{dx}$ on the left.

Example 9.3.2 Find the derivative of $f(x) = (2x+1)^2$ using the chain rule and compare the result to what you get if you expand $f(x)$ before differentiation.

Solution. To use the chain rule, we must identify the chain or composition that is involved. The last operation in this formula is the act of squaring. What do we square? This will be the way that we identify $u = 2x+1$. Then the final output is $y = u^2$. We can find the derivatives of each step in the chain:

$$\begin{cases} u = 2x+1 \\ y = u^2 \end{cases} \quad \Rightarrow \quad \begin{cases} \frac{du}{dx} = 2 \\ \frac{dy}{du} = 2u \end{cases}.$$

Consequently, we have

$$\frac{dy}{dx} = \left.\frac{dy}{du}\right|_{u=2x+1} \cdot \frac{du}{dx}.$$

The notation $u = 2x+1$ is simply a reminder that when writing the derivative $\frac{dy}{du} = 2u$ we will ultimately replace $u = 2x+1$.

$$f'(x) = \frac{dy}{dx} = (2u) \cdot (2) = 2(2x+1) \cdot 2 = 4(2x+1)$$

The other approach is to expand $f(x)$ to a form that is easier to differentiate.

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

This is a simple polynomial form that has a simple derivative:

$$f'(x) = 8x + 4.$$

We can see that this is actually the same as our earlier derivative if we factor out the common factor of 4. □

We could avoid the chain rule in the previous example because expanding the square of our expression could be calculated fairly simply. When this is not possible, the chain rule must be used.

Example 9.3.3 Find the derivative of $f(x) = 3(x^2 + 3x)^7$.

Solution. Our function has an intermediate formula $u = x^2 + 3x$ that is then raised to the 7th power and multiplied by 3. That is, if $y = f(x)$ then $y = 3u^7$. We would write this as a chain, along with their derivatives:

$$\begin{cases} u = x^2 + 3x \\ y = 3u^7 \end{cases} \quad \Rightarrow \quad \begin{cases} \frac{du}{dx} = 2x + 3 \\ \frac{dy}{du} = 21u^6 \end{cases}.$$

The chain rule implies

$$\begin{aligned} f'(x) = \frac{dy}{dx} &= \frac{dy}{du}\frac{du}{dx} \\ &= 21u^6 \cdot (2x+3) \\ &= 21(x^2+3x)^6(2x+3). \end{aligned}$$

Note that we had to substitute the formula for u to find our final result.

In the language of function composition, we could instead do this by writing $f(x)$ as a composition $f(x) = g(h(x))$:

$$\begin{aligned} h : x \mapsto u = x^2 + 3x \qquad & h(x) = x^2 + 3x \qquad h'(x) = 2x + 3 \\ g : u \mapsto y = 3u^7 \qquad & g(u) = 3u^7 \qquad g'(u) = 21u^6 \end{aligned}$$

The chain rule would be written:

$$\begin{aligned} f'(x) &= g'(h(x)) \cdot h'(x) \\ &= g'(x^2 + 3x) \cdot (2x + 3) \\ &= 21(x^2 + 3x)^6(2x + 3) \end{aligned}$$

□

Negative and rational powers are much simpler with the chain rule. Using negative powers in composition often helps us avoid needing the quotient rule.

Example 9.3.4 Find $f''(x)$ where $f(x) = \frac{3}{x^2+1}$.

Solution. The first derivative can be found using the quotient or reciprocal rule.

$$\begin{aligned} f'(x) &= 3\frac{d}{dx}\Big[\frac{1}{x^2+1}\Big] \\ &= 3 \cdot \frac{-2x}{(x^2+1)^2} \\ &= \frac{-6x}{(x^2+1)^2} \end{aligned}$$

We could also have done this using a chain rule. The relevant chain and associated derivatives are given:

$$\begin{cases} y = 3u^{-1} \\ u = x^2 + 1 \end{cases} \qquad \Rightarrow \qquad \begin{cases} \frac{dy}{du} = -3u^{-2} \\ \frac{du}{dx} = 2x \end{cases}$$

Consequently, we know $f'(x) = \frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$ and

$$f'(x) = -3(x^2 + 1)^{-2} \cdot (2x) = -6x(x^2 + 1)^{-2}.$$

To calculate the second derivative, we differentiate $f'(x)$. We could use either the quotient rule or the product rule with negative powers. In the first case, we find

$$\begin{aligned} f''(x) &= \frac{d}{dx}\Big[\frac{-6x}{(x^2+1)^2}\Big] \\ &= \frac{(x^2+1)^2\frac{d}{dx}[-6x] - (-6x)\frac{d}{dx}[(x^2+1)^2]}{(x^2+1)^4} \\ &= \frac{(x^2+1)^2(-6) + (6x) \cdot 2(x^2+1)(2x)}{(x^2+1)^4}, \end{aligned}$$

where we have used the chain rule on u^2 with $u = x^2 + 1$ to obtain

$$\frac{d}{dx}[(x^2 + 1)^2] = 2(x^2 + 1)(2x).$$

Notice that the numerator of $f''(x)$ has $x^2 + 1$ as a common factor, which cancels with one of the corresponding factors in the denominator. A simplified

version of $f''(x)$ is therefore given by

$$\begin{aligned} f''(x) &= \frac{-6(x^2+1)+(6x)\cdot 2(2x)}{(x^2+1)^3} \\ &= \frac{-6x^2-6+24x^2}{(x^2+1)^3} \\ &= \frac{18x^2-6}{(x^2+1)^3}. \end{aligned}$$

The other approach to finding the second derivative is to start with the product representation of $f'(x)$ and differentiate using the product rule. In order to differentiation $(x^2+1)^{-2}$, we use the chain rule on u^{-2} with $u = x^2+1$:

$$\frac{d}{dx}\Big[u^{-2}\Big] = -2u^{-3}\cdot\frac{du}{dx} = -2(x^2+1)^{-3}\cdot(2x).$$

This will give us

$$\begin{aligned} f''(x) &= \frac{d}{dx}\Big[(-6x)\cdot(x^2+1)^{-2}\Big] \\ &= \frac{d}{dx}\Big[-6x]\cdot(x^2+1)^{-2} + -6x\frac{d}{dx}\Big[(x^2+1)^{-2}\Big] \\ &= -6\cdot(x^2+1)^{-2} + -6x\cdot -2(x^2+1)^{-3}(2x) \\ &= -6\cdot(x^2+1)^{-2} + 24x^2(x^2+1)^{-3} \end{aligned}$$

Remembering that negative exponents correspond to powers in the denominator, we can see this formula requires a common denominator $(x^2+1)^3$ to simplify

$$\begin{aligned} f''(x) &= \frac{-6}{(x^2+1)^2} + \frac{24x^2}{(x^2+1)^3} \\ &= \frac{-6(x^2+1)+24x^2}{(x^2+1)^3} \\ &= \frac{18x^2-6}{(x^2+1)^3} \end{aligned}$$

We found the same answer both ways. Derivative rules are self-consistent. □

There may be times where the chain rule must be used more than once. Any time the last operation on an expression is a function acting on an expression, such as a power as opposed to arithmetic operations like sums or products joining two expressions, we need to use the chain rule.

Example 9.3.5 If $f(x) = (\sqrt{3x}+2)^4$, compute $f'(x)$.

Solution. The last operation in $f(x)$ is raising an expression to the power 4. The derivative will require a chain rule. The first step is to differentiate this last operation.

$$\begin{aligned} f'(x) &= \frac{d}{dx}\Big[(\sqrt{3x}+2)^4\Big] \\ &\underset{u=\sqrt{3x}+2}{=} \frac{d}{du}\Big[u^4\Big]\frac{d}{dx}\Big[\sqrt{3x}+2\Big] \\ &\underset{u=\sqrt{3x}+2}{=} 4u^3\cdot\frac{d}{dx}\Big[(3x)^{1/2}+2\Big] \end{aligned}$$

$$= 4(\sqrt{3x}+2)^3 \cdot \frac{d}{dx}\Big[(3x)^{1/2}+2\Big]$$

We need to continue by finding the derivative of the inner expression $u = \sqrt{3x}+2$. This is a sum, and the second term in a sum is a constant. The derivative of a constant is zero. We need to compute the derivative of $(3x)^{1/2}$, which is another composition. The expression $3x$ is raised to a power $\frac{1}{2}$. We need the chain rule one more time.

$$\begin{aligned}\frac{d}{dx}\Big[(3x)^{1/2}+2\Big] &= \frac{d}{dx}\Big[(3x)^{1/2}\Big]+0\\ &\underset{u=3x}{=} \frac{d}{du}\Big[u^{1/2}\Big]\frac{d}{dx}\Big[3x\Big]\\ &\underset{u=3x}{=} \frac{1}{2}u^{-1/2}\cdot 3\\ &= \frac{3}{2}(3x)^{-1/2} = \frac{3}{2\sqrt{3x}}\end{aligned}$$

Substituting this into our original formula for $f'(x)$, we find

$$\begin{aligned}f'(x) &= 4(\sqrt{3x}+2)^3 \cdot \frac{3}{2\sqrt{3x}}\\ &= \frac{6}{\sqrt{3x}}(\sqrt{3x}+2)^3.\end{aligned}$$

□

9.3.3 Related Rates and the Chain Rule

Derivative rules are fundamentally about relationships between instantaneous rates. The chain rule is no exception. The biggest difference in the rates that are related by the chain rule and other related rates problems is that the chain rule involves different independent variables for different steps in the chain.

Example 9.3.6 Consider a temperature dependent chemical reaction. At 20 degrees Celsius, the reaction generates a product at a rate of 30 grams per minute. For small changes in temperature, the reaction can generate an addition 5 grams per minute per degree increase in temperature. If the temperature is cooling at a rate of 0.05 degrees per minute, what is happening to the reaction?

Conceptually, we recognize some variables in this problem: the temperature T (in degrees Celsius), the time t (in minutes), and the reaction rate R (in grams per minute). Because temperature is changing in time, we know there is a map $t \mapsto T$. Similarly, we know that the reaction rate depends on temperature, there is another map $T \mapsto R$. In combination, we have a chain $t \mapsto T \mapsto R$.

We identify the values at the instant t in question. We know $T = 20$ and $\frac{dT}{dt} = -0.05$. (Why?) Similarly, we know $R = 30$ and $\frac{dR}{dT} = 5$. The chain rule tells us the rate of change of the final variable in the chain with respect to the original independent variable in the chain:

$$\frac{dR}{dt} = \frac{dR}{dT}\cdot\frac{dT}{dt} = 5\cdot -0.05 = -0.25.$$

That is, the reaction rate is *decreasing* at a rate of 0.25 grams per minute per minute. (R has units of grams per minute so $\frac{dR}{dt}$ has units of grams per minute per minute.) □

Example 9.3.7 As an ice cube melts, it maintains the shape of a cube. At one particular instant, each side of the cube is 30 mm and the volume of the cube is melting at a rate of $500\ \frac{\text{mm}^3}{\text{s}}$. What is the rate of change of the length of the sides at that instant?

Solution. Start by identifying the variables in the problem. The state of the ice cube is characterized by the time, the length of the sides, and the total volume. Let t be the time (in seconds), s the length of a side (in millimeters), and V the volume (in cubic millimeters).

Next identify the functions defining relations between the variables. We know that the length and volume are both functions of time, so we know there are maps $t \mapsto s$ and $t \mapsto V$. This is not a chain because t is the independent variable for both maps. We also know that the volume is a function of the length of a side, $s \mapsto V = s^3$. From this, we can identify a chain, $t \mapsto s \mapsto V$.

We finish by creating an equation relating our rates. Because our variables are related by a chain, the chain rule establishes this relationship:

$$\frac{dV}{dt} = \frac{dV}{ds}\frac{ds}{dt}.$$

The problem gives us $\frac{dV}{dt} = -500\ \frac{\text{mm}^3}{\text{s}}$. The equation $V = s^3$ is an explicit formula from which we can compute a derivative

$$\frac{dV}{ds} = 3s^2.$$

At the instant in question, $s = 30$ mm so that $\frac{dV}{ds} = 3(30)^2 = 2700\ \frac{\text{mm}^3}{\text{mm}}$. The related rates equation involved three rates, two of which we now know. Solving for $\frac{ds}{dt}$, we find

$$\frac{ds}{dt} = \frac{\frac{dV}{dt}}{\frac{dV}{ds}} = \frac{-500}{2700} = -\frac{5}{27}.$$

That is, the lengths of the sides are decreasing at a rate of $-\frac{5}{27}\ \frac{\text{mm}}{\text{s}}$. □

In some examples, there are multiple equations relating the variables. In that case, there will also be multiple equations relating their rates.

Example 9.3.8 Many water coolers have cups in the shape of a circular cone. The volume V of a cone can be calculated in terms of the radius r of the circular base and the height of the cone h by

$$V = \frac{1}{3}\pi r^2 h.$$

As water fills the cup, the volume of water creates a smaller cone than the cup but one with similar dimensions.

Suppose a cup has a height of 12 cm and a radius at the top of 5 cm. Water is filling the cup at a rate of $80\ \frac{\text{cm}^3}{\text{s}}$. When the cup is filled to a depth of 6 cm, how fast is the depth changing?

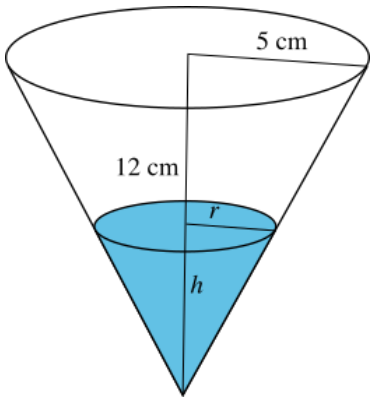


Figure 9.3.9 Illustration of partially filled cup of water in the shape of a cone.

Solution. We will work through two different approaches to solving this problem. The first method will be to consider two equations that relate our variables and create two equations for the related rates. The second method will use the two equations relating the variables to create a single function to create a related rates equation.

There are three basic dependent variables: the height of water in the cup, the radius of the circle at the top of the water level, and the volume of water in the cup. All of these change with respect to the independent variable of time. Let t measure time in seconds, let h measure the height of water, let r measure the radius at the top of the water level, and let V measure the volume of water in the cup. Interpreting the given information, we should note the values of variables at the instant in question. The units of how fast water is filling is a volume per unit time, which we interpret as saying $\frac{dV}{dt} = 80$. The depth of the water informs us that $h = 6$. The question asks us to determine $\frac{dh}{dt}$.

The volume of water is related to the radius and height by the equation

$$V = \frac{1}{3}\pi r^2 h.$$

In addition, we know that the radius and height must be similar dimensions to the radius and height of the cup itself. This means that the ratios of corresponding sides must be equal, giving a second equation

$$\frac{r}{5} = \frac{h}{12}.$$

If we solve for r, we find $r = \frac{5}{12}h$.

From the equations relating the dependent variables, we can differentiate to develop equations relating their rates of change. The volume is defined as a constant multiple of $\frac{1}{3}\pi$ with the product r^2h, and the derivative of r^2 requires the chain rule:

$$\begin{aligned}
\frac{dV}{dt} &= \frac{1}{3}\pi\frac{d}{dt}\Big[r^2h\Big] \\
&= \frac{1}{3}\pi\Big(\frac{d}{dt}\Big[r^2\Big]\cdot h + r^2\cdot\frac{dh}{dt}\Big) \\
&= \frac{1}{3}\pi(2r\frac{dr}{dt})\cdot h + \frac{1}{3}\pi r^2\cdot\frac{dh}{dt} \\
&= \frac{2}{3}\pi rh\frac{dr}{dt} + \frac{1}{3}\pi r^2\frac{dh}{dt}.
\end{aligned}$$

We can also differentiate the equation defining r to relate the rates for r and h:

$$\frac{dr}{dt} = \frac{d}{dt}\Big[\frac{5}{12}h\Big] = \frac{5}{12}\frac{dh}{dt}.$$

With these equations and the data, we can solve for $\frac{dh}{dt}$. Our related rates equation involves the variable r, for which we do not have a value. We can use the similar dimensions equation to solve for r,

$$r = \frac{5}{12}h = \frac{5}{12}(6) = \frac{5}{2}.$$

Substituting the values of variables and rates into the related rates equation for $\frac{dV}{dt}$, we find

$$80 = \frac{2}{3}\pi(\frac{5}{2})(6)\frac{dr}{dt} + \frac{1}{3}\pi(\frac{5}{2})^2\frac{dh}{dt}.$$

As this equation has both rates $\frac{dr}{dt}$ and $\frac{dh}{dt}$, we substitute into the equation our relation $\frac{dr}{dt} = \frac{5}{12}\frac{dh}{dt}$:

$$\begin{aligned} 80 = \frac{2}{3}\pi(\frac{5}{2})(6)(\frac{5}{12})\frac{dh}{dt} + \frac{1}{3}\pi(\frac{5}{2})^2\frac{dh}{dt} \\ 80 = \frac{25}{6}\pi\frac{dh}{dt} + \frac{25}{12}\pi\frac{dh}{dt} \\ 80 = \frac{75}{12}\pi\frac{dh}{dt} \\ \frac{80(12)}{75\pi} = \frac{dh}{dt} \\ \frac{dh}{dt} = \frac{64}{5\pi} \approx 4.074. \end{aligned}$$

Consequently, we conclude the height of water is rising at a rate just higher than 4 $\frac{\text{cm}}{\text{s}}$.

The second method uses substitution earlier in the process. Instead of substituting the rate of change from related rates, this approach seeks to write an equation so that V is only a function of h. (We choose h because it is that variable's rate of change that is desired.) Because $r = \frac{5}{12}h$, we can create a single equation relating V and h:

$$\begin{aligned} V &= \frac{1}{3}\pi r^2 h \\ &= \frac{1}{3}\pi(\frac{5}{12}h)^2(h) \\ &= \frac{1}{3}\pi(\frac{25}{144})h^3 \\ &= \frac{25}{432}\pi h^3. \end{aligned}$$

Once we have the equation relating volume and height of the water, we can differentiate to find a single related rates equation using the constant multiple rule and the chain rule for the power h^3:

$$\begin{aligned} \frac{dV}{dt} &= \frac{25}{432}\pi\frac{d}{dt}\Big[h^3\Big] \\ &= \frac{25}{432}\pi 3h^2\frac{dh}{dt} \\ &= \frac{25}{144}\pi h^2\frac{dh}{dt}. \end{aligned}$$

At this point, we can substitute our known values and solve for $\frac{dh}{dt}$:

$$\begin{aligned}\frac{dV}{dt} &= \frac{25}{144}\pi h^2 \frac{dh}{dt}\\ 80 &= \frac{25}{144}\pi (6)^2 \frac{dh}{dt}\\ 80 &= \frac{25}{4}\pi \frac{dh}{dt}\\ \frac{80(4)}{25\pi} &= \frac{dh}{dt}\\ \frac{dh}{dt} &= \frac{64}{5\pi}.\end{aligned}$$

□

9.3.4 Summary

- A composition or chain occurs when the output of one function acts as the input to another function.
- The derivative measures the limiting ratio of changes in the output to the input for small changes in the input. Consequently, in a composition or chain of functions, the overall rate of change is the product of the rates of change for each step.
- The chain rule states that

  $$\frac{d}{dx}\big[f\big(g(x)\big)\big] = f'(g(x)) \cdot g'(x).$$

  Represented as a chain $u = g(x)$ and $y = f(u)$ so that $y = f(g(x))$, the chain rule would be written

  $$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}.$$

  This is the derivative of the outer operation times the derivative of the inner expression.

9.3.5 Exercises

Use the given rates to find the unknown rate.

1. Given $\frac{dy}{du} = 4$ and $\frac{du}{dx} = -3$, find $\frac{dy}{dx}$.
Hint: Imagine a chain $x \mapsto u \mapsto y$ and apply the chain rule.

2. Given $\frac{dF}{dP} = 1.2$ and $\frac{dP}{dt} = 40$, find $\frac{dF}{dt}$.
Hint: $t \mapsto P \mapsto F$.

3. Given $\frac{dR}{dt} = 50000$ and $\frac{dp}{dt} = -2$, find $\frac{dR}{dp}$.
Hint: $t \mapsto p \mapsto R$.

4. Given $\frac{dR}{dt} = 50000$ and $\frac{dp}{dt} = -2$, find $\frac{dR}{dp}$.
Hint: $t \mapsto p \mapsto R$.

Compute the derivatives.

5. $\frac{d}{dx}[(3x+2)^3]$

6. $\frac{d}{dx}[(x^2+1)^5]$

7. $\frac{d}{dx}[(2x-5)^{-2}]$

8. $\frac{d}{dx}[\sqrt{4x+1}]$

9. $\frac{d}{dx}[\frac{3}{\sqrt{x^2+4}}]$

10. $\frac{d}{dx}[(x^3+2x)^{-2/3}]$

11. $\frac{d}{dx}[x^4(x^2+1)^3]$

12. $\frac{d}{dx}[x\sqrt{2x+1}]$

13. $\frac{d}{dx}[(3x+1)^4(2x-5)^3]$

14. $\frac{d}{dx}[\frac{3x}{(2x+1)^2}]$

15. For $f(x)=x^3(2x+1)^5$, find $f''(x)$.

16. For $g(x)=\frac{3}{x^2+1}$, find $g''(x)$.

17. For $h(x)=\sqrt{x^3-1}$, find $h''(x)$.

Use the values of $f(x)$ and $g(x)$ and their derivatives from the following table to calculate the indicated values.

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

18. For $h(x)=f(g(x))$, find $h(2)$ and $h'(2)$.

19. For $h(x)=g(f(x))$, find $h(2)$ and $h'(2)$.

20. For $h(x)=g(2x-3)$, find $h(3)$ and $h'(3)$.

21. For $h(x)=f(x^2)$, find $h(2)$ and $h'(2)$.

22. For $h(x)=f^2(x)=(f(x))^2$, find $h(1)$ and $h'(1)$.

23. For $h(x)=f(2g(x))$, find $h(0)$ and $h'(0)$.

Related Rates

24. A ripple in a pond spreads as a circle whose radius grows at a speed of 30 $\frac{\text{cm}}{\text{s}}$. At what rate is the area enclosed by the ripple increasing?

25. An oil spill in the ocean is spreading as a circle such that the total area is increasing at a constant rate. After 10 hours, the circle has a radius of 0.1 km. What is the instantaneous rate of change of the radius at this time?

26. A bacteria colony grows on its substrate in the shape of a circle. Your colleague suggests that the colony only grows along the outer edge such that the rate of change of the area should be proportional to the circumference. Show that this predicts a constant rate of change for the radius.

27. A spherical balloon is being filled with air at a rate of 0.5 cubic meters per minute. How fast is the radius increasing when the balloon has a radius of 20 cm?

28. A spherical balloon is being filled with air at a rate of 0.5 cubic meters per minute. At what radius will the balloon have its radius growing at a rate of 1 centimeter per second?

29. A pile of sand takes the form of a circular cone. As the sand falls, the pile always maintains the same slope so that the height and diameter have the same proportions. When the pile is 2 meters high, the diameter is 4 meters. If the sand pile at that instant is getting taller at a rate of 0.2 meters per minute, at what rate (cubic meters per minute) is sand being added to the pile?

9.4 The Derivative of Exponential Functions

We learned that the elementary exponential functions are of the form

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

for positive real numbers b. Because exponential functions involve powers, a common mistake students make is to use the power rule of derivatives. That rule only applies to power functions, where the independent variable is the base and the exponent is a constant. We will need a new rule for exponential functions.

In this section, we explore the derivative rule associate with exponential functions. As this is a new rule, we return to the definition of the derivative. We will learn that the derivative of an exponential function is proportional to the value of the function itself. The constant of proportionality is found using the base of the exponential.

9.4.1 Elementary Exponential Functions

The definition of the derivative allows us to develop a new differentiation rule. The key property necessary for the exponential function is a consequence of the properties of exponents,

$$\exp_b(x+y) = b^{x+y} = b^x \cdot b^y = \exp_b(x) \cdot \exp_b(y).$$

Using these properties

$$\begin{aligned}
\exp_b'(x) &= \lim_{h\to 0} \frac{\exp_b(x+h) - \exp_b(x)}{h} \\
&= \lim_{h\to 0} \frac{b^{x+h} - b^x}{h} \\
&= \lim_{h\to 0} \frac{b^x b^h - b^x}{h} \\
&= \lim_{h\to 0} \exp_b(x) \cdot \frac{b^h - 1}{h} \\
&= \exp_b(x) \cdot \lim_{h\to 0} \frac{b^h - 1}{h}.
\end{aligned}$$

This means that the derivative of b^x is just b^x times some number $L(b)$ that depends on b,

$$\frac{d}{dx}[b^x] = b^x \cdot L(b),$$

where $L(b)$ is calculated as the limit

$$L(b) = \lim_{h\to 0} \frac{b^h - 1}{h}.$$

Unfortunately, this is not a limit that we know as of yet. The following table illustrates some approximations for the value of this limit for a variety of bases.

h	$\frac{2^h-1}{h}$	$\frac{3^h-1}{h}$	$\frac{5^h-1}{h}$	$\frac{10^h-1}{h}$
$h = 0.1$	0.71773	1.16123	1.74619	2.58925
$h = 0.01$	0.69555	1.10467	1.62246	2.32930
$h = 0.001$	0.69339	1.09922	1.61073	2.30524
$h = 0.0001$	0.69317	1.09867	1.60957	2.30285
$h \to 0$	0.69315	1.09861	1.60944	2.30259

Every positive real number b has such a limit, $L(b)$. This limit corresponds to the slope of the elementary exponential function $y = b^x$ at the point $x = 0$,

$$L(b) = \exp_b'(0) = \lim_{h\to 0} \frac{b^{0+h} - b^0}{h} = \lim_{h\to 0} \frac{b^h - 1}{h}.$$

For the four bases used above, this corresponds to the following derivatives:

$$\begin{aligned}
\frac{d}{dx}[2^x] &= L(2) \cdot 2^x \approx 0.69315 \cdot 2^x \\
\frac{d}{dx}[3^x] &= L(3) \cdot 3^x \approx 1.09861 \cdot 3^x \\
\frac{d}{dx}[5^x] &= L(5) \cdot 5^x \approx 1.60944 \cdot 5^x \\
\frac{d}{dx}[10^x] &= L(10) \cdot 10^x \approx 2.30259 \cdot 10^x
\end{aligned}$$

We can see from the table that $L(2) \approx 0.69315$ and $L(3) \approx 1.09861$. This suggests that there is a particular base b between 2 and 3 such that $L(b) = 1$. Such a value does exist, using $b \approx 2.71828183$. This value has the property that the elementary exponential function and its derivative are exactly equal. The base is called the **natural base** and is given the special symbol e. Consequently,

$$\frac{d}{dx}[e^x] = e^x.$$

Definition 9.4.1 The number e is that positive value such that

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

◊

Theorem 9.4.2

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

Every exponential function is proportional to its derivative. This means that at every point on the graph $y = b^x$, the ratio of the slope to the y-value is always the same constant. The interactive graph in Figure 9.4.3 illustrates this principle. The proportionality constant, $L(b)$, has been defined by a limit. We will soon discover another way to find $L(b)$ for these other bases.

Specify static image with `@preview` attribute,
Or create and provide automatic screenshot as
`images/interactive-exponential-proportional-preview.png`
via the `mbx` script

Figure 9.4.3 A graph of $y = b^x$ that illustrates the proportionality relation between the slope and the y-value.

9.4.2 The Chain Rule with Exponentials

Because the exponential function $\exp(x) = e^x$ is defined with the natural base, $\exp'(x) = e^x$ is the same as the original function, $\exp' = \exp$. Combining

this with the chain rule, we get a generalized derivative of compositions with exponentials,

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

Example 9.4.4 Find the derivatives of the following functions.

1. $f(x) = e^{5x}$
2. $g(x) = e^{x^3}$
3. $h(x) = e^{x^2-4x}$

Solution. In each case, we will identify the formula $u(x)$ and then apply the chain rule.

1. For $f(x) = e^{5x}$, we have $u(x) = 5x$ so that $f(x) = e^u$. We will use $u'(x) = 5$. The chain rule gives:

$$\begin{aligned} f'(x) &= \frac{d}{dx}[e^{5x}] \\ &\underset{(u=5x)}{=} \frac{d}{dx}[e^u] \underset{(u=5x)}{=} e^u \frac{du}{dx} \\ &= e^{5x} \cdot 5 = 5e^{5x} \end{aligned}$$

   So $f'(x) = 5e^{5x}$.

2. The function $g(x) = e^{x^3}$ involves a composition with $u(x) = x^3$ such that $u'(x) = 3x^2$.

$$\begin{aligned} g'(x) &= \frac{d}{dx}[e^{x^3}] \\ &\underset{(u=x^3)}{=} \frac{d}{dx}[e^u] \underset{(u=x^3)}{=} e^u \frac{du}{dx} \\ &= e^{x^3} \cdot (3x^2) = 3x^2 e^{x^3} \end{aligned}$$

   Thus $g'(x) = 3x^2 e^{x^3}$.

3. The function $h(x) = e^{x^2-4x}$ involves a composition with $u(x) = x^2 - 4x$ such that $u'(x) = 2x - 4$.

$$\begin{aligned} h'(x) &= \frac{d}{dx}[e^{x^2-4x}] \\ &\underset{(u=x^2-4x)}{=} \frac{d}{dx}[e^u] \underset{(u=x^2-4x)}{=} e^u \frac{du}{dx} \\ &= e^{x^2-4x} \cdot (2x-4) = (2x-4)e^{x^2-4x} \end{aligned}$$

   Thus $h'(x) = (2x - 4)e^{x^2-4x}$.

□

9.4.3 Other Exponential Bases

Every function involving a positive base raised to an exponent can be rewritten using the natural exponential function $\exp(x) = e^x$. Recall that the exponential and the logarithm are inverse functions so that for every number $u > 0$ we

have

$$\exp(\ln(u)) = e^{\ln(u)} = u.$$

This identity and the rules for logarithms show that any power u^w with $u > 0$ can be rewritten as

$$u^w = \exp(\ln(u^w)) = e^{\ln(u^w)} = e^{w \cdot \ln(u)}.$$

Example 9.4.5 Rewrite each of the following in terms of the natural exponential function.

1. $f(x) = 2^x$
2. $g(x) = 5^x$
3. $p(x) = x^{3/4}$
4. $r(x) = x^x$

Solution.

1. $f(x) = 2^x = \exp(\ln(2^x)) = e^{\ln(2^x)} = e^{x \ln(2)}$
2. $g(x) = 5^x = \exp(\ln(5^x)) = e^{\ln(5^x)} = e^{x \ln(5)}$
3. $p(x) = x^{3/4} = \exp(\ln(x^{3/4})) = e^{\ln(x^{3/4})} = e^{\frac{3}{4} \ln(x)}$
4. $r(x) = x^x = \exp(\ln(x^x)) = e^{\ln(x^x)} = e^{x \ln(x)}$

□

Because every exponential can be rewritten in terms of the natural exponential, we have a special result for all other exponential functions,

$$b^x = e^{x \cdot \ln(b)}.$$

Using the chain rule with $u = x \cdot \ln(b)$ and $\frac{du}{dx} = \ln(b)$, we discover

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

Because we already had a formula $\frac{d}{dx}[b^x] = L(b) \cdot b^x$, we now have an exact expression for the limit $L(b)$:

$$L(b) = \lim_{h \to 0} \frac{b^h - 1}{h} = \ln(b).$$

Theorem 9.4.6

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

9.4.4 Power Function or Exponential Function

One of the challenges for a calculus novice is identifying which rule applies. It is essential that you can distinguish between an exponential function and a power function.

Recall that a power function has a constant power while an exponential function has a constant as the base. Furthermore, don't be fooled by numbers that look like other symbols. For example, x^e is a power function since the power is the constant value e (the natural exponential base):

$$\frac{d}{dx}[x^e] = ex^{e-1}.$$

Similarly, $x^{\sqrt{2}}$ is a power function because the power $\sqrt{2}$ is a number (even though it is represented as a formula, it does not have any variables):

$$\frac{d}{dx}[x^{\sqrt{2}}] = \sqrt{2}x^{\sqrt{2}-1}.$$

Furthermore, you must look at a formula and determine which operation determines the differentiation rule that is required (constant multiple, sum, product, quotient or chain). This is always based on the last operation to be applied under the rules of order of operations. When the number of steps is small, you should be able to write the derivative down directly. When the number of steps is large, you might need to use the differentiation operator to allow yourself the chance to work part of the way and indicate that there are still steps remaining.

Example 9.4.7 Find the derivative $\frac{d}{dx}[(x^2+4)^5 e^{5x^3}]$.

Solution. Start by identifying the final operation. In this problem, the function $f(x) = (x^2+4)^5 e^{5x^3}$ is a product of $(x^2+4)^5$ and e^{5x^3}. So we begin by using the product rule. If you want to emphasize this without having to write down the derivatives of the factors, then we use the differentiation operator:

$$f'(x) = \frac{d}{dx}[(x^2+4)^5] \cdot e^{5x^3} + (x^2+4)^5 \cdot \frac{d}{dx}[e^{5x^3}].$$

Notice that the differentiation operator is pointing out where we still need to find derivatives in order to complete the problem.

The first term $(x^2+4)^5$ should be recognized as a composition with a power function u^5 where $u = x^2+4$. We will use the chain rule:

$$\frac{d}{dx}[(x^2+4)^5] \underset{(u=x^2+4)}{=} \frac{d}{du}[u^5] \cdot \frac{du}{dx} = 5(x^2+4)^4 \cdot (2x).$$

The second term e^{5x^3} should be recognized as a composition with the exponential function e^u where $u = 5x^3$. Again, the chain rule guides us:

$$\frac{d}{dx}[e^{5x^3}] \underset{(u=5x^3)}{=} \frac{d}{du}[e^u] \cdot \frac{du}{dx} = e^{5x^3} \cdot (15x^2).$$

The previous paragraph represents work that you either think through mentally or write out as scratch work. Putting the pieces together gives us the overall answer.

$$\begin{aligned} f'(x) = \frac{d}{dx}[(x^2+4)^5 e^{5x^3}] &= \frac{d}{dx}[(x^2+4)^5] \cdot e^{5x^3} + (x^2+4)^5 \cdot \frac{d}{dx}[e^{5x^3}] \\ &= 5(x^2+4)^4(2x) \cdot e^{5x^3} + (x^2+4)^5 \cdot e^{5x^3} \cdot (15x^2) \\ &= 10x(x^2+4)^4 e^{5x^3} + 15x^2(x^2+4)^5 e^{5x^3} \end{aligned}$$

As we start to use our derivatives in applications, we will often need to factor our formulas. To illustrate this principle, we identify all of the common factors of the terms.

$$\begin{aligned} f'(x) &= 5x(x^2+4)^4 e^{5x^3} \cdot (2 + 3x(x^2+4)) \\ &= 5x(x^2+4)^4 e^{5x^3}(2 + 12x + 3x^2) \end{aligned}$$

□

9.4.5 Another Limit Defining e

When we defined the number e, it was done so that $\frac{d}{dx}\left[e^x\right] = e^x$. The natural base was then chosen so that

$$\lim_{x\to 0} \frac{e^x - 1}{x} = 1.$$

Suppose that we instead looked for a function $f(x)$ so that $f(0) = 1$ and $f'(x) = f(x)$. An equation for an unknown function involving derivatives is called a differential equation. We already know that $f(x) = e^x$ is the solution to this differential equation. However, we will use the differential equation to gain some additional insights into the function.

We begin by considering how we might approximate our function using only the differential equation. Consider the interval $[0, x]$ and create a partition with n subintervals,

$$x_k = \frac{kx}{n}.$$

We will recursively calculate a sequence of values $y_k \approx f(x_k)$. From the definition of an accumulation function, we know

$$f(x_{k+1}) = f(x_k) + \int_{x_k}^{x_{k+1}} f'(z)\, dz.$$

We further know $f'(z) = f(z)$ by our differential equation. So if $\Delta x = \frac{x}{n}$ is sufficiently small, we can approximate $f'(z)$ on the interval $[x_k, x_{k+1}]$ by a constant value $f'(z) \approx f'(x_k)$. This method for approximating the next value of the function defined by a differential equation,

$$f(x_{k+1}) \approx f(x_k) + f'(x_k)(x_{k+1} - x_k) = f(x_k) + f'(x_k)\Delta x,$$

is called the **Euler method**.

Our differential equation $f'(x) = f(x)$ allows us to find a value for $f'(x_k) = f(x_k)$. For our differential equation, the Euler method approximation gives us

$$f(x_{k+1}) \approx f(x_k) + f(x_k)(x_{k+1} - x_k) = f(x_k) \cdot (1 + \Delta x).$$

The approximation is then characterized by the recursive sequence,

$$\begin{aligned} y_0 &= 1, \\ y_{k+1} &= y_k(1 + \Delta x), \end{aligned}$$

which we identify as a geometric sequence with explicit formula

$$y_k = (1 + \Delta x)^k.$$

We also know that $f(x) = e^x$ is the solution. Since $x_n = x$, the Euler method approximation shows that $f(x_n) = e^x$ will be approximated by $y_n = (1 + \Delta x)^n$. Because $\Delta x = \frac{x}{n}$, we can write $n = \frac{x}{\Delta x}$ to obtain

$$f(x) = e^x \approx (1 + \Delta x)^{x/\Delta x} = \left((1 + \Delta x)^{1/\Delta x}\right)^x.$$

It would appear that e can be approximated by

$$e \approx (1 + \Delta x)^{1/\Delta x}$$

when Δx is sufficiently small. We state this heuristic result as the following unproved theorem.

Theorem 9.4.8

$$e = \lim_{h\to 0}(1+h)^{1/h} = \lim_{n\to\infty}(1+\tfrac{1}{n})^n.$$

We don't have a proof at this point because without first knowing the function e^x, we do not have a clear definition for nonrational powers to justify the following limit steps:

$$\begin{aligned} f(x) &= \lim_{\Delta x\to 0}\left[\left((1+\Delta x)^{1/\Delta x}\right)^x\right] \\ &= \left(\lim_{\Delta x\to 0}\left[(1+\Delta x)^{1/\Delta x}\right]\right)^x \\ &= e^x. \end{aligned}$$

Mathematically, the best way to prove that these limits are valid is actually to work with the inverse function, the natural logarithm.

9.5 Summary

- Given any positive base $b > 0$, we know

$$\frac{d}{dx}[b^x] = b^x \cdot L(b)$$

where $L(b)$ is defined by a limit

$$L(b) = \lim_{x \to 0} \frac{b^x - 1}{x}.$$

- The number e is the natural base such that $L(e) = 1$,

$$\lim_{x \to 0} \frac{e^x - 1}{x} = 1.$$

Consequently,

$$\frac{d}{dx}\Big[e^x\Big] = e^x.$$

- The general derivative rule for exponentials applies the chain rule,

$$\frac{d}{dx}\Big[e^u\Big] = e^u \cdot \frac{du}{dx}.$$

- Expressing other formulas involving powers in terms of e,

$$u^v = e^{v \cdot \ln(u)},$$

we can show

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

This also shows that

$$\lim_{x \to 0} \frac{b^x - 1}{x} = \ln(b).$$

- The function $f(x) = e^x$ is the solution to a differential equation $f'(x) = f(x)$ with $f(0) = 1$.

- $\lim_{h \to 0}(1 + h)^{1/h} = e$ and $\lim_{n \to \infty}(1 + \frac{1}{n})^n = e$.

9.5.1 Exercises

Foundations

1. Use the definition to find $L(\frac{1}{2})$ using a table to approximate the limit. Compare the result to $\ln(\frac{1}{2})$.

2. Use the definition to find $L(4)$ using a table to approximate the limit. Compare the result to $2\ln(2)$.

3. Find the tangent line of $y = 3^x$ at $x = 0$, using an exact expression.

Find the indicated derivatives.

4. $\frac{d}{dx}\Big[e^{4x}\Big]$

5. $\frac{d}{dx}\left[3^{2x}\right]$
6. $\frac{d}{dt}\left[5e^{-3t}\right]$
7. $\frac{d}{dx}\left[\frac{3}{e^{\frac{1}{2}x}}\right]$
8. $\frac{d}{ds}\left[\frac{1}{3e^{-5s}}\right]$
9. $\frac{d}{dx}\left[x \cdot 2^{-x}\right]$
10. $\frac{d}{dx}\left[e^{x^2}\right]$
11. $\frac{d}{dx}\left[e^{\sqrt{x}}\right]$
12. $\frac{d}{dx}\left[e^{e^x}\right]$
13. $\frac{d}{dx}\left[e^{(2x+1)^4}\right]$
14. $\frac{d}{dx}\left[(e^{3x}-1)^5\right]$
15. $\frac{d}{dt}\left[\frac{2}{e^{-t}+1}\right]$
16. $\frac{d^2}{dx^2}\left[4xe^{-2x}\right]$
17. $\frac{d^2}{dx^2}\left[x^2e^{5x}\right]$
18. $\frac{d^2}{dx^2}\left[e^{-x^2+3x}\right]$

Differential Equations

19. Show that $y(t) = Ae^{kt}$, where A and k are constants, is a solution to the differential equation $\frac{dy}{dt} = ky$ for any value of A. That is, using the proposed formula for $y(t)$, compute $\frac{dy}{dt}$ and $k \cdot y$ and show that they are equal.
20. Find a solution for the differential equation $\frac{dy}{dt} = 2y$ with an initial value $y(0) = 200$. Use the proposed formula from Exercise 9.4.7.19 and solve for the value A which also satisfies the initial value.
21. A population grows at a rate that is proportional to the current population size,
$$\frac{dP}{dt} = k \cdot P.$$
If the population P is currently 2000 individuals and is growing at an instantaneous rate of 40 individuals per day, find the value k and solve the differential equation. Use the proposed formula from Exercise 9.4.7.19. What will be the population size in one week?
22. A radioactive substance decays at a rate that is proportional to the current mass of the substance,
$$\frac{dM}{dt} = -k \cdot M.$$

If the mass M is currently 50 milligrams and is decaying at an instantaneous rate of 2 micrograms per second, find the value k and solve the differential equation. Use the proposed formula from Exercise 9.4.7.19. What will be the mass of the radioactive substance after one day?

9.6 Implicit Differentiation and Derivatives of Inverse Functions

When we developed the power rule for derivatives, we were able to prove the rule was true for positive integers as well as for the special case of integer multiples of $\frac{1}{2}$. Other reciprocal powers like $\frac{1}{3}$, $\frac{1}{4}$, or $\frac{1}{5}$ are the inverses of the corresponding integer powers 3, 4, or 5, respectively. Having established the differentiation rule for exponential functions, we want to find the rule for their inverses, the logarithm functions.

In preparation for differentiation rules of inverse functions, this section introduces the concept of an implicit function. Implicit functions and the chain rule result in the process of implicit differentiation, which creates an equation involving the derivative of the implicit function. We then use this process to complete the rules of differentiation.

9.6.1 Implicit Functions

A function is defined explicitly when a formula for the output is given in terms of the input. If we define a dependent variable using such a function, say $y = f(x)$, then we say y is defined explicitly as a function of x.

On the other hand, we also encounter situations where two variables are related through an equation but not in a way that is an explicit formula. When the graph of the equation defines a curve, say in the (x, y) plane, we say that the equation defines y as an **implicit function** of x. (It also defines x as an implicit function of y, depending on which variable we wish to consider as the dependent variable.)

Example 9.6.1 The equation $2x - 3y = 6$ has a graph that is a line. This equation defines y as an implicit function of x. If we solve for y to find

$$y = \frac{2x - 6}{3},$$

then we have now defined y as an explicit function of x. □

A graph does not represent a function when it fails the vertical line test. In spite of this, connected segments of the graph that individually pass the vertical line test can still be treated as functions for which we seek to find the derivative. It is in this context that we are interested in implicit functions.

Example 9.6.2 The equation $x^2 + y^2 = 4$ has a graph that is a circle centered at $(0, 0)$ and with radius 2. The graph fails the vertical line test which means that the relation fails to define y as a function of x. However, the top half of the circle considered separately is a function, as is the bottom half of the circle. We see this when we attempt to solve the equation for y and obtain

$$y^2 = 4 - x^2,$$
$$y = \pm\sqrt{4 - x^2}.$$

Each of the branches of the graph, $y = \sqrt{4 - x^2}$ and $y = -\sqrt{4 - x^2}$, defines y as an explicit function of x. The original equation $x^2 + y^2 = 4$ defines both of these functions implicitly without requiring solving for y. □

When an equation involving two variables (like x and y) has a graph that consists of curves, the connected components of those curves that individually

pass the vertical line test define the dependent variable (e.g., y) as an **implicit function** of the independent variable (e.g., x).

9.6.2 Implicit Differentiation

Once we recognize that an equation defines y as an implicit function of x, we can compute the derivative $y' = \frac{dy}{dx}$ using a process called implicit differentiation. Recall that when a dependent variable y is a function of x, computing a derivative of a function of y requires an application of the chain rule. We also must use any other differentiation rule as appropriate.

Example 9.6.3 Suppose that y is a function of x. Find the derivatives of the following expressions in terms of x, y and y'.

1. $\frac{d}{dx}[y^3]$
2. $\frac{d}{dx}[x^2y]$
3. $\frac{d}{dx}[xe^{x+y}]$

Solution. We find the derivative by recognizing how the expression is computed and using the appropriate rules of differentiation.

1. The expression y^3 is a composition $(y(x))^3$ so that the chain rule along with the power rule allow us to find the derivative:

$$\frac{d}{dx}[y^3] = 3y^2 \cdot y'.$$

2. The expression x^2y is a product of x^2 and $y(x)$, so the derivative will require using the product rule of derivatives.

$$\frac{d}{dx}[x^2y] = \frac{d}{dx}[x^2] \cdot y + x^2 \cdot \frac{d}{dx}[y] = 2xy + x^2y'$$

3. The expression xe^{x+y} is a product of x and e^{x+y}, so we start by using the product rule. To differentiate e^{x+y}, we need the chain rule for e^u where $u = x + y$. Finally, when we differentiate y, we get the function y'.

$$\begin{aligned}\frac{d}{dx}[xe^{x+y}] &= \frac{d}{dx}[x]e^{x+y} + x\frac{d}{dx}[e^{x+y}] \\ &= e^{x+y} + xe^{x+y}\frac{d}{dx}[x+y] \\ &= e^{x+y} + xe^{x+y}(1+y') \\ &= e^{x+y}(1+x+xy')\end{aligned}$$

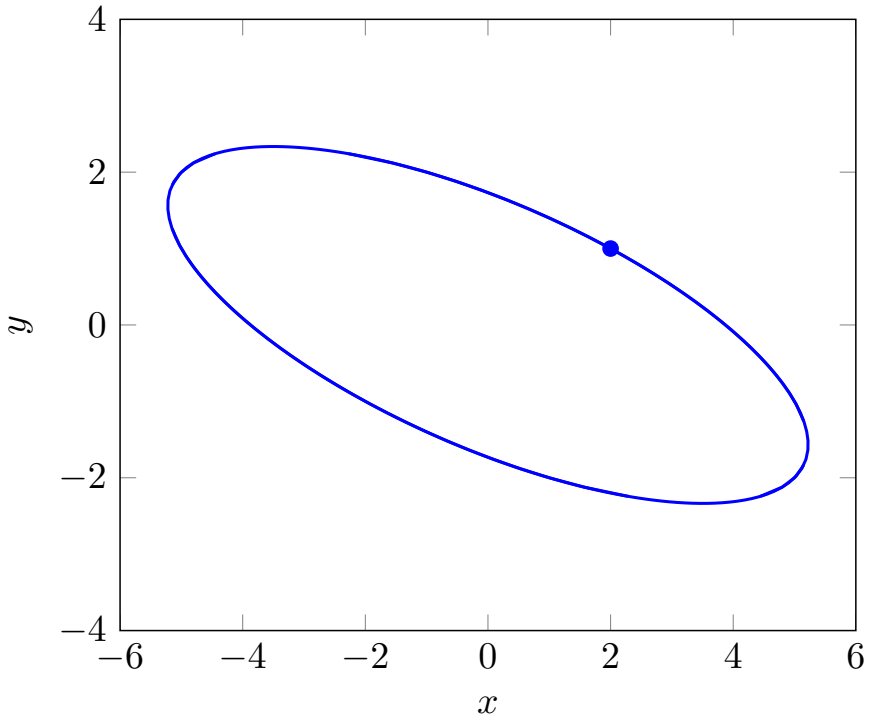
□

Implicit differentiation builds on the idea that if $f(x) = g(x)$ for all x in an interval, then $f'(x) = g'(x)$ on the same interval. That is, if two functions are equal then their derivatives must be equal. So consider an equation in x and y that defines y as an implicit function of x. We can think of the left side of the equation as defining one function (like the $f(x)$ in the earlier sentence) and the right side of the equation as defining a second function (like the $g(x)$). Then the derivatives of the two sides of the equations must also be equal.

Implicit differentiation uses the following steps.

1. Start with an equation involving your two variables, say x and y. It may be desirable to find an equivalent equation for which derivatives are easier to compute.
2. Create a new equation by differentiating each side of the equation. The dependent variable must be treated as an implicit function so that the new equation involves the variables x and y and the derivative $y' = \frac{dy}{dx}$).
3. Solve the new equation for $y' = \frac{dy}{dx}$ as a function of x and y. If the slope at a particular point is desired, substitute the values of x and y to find a value for y'.

Example 9.6.4 The equation $x^2 + 5y^2 = 15 - 3xy$ defines an ellipse, shown below. What is the slope of the curve at the point $(2, 1)$?



Solution. We recognize y as an implicit function of x and differentiate the two functions in the equation to create a new equation.

$$\frac{d}{dx}[x^2 + 5y^2] = \frac{d}{dx}[15 - 3xy]$$
$$2x + 10yy' = 0 - (3 \cdot y + 3x \cdot y')$$
$$2x + 10yy' = -3y - 3xy'$$

From this new equation, we solve for y' by moving all terms with y' to the same side of the equation.

$$2x + 3y = -3xy' - 10yy'$$

Next, factor out the common factor of y' and the solve for y' with division.

$$2x + 3y = -(3x + 10y)y'$$
$$-\frac{2x + 3y}{3x + 10y} = y'$$

This gives us the formula for the slope at any point in terms of x and y,

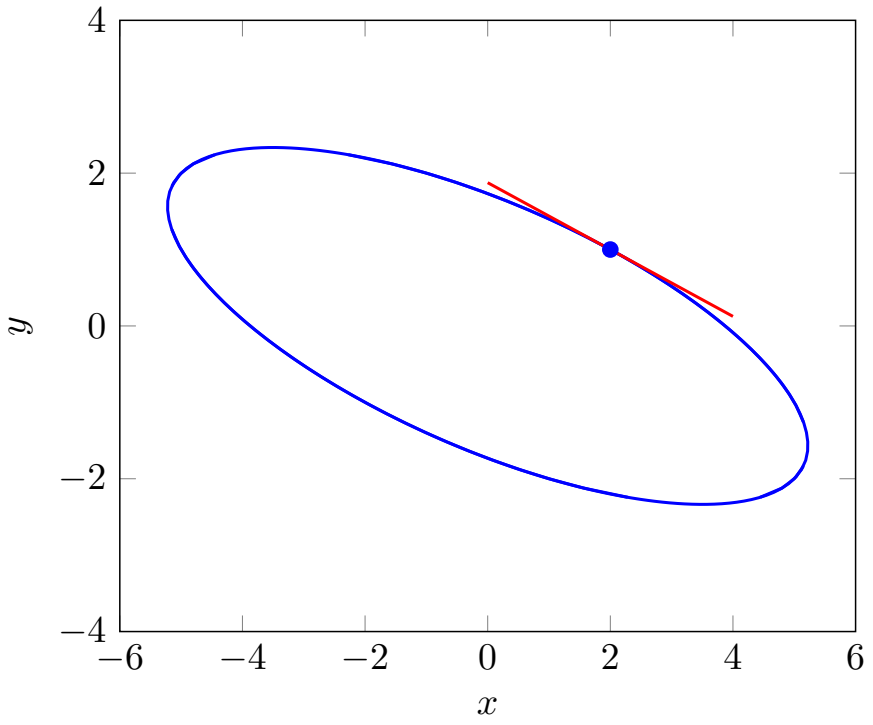
$$\frac{dy}{dx} = y' = -\frac{2x+3y}{3x+10y}.$$

To find the actual slope at the point $(x, y) = (2, 1)$, we use the values $x = 2$ and $y = 1$:

$$\left.\frac{dy}{dx}\right|_{(x,y)=(2,1)} = -\frac{2(2)+3(1)}{3(2)+10(1)} = -\frac{7}{16} = -\frac{7}{16}.$$

We could also find the equation of the tangent line knowing this information,

$$y = \frac{-7}{16}(x-2)+1.$$



□

9.6.3 Derivatives of Inverse Functions

Suppose that we know a function $f(x)$ and its derivative $f'(x)$. We are now interested in knowing how this information might relate to its inverse. In general, the function f does not necessarily have an inverse function unless it happens to be one-to-one. So suppose that f has an inverse function f^{-1}.

The equation for the graph of the inverse function is $y = f^{-1}(x)$. By virtue of being an inverse function, this equation is equivalent to the inverse equation

$$f(y) = x.$$

We can use this equation and the ideas of implicit differentiation to find the derivative of the inverse function,

$$\frac{d}{dx}[f^{-1}(x)] = \frac{dy}{dx} = y'.$$

Differentiating the left side of the inverse equation and the chain rule leads to an implicit differentiation equation

$$f'(y) \cdot y' = 1,$$

from which we can solve for y' to get

$$y' = \frac{dy}{dx} = \frac{1}{f'(y)}.$$

This result is saying that the slope for the inverse function is related to the slope of the original function. If the graph of f has a point (a, b) and a derivative $f'(a)$, then the graph of the inverse function includes the corresponding point (b, a) and has the reciprocal rate of change $\left.\frac{df^{-1}}{dx}\right|_b = \frac{1}{f'(a)}$. To find the rate of change of an inverse function, we need to identify the corresponding point for the original function, $a = f^{-1}(b)$. This is formally stated in the following theorem, with x as a variable in place of b.

Theorem 9.6.5 *Let f^{-1} be the inverse of a function f for which we know $\frac{d}{dx}[f(x)] = f'(x)$. Then*

$$\frac{d}{dx}[f^{-1}(x)] = \frac{1}{f'(f^{-1}(x))}.$$

This first example illustrates the principle for a specific point.

Example 9.6.6 A function $f(x) = x^3 + 3x$ has an inverse function because it is one-to-one, but the formula for the inverse function $f^{-1}(x)$ is not easy to find. Because $f(1) = 4$ we know $f^{-1}(4) = 1$. Find the equation of the tangent line to $y = f^{-1}(x)$ at $x = 4$.

Solution. This problem requires using the theorem for derivatives of inverse functions. We know that the original function $f(x) = x^3 + 3x$ has a derivative $f'(x) = 3x^2 + 3$. Consequently, the graph of f has a tangent line with slope $f'(1) = 3(1^2) + 3 = 6$ at the point $(1, 4)$. The inverse function $y = f^{-1}(x)$ must then have a corresponding point $(4, 1)$ and tangent line with slope $\frac{1}{6}$.

Formally, the theorem for derivatives of inverse functions states that for $y = f^{-1}(x)$,

$$\left.\frac{dy}{dx}\right|_{x=4} = \frac{1}{f'(f^{-1}(4))}.$$

Because $f^{-1}(4) = 1$ and $f'(1) = 6$, we know,

$$\left.\frac{dy}{dx}\right|_{x=4} = \frac{1}{f'(1)} = \frac{1}{6}.$$

Knowing the slope and the point, we can find the equation for the tangent line,

$$y = 1 + \frac{1}{6}(x - 4).$$

□

When finding the derivative of an inverse function with the goal of finding a formula, we need to simplify the expression $f'(f^{-1}(x))$. In many cases, this composition will simplify nicely.

Example 9.6.7 The functions $f(x) = x^2$ for $x \geq 0$ and $f^{-1}(x) = \sqrt{x}$ are inverse functions. Use the derivative of an inverse function to find $\frac{d}{dx}[\sqrt{x}]$.

Solution. The original equation for the square root is $y = \sqrt{x}$, which is

equivalent to the inverse equation,

$$y^2 = x.$$

Implicit differentiation leads to

$$2yy' = 1 \quad \Rightarrow \quad y' = \frac{1}{2y}.$$

Using the original inverse $y = \sqrt{x}$, this simplifies to

$$\frac{d}{dx}[\sqrt{x}] = y' = \frac{1}{2\sqrt{x}}.$$

Alternatively, just using the theorem for derivatives of inverse functions with $f(x) = x^2$ and $f^{-1}(x) = \sqrt{x}$, we have $f'(x) = 2x$ so that

$$\frac{d}{dx}[\sqrt{x}] = \frac{1}{f'(f^{-1}(x))} = \frac{1}{2\sqrt{x}}.$$

□

Example 9.6.8 The functions $f(x) = e^x$ and $f^{-1}(x) = \ln x$ are inverse functions. Use the derivative of an inverse function to find $\frac{d}{dx}[\ln(x)]$.

Solution. The original equation for the logarithm is $y = \ln(x)$, defined for $x > 0$, which is equivalent to the inverse equation,

$$e^y = x.$$

Implicit differentiation leads to

$$e^y y' = 1 \quad \Rightarrow \quad y' = \frac{1}{e^y}.$$

Using the original inverse $y = \ln(x)$, this simplifies to

$$\frac{d}{dx}[\ln(x)] = y' = \frac{1}{e^{\ln(x)}} = \frac{1}{x}.$$

The theorem for derivatives of inverse functions with $f(x) = e^x$ and $f^{-1}(x) = \ln(x)$, we have $f'(x) = e^x$ so that

$$\frac{d}{dx}[\ln(x)] = \frac{1}{f'(f^{-1}(x))} = \frac{1}{e^{\ln(x)}} = \frac{1}{x}.$$

□

The previous example is important. We summarize the result as a theorem. The implicit differentiation argument required $x > 0$. We can extend the result by considering the logarithm of the absolute value of x.

Theorem 9.6.9 Derivative of Natural Logarithm.

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

Proof. For $x > 0$, we have $|x| = x$ and $\ln(|x|) = \ln(x)$. Implicit differentiation showed $\frac{d}{dx}[\ln(|x|)] = \frac{1}{x}$. For $x < 0$, we have $|x| = -x$ and $\ln(|x|) = \ln(-x)$.

Differentiation requires the chain rule,

$$\begin{aligned}\frac{d}{dx}[\ln(|x|)] &= \frac{d}{dx}[\ln(-x)] \\ &\underset{u=-x}{=} \frac{1}{u}\cdot\frac{du}{dx} \\ &= \frac{1}{-x}\cdot(-1) = \frac{1}{x}.\end{aligned}$$

Therefore, the differentiation rule is true for all $x \neq 0$ ■

Using the chain rule gives us a more general differentiation rule,

$$\frac{d}{dx}[\ln(|u|)] = \frac{1}{u}\frac{du}{dx}.$$

This is summarized as a theorem.

Theorem 9.6.10 General Derivative of Natural Logarithm.

$$\frac{d}{dx}[\ln(|f(x)|)] = \frac{f'(x)}{f(x)}$$

9.6.4 Summary

- An equation in two variables generally defines a curve in the plane. When the equation is solved for one of the variables, the equation defines that dependent variable as an explicit function of the other.
- When an equation defines a curve but is not solved for one of the variables, we can still treat a dependent variable as an implicit function of the other. The curve overall may not satisfy the vertical line test for a function, but isolated segments of the curve could.
- Implicit differentiation treats a dependent variable as an implicit function and creates an equation for the derivative by differentiating both sides of the equation and applying the chain rule for any functions of the dependent variable.
- The equation for the derivative coming from implicit differentiation will typically depend on both variables.
- Finding the derivative of an inverse function $y = f^{-1}(x)$ is found by writing the equivalent inverse equation $x = f(y)$ and using implicit differentiation. This gives

$$\frac{df^{-1}}{dx} = \frac{1}{f'(y)} = \frac{1}{f'(f^{-1}(x))}.$$

- If $y = f(x)$ has a point $(x, y) = (a, b)$ with $\frac{df}{dx}(a) = m$, then $y = f^{-1}(x)$ has a corresponding point $(x, y) = (b, a)$ with $\frac{df^{-1}}{dx}(b) = \frac{1}{m}$.
- Because the natural logarithm is the inverse function of the natural exponential, we have

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

defined only for $x > 0$. Using $|x| = -x$ for $x < 0$, the chain rule gives us an extension for all $x \neq 0$,

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

The general application gives

$$\frac{d}{dx}[\ln(|f(x)|] = \frac{f'(x)}{f(x)}.$$

9.6.5 Exercises

9.7 Logarithmic Differentiation

Knowing the derivative of the logarithm, the chain rule, and the properties of logarithms, we can use the logarithm to find derivatives of formulas involving otherwise awkward products, quotients, or powers. If $f(x)$ involves a product, then $\ln(|f(x)|)$ involves a sum. If $f(x)$ involves a quotient, then $\ln(|f(x)|)$ involves a difference. And if $f(x)$ involves a power, then $\ln(|f(x)|)$ involves a product. Because the differentiation rules of sums and differences are simpler than for products and quotients, introducing a logarithm on an equation can often simplify the required work.

9.7.1 Logarithmic Differentiation

The product rule and quotient rules for derivatives can be considered to be consequences of the logarithm's derivative along with the chain rule. For example, consider a product $y = f(x)g(x)$. If we create an equivalent equation by applying the logarithm of the absolute value to both sides, we obtain a formula that can be expanded using the properties of the logarithm.

$$\ln(|y|) = \ln(|f(x)g(x)|) = \ln(|f(x)|) + \ln(|g(x)|).$$

This also used the property of absolute values $|a \cdot b| = |a| \cdot |b|$. Implicit differentiation gives an equation that relates the derivatives:

$$\frac{y'}{y} = \frac{f'(x)}{f(x)} + \frac{g'(x)}{g(x)}.$$

Multiplying through by $y = f(x)g(x)$, we obtain the product rule

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

A similar argument can be used to derive the quotient rule.

The argument from the previous paragraph is typical of the process we call **logarithmic differentiation**.

1. Introduce a dependent variable such as y equal to the expression to differentiate.
2. Apply the logarithm of the absolute value to both sides of that equation.
3. Use the properties of logarithms and absolute values to expand the formula until there are no more logarithms of products, quotients, or powers.
4. Use implicit differentiation and the general derivative of the logarithm to solve for y'.

When an expression has more than two factors or complicated powers, using logarithmic differentiation can often simplify the work of finding the derivative.

Example 9.7.1 Use logarithmic differentiation to compute

$$\frac{d}{dx}[x^3 e^{3x}(2x+1)^5].$$

Solution. Start by creating an equation involving a dependent variable.

$$y = x^3 e^{3x}(2x+1)^5$$

Apply the logarithm of the absolute value to both sides:

$$\ln(|y|) = \ln(|x^3 e^{3x}(2x+1)^5|).$$

Use the properties of logarithms to expand the right side:

$$\ln(|y|) = 3\ln(|x|) + \ln(|e^{3x}|) + 5\ln(|2x+1|).$$

Now use implicit differentiation and simplify:

$$\frac{y'}{y} = 3 \cdot \frac{1}{x} + \frac{3e^{3x}}{e^{3x}} + 5 \cdot \frac{2}{2x+1}$$
$$\frac{y'}{y} = \frac{3}{x} + 3 + \frac{10}{2x+1}$$
$$y' = y \cdot \left(\frac{3}{x} + 3 + \frac{10}{2x+1}\right)$$

We finish by substituting the original expression for y:

$$y' = x^3 e^{3x}(2x+1)^5 \cdot \left(\frac{3}{x} + 3 + \frac{10}{2x+1}\right).$$

□

9.7.2 Proving the Power Rule

Logarithmic differentiation allows us to differentiate additional functions for which other rules may not apply. We start by proving the power rule for arbitrary powers. Using the known differentiation rules and the definition of the derivative, we were only able to prove the power rule in the case of integer powers and the special case of rational powers that were multiples of $\frac{1}{2}$. Logarithmic differentiation gives us a tool that will prove it generally.

Theorem 9.7.2 *For any power p, $\frac{d}{dx}[x^p] = px^{p-1}$.*

Proof. Start with the equation $y = x^p$. Create an equivalent equation by taking the logarithm of the absolute value of both sides,

$$\ln(|y|) = \ln(|x^p|) = \ln(|x|^p).$$

Using the properties of logarithms, this can be rewritten

$$\ln(|y|) = p\ln(|x|).$$

We now use implicit differentiation and differentiate both sides of the equation:

$$\frac{d}{dx}[\ln(|y|)] = \frac{d}{dx}[p\ln(|x|)]$$
$$\frac{1}{y} \cdot \frac{dy}{dx} = p \cdot \frac{1}{x}.$$

Solving for the derivative, we have

$$\frac{dy}{dx} = p \cdot \frac{y}{x}.$$

Substituting the original equation $y = x^p$, we obtain the power rule

$$\frac{dy}{dx} = px^{p-1}.$$

■

We can also use logarithmic differentiation to find derivatives of functions

represented by powers that are neither power functions nor exponential functions. Recall that a power function must have a constant exponent and an exponential function must have a constant base. If the base and exponent both involve variables, then we are dealing with a function for which we have no differentiation rule.

Example 9.7.3 Find $\frac{d}{dx}[x^x]$.

Solution. The function $y = x^x$ is not an exponential or a power function. Using logarithmic differentiation, we first rewrite the equation

$$\ln(|y|) = \ln(|x^x|) = \ln(|x|^x) = x\ln(|x|).$$

The new expression involves only operations for which we have valid differentiation rules:

$$\begin{aligned}\frac{y'}{y} &= \frac{d}{dx}[x\ln(|x|)] \\ &= 1 \cdot \ln(|x|) + x \cdot \frac{1}{x} \\ &= \ln(|x|) + 1.\end{aligned}$$

Multiply by y and rewrite the formula gives

$$y' = y \cdot (\ln(|x|) + 1) = x^x(\ln(|x|) + 1).$$

The original function $y = x^x$ is not continuous for $x < 0$. This is because a power of a negative number is only defined at special rational powers. For example, $(-\frac{1}{2})^{-1/2}$ is not a real number but $(-2)^{-2}$ is. The function x^x is undefined almost everywhere for $x < 0$ and therefore y' does not exist for $x < 0$. Consequently, the absolute value is not actually necessary if we state

$$\frac{d}{dx}[x^x] = x^x(\ln(x) + 1), \quad x > 0.$$

□

9.7.3 Summary

- To differentiate an expression $f(x)$ using logarithmic differentiation:
  1. Create an equation $y = f(x)$.
  2. Apply the logarithm of the absolute value to both sides of that equation, $\ln(|y|) = \ln(|f(x)|)$.
  3. Use the properties of logarithms and absolute values to expand the formula $\ln(|f(x)|)$.
  4. Use implicit differentiation and solve for y'.
- Logarithmic differentiation is useful when working with a product or quotient for which the direct rules will be cumbersome, as well as for any power.

9.7.4 Exercises

Use logarithmic differentiation to compute the derivatives.

1. $\frac{d}{dx}\left[4(x+1)^3(2x-1)^4(x^2+3)^5\right]$

2. $\dfrac{d}{dx}\left[\dfrac{2\sqrt{x}(x-1)^3}{(x^3+3x)^5}\right]$

3. $\dfrac{d}{dx}\left[(x^2+1)^x\right]$

4. $\dfrac{d}{dx}\left[(2x+1)^{\ln(x)}\right]$

5. $\dfrac{d}{dx}\left[x^{(e^x))}\right]$

General consequences of logarithmic differentiation.

6. Use logarithmic differentiation to prove the quotient rule.

7. Use logarithmic differentiation to establish a product rule for three factors. That is, find $\frac{d}{dx}[f(x)g(x)h(x)]$.

Chapter 10

Derivatives and Integrals

10.1 Antiderivatives

We have previously studied the differentiation operator. Given a function relationship between two variables $x \overset{f}{\mapsto} Q$, the derivative f' is the function relating x to the rate of change $\frac{dQ}{dx}$. Differentiation is the operation that goes maps $f \overset{\frac{d}{dx}}{\mapsto} f'$. Because f' is itself a function, we can apply differentiation again $f' \overset{\frac{d}{dx}}{\mapsto} f''$. This process can repeat indefinitely.

Consider as an example $f(x) = x^4 + 2x^2 - 3x$. There is a sequence of functions corresponding to the derivatives:

$$\begin{aligned} f(x) &= x^4 + 2x^2 - 3x, \\ f'(x) &= 4x^3 + 4x - 3, \\ f''(x) &= 12x^2 + 4, \\ f^{(3)}(x) &= 24x, \\ f^{(4)}(x) &= 24, \\ f^{(5)}(x) &= 0, \\ f^{(6)}(x) &= 0. \end{aligned}$$

This pattern continues with $f^{(n)}(x) = 0$ for $n = 5, 6, 7, \ldots$.

As the example above illustrates, given a function we can find its derivative. One of the major themes of mathematics is the idea of inverse operations. Is there an inverse operation to differentiation? That is, given $f(x)$, instead of computing $f'(x)$, can we find a function $F(x)$ so that $F(x) \overset{\frac{d}{dx}}{\mapsto} f(x)$? This inverse operation, using $f(x)$ to find $F(x)$, is called **antidifferentiation**.

In this section, we define antiderivatives. We discuss why a function has infinitely many different antiderivatives. Based on the First Part of the Fundamental Theorem of Calculus, we recognize that accumulation functions are special examples of antiderivatives for continuous rates of accumulation. Motivated by this observation, we introduce the indefinite integral as the notation for antidifferentiation. Examples will illustrate how we use our known differentiation rules to develop corresponding antidifferentiation rules.

10.1.1 Terminology

Definition 10.1.1 Antiderivatives. Given a function $f(x)$, we say that $F(x)$ is an **antiderivative** of $f(x)$ if $f(x)$ is the derivative of $F(x)$. That is, $F'(x) = f(x)$. ◊

The derivative of any constant is zero, so adding a constant to a function creates a new function that has the same derivative as the original. This means that differentiation is not one-to-one.

Example 10.1.2 Compare the following derivatives:

$$\begin{aligned} \frac{d}{dx}[x^2 + 3x] &= 2x + 3, \\ \frac{d}{dx}[x^2 + 3x - 1] &= 2x + 3, \\ \frac{d}{dx}[x^2 + 3x + 4] &= 2x + 3. \end{aligned}$$

Each of the functions have the same derivative. We say that x^2+3x, x^2+3x-1, and $x^2 + 3x + 4$ are all antiderivatives of $2x + 3$. More generally, we know

$x^2 + 3x + C$ will be an antiderivative for *any* constant value C. □

If we know that a function $F(x)$ is an antiderivative of $f(x)$, then we know that all functions of the form $F(x) + C$, where C is a constant, are also antiderivatives. This shows that infinitely many different functions have the same derivative. We call all such functions **antiderivatives**.

We will later prove the following theorem. It states that the *only* way that two antiderivatives can be different is that they differ by a constant. The proof of the theorem will use a Mean Value Theorem for derivatives.

Theorem 10.1.3 *Suppose that $F(x)$ and $G(x)$ are both antiderivatives of $f(x)$ on an interval I. That is, for all $x \in I$ we have*

$$F'(x) = G'(x) = f(x).$$

Then there is a constant C so that for all $x \in I$, $G(x) = F(x) + C$.

Consequently, knowing just one antiderivative allows us to determine all possible antiderivatives by adding some constant. Suppose $F(x)$ is an antiderivative of $f(x)$. Then any other antiderivative must be $F(x) + C$ for some constant C. If we leave the constant as an unspecified parameter, we call this the **general antiderivative**. Graphically, different antiderivatives correspond to a vertical translation of the graph. That is, all antiderivatives have the same graph shifted up or down relative to one another.

In the case that $f(x)$ is continuous on some interval I, we can define an accumulation function starting at any convenient point $a \in I$,

$$A(x) = \int_a^x f(z)\,dz.$$

By the Part One of the Fundamental Theorem of Calculus, we know that $A'(x) = f(x)$. That is, $A(x)$ is itself an antiderivative of $f(x)$ and any other antiderivative could be written $F(x) = \int_a^x f(z)\,dz + C$.

Owing to this close connection between antiderivatives and integrals, the standard notation for finding antiderivatives is with the integral symbol using an indefinite integral. An indefinite integral will not have any limits of integration, uses the same variable of integration as the independent variable, and refers to antiderivatives rather than definite integrals.

Definition 10.1.4 Indefinite Integrals. Given a function $f(x)$, the **indefinite integral** of $f(x)$ with respect to x, written $\int f(x)\,dx$, is the general antiderivative of $f(x)$. That is, if $F(x)$ is any antiderivative such that $F'(x) = f(x)$, then

$$\int f(x)\,dx = F(x) + C.$$

◊

Using our earlier example, we can write the indefinite integral of $2x + 3$ as

$$\int 2x + 3\,dx = x^2 + 3x + C.$$

The indefinite integral represents the infinite family of all antiderivatives of $2x + 3$.

10.1.2 Examples

For the most part, finding antiderivatives corresponds to recognizing how a function might have been computed as a derivative. Every statement about

differentiation has an equivalent statement about integrals. To check whether a proposed function is an antiderivative, we calculate its derivative and compare that with the function inside the integral.

Example 10.1.5 Find $\dfrac{d}{dx}\left[(3x+5)^4\right]$ and then write down the equivalent statement as an integral.

Solution. The last operation in the expression $(3x+5)^4$ is the power acting on the expression $u = 3x+5$. The derivative requires a chain rule:

$$\begin{aligned}\frac{d}{dx}\left[(3x+5)^4\right] &= 4u^3\frac{du}{dx}\\ &= 4(3x+5)^3(3)\\ &= 12(3x+5)^3.\end{aligned}$$

Once we know the derivative, we can write the equivalent integral

$$\int 12(3x+5)^3\,dx = (3x+5)^4 + C.$$

This says that $(3x+5)^4$ is an antiderivative of $12(3x+5)^3$, along with that same formula plus any constant. □

We must learn to recognize which differentiation rules would result in a particular formula for a given function. Because differentiation is a linear operator, antidifferentiation is as well.

Theorem 10.1.6 *If $F(x)$ is an antiderivative of $f(x)$ and $G(x)$ is an antiderivative of $g(x)$, then for any constants c_1 and c_2, $c_1F(x)+c_2G(x)$ is an antiderivative of $c_1f(x)+c_2g(x)$. We write*

$$\int [c_1f(x)+c_2g(x)]dx = c_1\int f(x)\,dx + c_2\int g(x)\,dx.$$

If the integrand $f(x)$ is expressed as a sum of terms, we typically first try to find antiderivatives of each term.

Example 10.1.7 Find $\displaystyle\int 4x^3 - 2e^{2x}\,dx$.

Solution. We are looking for a function $F(x)$ for which $F'(x) = 4x^3 - 2e^{2x}$. From experience computing derivatives, we know

$$\begin{aligned}\frac{d}{dx}[x^4] &= 4x^3,\\ \frac{d}{dx}[e^{2x}] &= 2e^{2x}.\end{aligned}$$

This suggests we should use the difference $F(x) = x^4 - e^{2x}$. We verify by differentiation:

$$F'(x) = \frac{d}{dx}[x^4 - e^{2x}] = 4x^3 - 2e^{2x}.$$

This verifies that $F(x)$ is an antiderivative of $4x^3 - 2e^{2x}$. The general antiderivative is written as the indefinite integral,

$$\int 4x^3 - 2e^{2x}\,dx = x^4 - e^{2x} + C.$$

□

Most derivative rules do not result in a product of expressions. The product

rule for derivatives results in the sum of two products. The quotient rule results in in difference of quotients. Only the chain rule creates a derivative by multiplying two expresions together. Consequently, if we see an integrand with expressions multiplied together, we should consider whether we would benefit from expanding the product as a sum.

Example 10.1.8 Find $\int x^2(x^2-3)\,dx$.

Solution. The function $f(x) = x^2(x^2-3)$ is a product that can be expanded to a sum using the distributive property.

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

Our experience with the power rule suggests that we should be able to integrate this expression. We know

$$\frac{d}{dx}[x^5] = 5x^4.$$

To eliminate the unwanted constant multiple of 5, we can multiply both sides by $\frac{1}{5}$ to get

$$\frac{d}{dx}\left[\frac{1}{5}x^5\right] = x^4.$$

This suggests an antiderivative

$$F(x) = \frac{1}{5}x^5 - x^3.$$

We verify using regular differentiation rules:

$$\begin{aligned} F'(x) &= \frac{d}{dx}[\frac{1}{5}x^5 - x^3] \\ &= \frac{1}{5}(5x^4) - 3x^2 \\ &= x^4 - 3x^2 = f(x). \end{aligned}$$

We have found

$$\int x^2(x^2-3)\,dx = \frac{1}{5}x^5 - x^3 + C.$$

□

Just as it is useful to collect and learn the basic building blocks for differentiation, we can collect and learn basic building blocks for integration. Each derivative rule has its equivalent statement about antiderivatives. If we incorporate the chain rule, we extend each of the elementary rules to generalized rules.

1. Power Rule: For any power $n \neq -1$,
$$\int x^n\,dx = \frac{1}{n+1}x^{n+1} + C.$$
2. Generalized Power Rule: For any power $n \neq -1$ and expression u,
$$\int u^n \cdot \frac{du}{dx}\,dx = \frac{1}{n+1}u^{n+1} + C.$$

3. Logarithm Rule:
$$\int \frac{1}{x}\,dx = \ln(|x|) + C.$$

4. Generalized Logarithm Rule: For any expression u,
$$\int \frac{u'}{u}\,dx = \ln(|u|) + C.$$

5. Elementary Exponential Rule: For any real value $k \neq 0$,
$$\int e^{kx}\,dx = \frac{1}{k}e^{kx} + C.$$

6. Generalized Exponential Rule: For any expression u,
$$\int e^u \cdot \frac{du}{dx}\,dx = e^u + C.$$

Example 10.1.9 $\int x^2 e^{x^3}\,dx$

Solution. Because the integrand has a product of expressions, we should begin by looking to see if the problem involves the chain rule. The exponential term e^{x^3} involves the expression $u = x^3$ which has a derivative $u' = 3x^2$. Notice that the other factor in the problem, x^2, differs from u' only by a constant multiple. That is, we can recognize our problem as a generalized exponential

$$\begin{aligned}\int x^2 e^{x^3}\,dx &= \int \frac{1}{3}(3x^2)e^{x^3}\,dx \\ &\underset{u=x^3}{=} \int \frac{1}{3}e^u \cdot \frac{du}{dx}\,dx \\ &\underset{u=x^3}{=} \frac{1}{3}e^u + C \\ &= \frac{1}{3}e^{x^3} + C.\end{aligned}$$

□

10.1.3 Finding a Particular Antiderivative

Adding a constant to a function represents a graphical transformation of a vertical shift. Consequently, different antiderivatives have the same graph shifted vertically from one another. Consider the function $f(x) = x^2 - 4x$. Integration gives us

$$\int x^2 - 4x\,dx = \frac{1}{3}x^3 - 2x^2 + C.$$

The function $F(x) = \frac{1}{3}x^3 - 2x^2$ has the derivative $F'(x) = x^2 - 4x$, as does every function $F(x) + C$.

The following dynamic graph has a slider for the integration constant C. Notice that changing the value of C shifts the graph up or down. See if you can find a value so that the graph $y = F(x) + C$ goes through $(x, y) = (3, 2)$.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 10.1.10 $y = F(x) + C$

We can solve for the integration constant to find a particular antiderivative passing through a given point.

Example 10.1.11 Find the constant C so that $F(x) = \frac{1}{3}x^3 - 2x^2 + C$ satisfies $F(3) = 2$.

Solution. Substitute the value $x = 3$ into the equation for $F(x)$.

$$\begin{aligned} F(3) &= \frac{1}{3}(3^3) - 2(3^2) + C \\ &= 9 - 18 + C \\ &= -9 + C \end{aligned}$$

Because we want $F(3) = 2$, we create the equation

$$-9 + C = 2$$

so that we can solve for C to get $C = 11$. □

Example 10.1.12 Find a function $P(t)$ so that $P'(t) = 20e^{-2t} + 3t$ and $P(0) = 50$.

Solution. Start by finding the general antiderivative.

$$\int [20e^{-2t} + 3t]dt = -10e^{-2t} + \frac{3}{2}t^2 + C$$

We therefore see that $P(t) = -10e^{-2t} + \frac{3}{2}t^2 + C$. Now we substitute $t = 0$ and $P(0) = 50$ to solve for C.

$$\begin{aligned} P(0) &= -10e^0 + \frac{3}{2}(0^2) + C \\ 50 &= -10 + C \\ 60 &= C \end{aligned}$$

Having found $C = 60$, we can conclude

$$P(t) = -10e^{-2t} + \frac{3}{2}t^2 + 60.$$

□

Because the derivative represents a rate of change, finding particular antiderivatives is equivalent to finding a quantity as a function of an independent variable when we know the rate of change as a function and we know an initial value.

Example 10.1.13 A cup of coffee starts at a temperature of 160 degrees Fahrenheit. The temperature changes at a rate of change (degrees per minute) modeled by the formula $-3.6e^{-0.04t}$ where t is the time in minutes. Find the temperature as a function of time.

Solution. Let T represent the temperature of the cup of coffee in degrees Fahrenheit. Our given information shows that

$$\frac{dT}{dt} = -3.6e^{-0.04t}.$$

The temperature T must be an antiderivative of this formula,

$$\begin{aligned} T &= \int -3.6e^{-0.04t}\,dt \\ &= \frac{-3.6}{-0.04}e^{-0.04t} + C \\ &= 90e^{-0.04t} + C. \end{aligned}$$

To find the value of C, substitute $t = 0$ and $T = 160$.

$$\begin{aligned} T &= 90e^{-0.04t} + C \\ 160 &= 90e^{0} + C \\ 160 &= 90 + C \\ 70 &= C \end{aligned}$$

Consequently, we have $T = 90e^{-0.04t} + 70$. □

10.1.4 Summary

1. An antiderivative of $f(x)$ is any function $F(x)$ so that $\frac{d}{dx}[F(x)] = f(x)$. If $F(x)$ is an antiderivative of $f(x)$, then so is $F(x) + C$ for any value of C.

2. The Fundamental Theorem of Calculus guarantees that every continuous function has an antiderivative. In particular, if $f(x)$ is continuous on an interval I with $a \in I$, then the accumulation function

$$A(x) = \int_a^x f(z)\,dz$$

is an antiderivative on the interval I.

3. We use the **indefinite integral** as the operator for antidifferentiation. For a function $f(x)$ with antiderivative $F(x)$, we write

$$\int [f(x)]\,dx = F(x) + C$$

where C (or any other chosen symbol) represents an arbitrary **constant of integration**.

4. The constant of integration graphically represents an arbitrary vertical shift of the graph of a function. Given any point representing an initial value, we can solve for the constant of integration so that there is the graph of an antiderivative which passes through the given point.

10.1.5 Exercises

Calculate the specified derivative and then write the equivalent indefinite integral.

1. $\dfrac{d}{dx}\left[2x^4\right]$

2. $\dfrac{d}{dx}\left[(2x+3)^5\right]$

3. $\dfrac{d}{dx}\left[\sqrt{x^2+3}\right]$

4. $\dfrac{d}{dx}\left[\ln(|x^2-4x|)\right]$

5. $\dfrac{d}{dx}\left[e^{3x^4}\right]$

6. $\dfrac{d}{dx}\left[x^2e^{-3x}\right]$

7. $\dfrac{d}{dx}\left[x\ln(|x|)\right]$

8. $\dfrac{d}{dx}\left[\dfrac{x-1}{x-3}\right]$

Compute the indefinite integral by finding the general antiderivative. Some integrands need to be rewritten before integration.

9. $\displaystyle\int -3x^5+2x^2+3\,dx$

10. $\displaystyle\int 2x-4x^{-1}+5x^{-3}\,dx$

11. $\displaystyle\int x^3(3x^2-4x+7)\,dx$

12. $\displaystyle\int (x+4)(x-8)\,dx$

13. $\displaystyle\int \frac{x^2+4x-5}{3x^2}\,dx$

14. $\displaystyle\int e^{2x}\,dx$

15. $\displaystyle\int 4e^{-3x}\,dx$

16. $\displaystyle\int xe^{x^2}\,dx$

17. $\displaystyle\int 2x^3e^{-x^4}\,dx$

18. $\displaystyle\int \frac{1}{x+3}\,dx$

19. $\displaystyle\int \frac{3}{2x+1}\,dx$

20. $\displaystyle\int \frac{x}{x^2+4}\,dx$

21. $\displaystyle\int \frac{e^{2x}}{e^{2x}+1}\,dx$

22. $\displaystyle\int -xe^{-x}+e^{-x}\,dx$

23. $\displaystyle\int \frac{2xe^{2x}-e^{2x}}{x^2}\,dx$

Use the given information to find the particular function.

24. Find $f(x)$ if $f'(x) = 2x - 5$ with $f(1) = 4$.

25. Find $g(x)$ if $g'(x) = 3e^{-3x}$ with $g(0) = 2$.

26. The velocity of a vehicle on track that runs left to right is $v(t) = \frac{1}{2}t^2 - 8t + 24$. If the vehicle is at a position $s = 0$ when $t = 1$, find the position $s(t)$ as a function of time.

27. A population changes at a rate defined by $R(t) = 0.24t^2 - 24t + 216$, where t is measured in years. If the population is $P = 120000$ when $t = 0$, find the population as a function of time.

28. A radiation detector absorbs radiation at a rate of $R(t) = 5e^{-0.1t}$ (grays per minute). Find the total amount of radiation absorbed by the detector as a function of time t (minutes) since $t = 0$.

10.2 Differentiable Functions

When we learned about the definite integral and defined accumulation functions, we were able to characterize the behavior of those functions by considering the behavior of the rates of accumulation. In particular, we learned that the monotonicity of a function depended on the sign of the rate of accumulation, and the concavity of a function depended, in turn, on the monotonicity of the rate. These concepts have a natural analogue for functions in terms of their derivatives, even when a function is not defined as an accumulation function.

In this section, we introduce the major theorems relating to differentiability. Differentiability is a property of a function characterized by where the derivative is defined. We first learn that local extremes can only occur at critical points, or points where the derivative equals zero or is not defined. We then learn about Rolle's theorem, which is a theorem guaranteeing a point with zero derivative. Rolle's theorem principal value is in proving the Mean Value Theorem for Derivatives. The Mean Value Theorem plays a prominent role in characterizing the behavior of functions. In particular, it will be used to prove that antiderivatives can only differ by constants.

10.2.1 Differentiability of Functions

Recall that continuity and differentiability are properties of functions. To say that a function is continuous at a point means that the function itself has a value at that point and that the limits of the function from both the left and the right converge to the same value. The property of continuity essentially characterizes the idea that the graph of the function is connected at the given point. In a similar way, differentiability is a property that the limit defining the derivative at a point is defined. Differentiability guarantees that a function has a linear tangent line approximation.

Now that we know how to compute derivatives with the rules of differentiation, we can consider when these functions are differentiable. As an example, consider power functions $f(x) = x^p$. When p is an irrational number, this is defined in terms of the exponential $f(x) = e^{p\ln(x)}$, so that the domain is $x > 0$. However, when p is a rational number $p = \frac{k}{n}$ for integers k and n with $n > 0$, then $f(x) = x^{k/n}$ is defined by the nth roots of x,

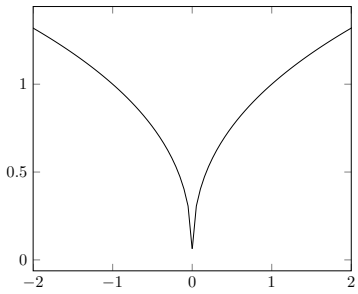
$$f(x) = x^{k/n} = (\sqrt[n]{x})^k.$$

For odd values n, the root $\sqrt[n]{x}$ is defined for all values of x. However, $f(0)$ is only defined if $k \geq 0$.

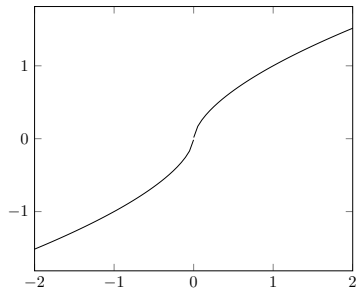
What about the derivative? We have

$$f'(x) = \frac{k}{n}x^{(k-n)/n}.$$

If $k \geq n$, then $f'(0)$ will exist. However, if $k < n$ corresponding to $0 < p = \frac{k}{n} < 1$, then $f'(0)$ will not exist. This is an example of a nondifferentiable function. Graphically, the tangent line at the point is vertical so that the slope is undefined with infinite limits.



(a) $p = \frac{2}{5}$



(b) $p = \frac{3}{5}$

Figure 10.2.1 Examples of power functions that are nondifferentiable at $x = 0$.

When we first introduced the concept of differentiability, we used piecewise functions to provide examples of nondifferentiable functions. In those examples, we used the definition of the derivative. With the rules of differentiation, we can determine differentiability more directly.

Theorem 10.2.2 *Suppose that a function $f(x)$ is defined piecewise around $x = a$ so that for some $\delta > 0$,*

$$f(x) = \begin{cases} f_\ell(x), & a - \delta < x < a, \\ f(a), & x = a, \\ f_r(x), & a < x < a + \delta. \end{cases}$$

If $f_\ell(x)$ and $f_r(x)$ in their natural domains are both continuous and differentiable at $x = a$, then f is differentiable at $x = a$ if and only if $f_\ell(a) = f_r(a) = f(a)$ and $f'_\ell(a) = f'_r(a)$.

Proof. Because f_ℓ and f_r are continuous, the requirements of continuity that

$$\lim_{x \to a^-} f(x) = f(a) \quad \text{and} \quad \lim_{x \to a^+} f(x) = f(a)$$

are replaced by $f_\ell(a) = f(a)$ and $f_r(a) = f(a)$. Similarly, the calculation of the derivative using the definition reduces to the values of the derivatives of f_ℓ and f_r at $x = a$:

$$\begin{aligned} \lim_{h \to 0^-} \frac{f(a+h) - f(a)}{h} &= \lim_{h \to 0^-} \frac{f_\ell(a+h) - f_\ell(a)}{h} \\ &= f'_\ell(a), \\ \lim_{h \to 0^+} \frac{f(a+h) - f(a)}{h} &= \lim_{h \to 0^+} \frac{f_r(a+h) - f_r(a)}{h} \\ &= f'_r(a). \end{aligned}$$

For the two sided limit to exist, and thus for the derivative itself to exist, the left- and right-side limits must agree, $f'_\ell(a) = f'_r(a)$. Then $f'(a) = f'_\ell(a) = f'_r(a)$. ■

Example 10.2.3 Determine the values of a and b so that the function

$$f(x) = \begin{cases} x^2 - 2x, & x \le 2, \\ -2x^2 + ax + b, & x > 2, \end{cases}$$

is differentiable at $x = 2$.

Solution. The function used for $x < 2$ is $f_\ell(x) = x^2 - 2x$, and the function used for $x > 2$ is $f_r(x = -2x^2 + ax + b$. The derivatives are found using

differentiation rules:

$$f'_\ell(x) = 2x - 2,$$
$$f'_r(x) = -4x + a.$$

The requirement for continuity will give us one equation, which we simplify: which becomes

$$f_\ell(2) = f_r(2)$$
$$2^2 - 2(2) = -2(2^2) + a(2) + b$$
$$0 = 2a + b - 8$$
$$2a + b = 8.$$

This means that so long as $b = 8 - 2a$, $f(x)$ will be continuous at $x = 2$. However, it may or may not be differentiable, depending on whether the derivatives match.

The requirement that the left- and right-sided derivatives are equal gives us a second equation, which we also simplify:

$$f'_\ell(2) = f'_r(2)$$
$$2(2) - 2 = -4(2) + a$$
$$2 = -8 + a$$
$$a = 10.$$

Once we know $a = 10$, we can substitute that into the first equation to find b:

$$b = 8 - 2a$$
$$b = 8 - 2(10)$$
$$b = -12.$$

Consequently, $f(x)$ will be differentiable at $x = 2$ if and only if $a = 10$ and $b = -12$.

Specify static image with `@preview` attribute,
Or create and provide automatic screenshot as
`images/interactive-piecewise-differentiable-preview.png`
via the `mbx` script

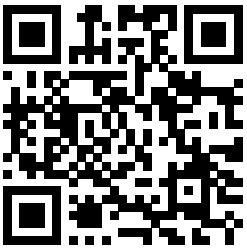


Figure 10.2.4 A graph of $f(x)$ where parameters a and b can be changed dynamically.

□

10.2.2 Consequences of Differentiability

There are a number of important consequences of a function being differentiable. These consequences are stated as mathematical theorems. The first such theorem focuses on differentiability at local extreme values.

Theorem 10.2.5 Fermat's Theorem. *If f has a local extreme at $x = a$ and $f'(a)$ exists, then $f'(a) = 0$.*

Proof. Suppose that f has a local maximum at $x = a$. Then there is some value $\delta > 0$ so that if $a - \delta < x < a + \delta$, we must have $f(x) \leq f(a)$. For $-\delta < h < 0$, we therefore have $f(a + h) - f(a) \leq 0$ so that dividing by $h < 0$ gives

$$\frac{f(a+h) - f(a)}{h} \geq 0.$$

This implies that

$$\lim_{h \to 0^-} \frac{f(a+h) - f(a)}{h} \geq 0.$$

For $0 < h < \delta$, we also have $f(a+h) - f(a) \leq 0$ so that dividing by $h > 0$ gives

$$\frac{f(a+h) - f(a)}{h} \leq 0.$$

Thus, we have

$$\lim_{h \to 0^+} \frac{f(a+h) - f(a)}{h} \leq 0.$$

If $f'(a)$ exists, these limits must equal and $f'(a) = 0$.

If f has a local minimum at $x = a$, the argument is similar. ■

If we are looking for extreme values of a function, we can ignore all points where $f'(x)$ exists but $f'(x) \neq 0$. The only points in the domain of f that might be considered are where $f'(x)$ does not exist or where $f'(x) = 0$ and f has a horizontal tangent line. We call such points the **critical points** of f.

Definition 10.2.6 The **critical points** of a function f are all values in the domain of f such that $f'(x)$ does not exist or $f'(x) = 0$. ◊

The second theorem combines the Extreme Value Theorem with Fermat's Theorem. If a function is continuous on a closed interval $[a, b]$, then it must achieve both a maximum and a minimum value. If that function has $f(a) = f(b)$, then one of the extreme values must occur inside the interval at some point $c \in (a, b)$. If the function is also differentiable, then we must have $f'(c) = 0$. This result is named Rolle's theorem.

Theorem 10.2.7 Rolle's Theorem. *If f is continuous on $[a, b]$ and differentiable on (a, b) and $f(a) = f(b)$, then there must be some value $c \in (a, b)$ so that $f'(c) = 0$.*

Proof. The argument is given in the paragraph preceding the theorem. The hypothesis of continuity allows us to apply the Extreme Value Theorem. The hypothesis of differentiability allows us to apply Fermat's Theorem to the local extreme that was guaranteed at the point between a and b. ■

The consequence of Rolle's theorem is that if a function starts and ends at the same value over an interval, it must turn around somewhere with a horizontal tangent.

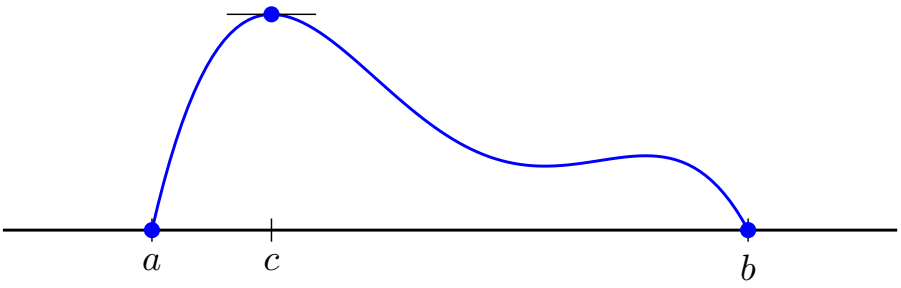


Figure 10.2.8 A graphical illustration of Rolle's theorem. Note that extreme values have horizontal tangents.

Rolle's theorem is not usually applied on its own. It is most often referenced in the context of proving more useful theorems. The third theorem about differentiability applies Rolle's theorem to create the Mean Value Theorem for derivatives in relation to the average rate of change. Recall that the average rate of change,

$$\left.\frac{\Delta f}{\Delta x}\right|_{[a,b]} = \frac{f(b)-f(a)}{b-a},$$

is the slope of the line, called a **secant line**, that joins the points $(a, f(a))$ and $(b, f(b))$. The Mean Value Theorem guarantees that a continuous and differentiable function will have some point at which the tangent line has the same slope as the secant line over the given interval.

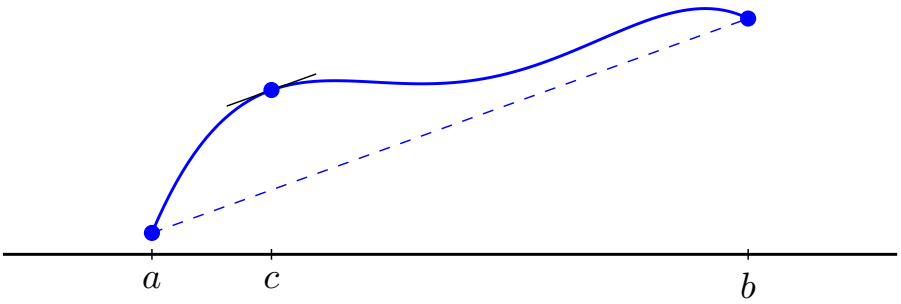


Figure 10.2.9 A graphical illustration of the Mean Value theorem. Note that at the point furthest from the secant line (dashed), the slope matches that of the secant line.

Theorem 10.2.10 Mean Value Theorem. *If f is continuous on $[a,b]$ and differentiable on (a,b), then there must be some value $c \in (a,b)$ so that*

$$f'(c) = \left.\frac{\Delta f}{\Delta x}\right|_{a,b} = \frac{f(b)-f(a)}{b-a}.$$

Alternatively, we sometimes rewrite this as

$$f(b)-f(a) = f'(c)\cdot(b-a).$$

Proof. Let $s(x)$ be the linear function corresponding to this secant line. That is, $s(a) = f(a)$ and $s(b) = f(b)$ and $s(x)$ has the constant slope

$$s'(x) = \frac{f(b)-f(a)}{b-a}.$$

We now define $g(x) = f(x) - s(x)$. Since $s(a) = f(a)$ and $s(b) = f(b)$, we have $g(a) = g(b) = 0$. If f is continuous and differentiable, then so is g. Rolle's theorem guarantees that $g'(c) = f'(c) - s'(c) = 0$ for some value $c \in (a,b)$.

Thus, $f'(c) = s'(c) = \left.\dfrac{\Delta f}{\Delta x}\right|_{[a,b]}$. ■

10.2.3 Applications of the Mean Value Theorem

The Mean Value Theorem for derivatives allows us to know that the average rate of change of a differentiable function between any two points will be equal to the instantaneous rate of change at some point within the interval. Consequently, if we know properties of the derivative on entire intervals, that can provide information about how the function is changing on the interval. In particular, we learn that the sign of a derivative can be used to determine monotonicity of a function.

Theorem 10.2.11 Monotonicity of Differentiable Functions. *Suppose that f is a differentiable on an interval I (open or closed).*

- *If $f'(x) > 0$ for all $x \in I$, then $f(x)$ is increasing on I.*
- *If $f'(x) < 0$ for all $x \in I$, then $f(x)$ is decreasing on I.*
- *If $f'(x) = 0$ for all $x \in I$, then $f(x)$ is constant on I.*

If the interval I is open but f is continuous up to and including the end-points, then the conclusion can be extended to include the end-points as well.

Proof. Consider any two points $a, b \in I$ with $a < b$. Because f is differentiable on I, we know that f is continuous and differentiable on the subinterval $[a, b]$. The Mean Value Theorem guarantees the existence of a point $c \in (a, b)$ such that

$$f(b) - f(a) = f'(c) \cdot (b - a).$$

Now assume that $f'(x) > 0$ for all $x \in I$. Then $f'(c) > 0$ and $b - a > 0$, guaranteeing that $f(b) - f(a) > 0$. That is, $f(b) > f(a)$. This is what is needed to show that f is increasing on I.

Next assume that $f'(x) < 0$ for all $x \in I$. Then $f'(c) < 0$ while $b - a > 0$, guaranteeing that $f(b) - f(a) < 0$. That is, $f(b) < f(a)$, which shows that f is decreasing on I.

Finally assume that $f'(x) = 0$ for all $x \in I$. Then $f'(c) = 0$, implying that $f(b) - f(a) = 0$. That is, $f(b) = f(a)$, which shows that f is constant on I. ■

We can now justify doing the same sign analysis work using a derivative as we did for the rate of accumulation functions. What is different from then? Our previous justification required that the function could be written as an accumulation function with a known rate of accumulation. Now, we can do the same type of sign analysis with any function for which we can determine the derivative.

Because the second derivative gives the rate of change of the first derivative, we can use sign analysis of $f''(x)$ to describe concavity of $f(x)$.

Theorem 10.2.12 Concavity of Twice-Differentiable Functions. *Given a function f for which f' and f'' are defined on an interval I.*

- *If $f''(x) > 0$ for all $x \in I$, then $f(x)$ is concave up on I.*
- *If $f''(x) < 0$ for all $x \in I$, then $f(x)$ is concave down on I.*
- *If $f''(x) = 0$ for all $x \in I$, then $f(x)$ is linear on I.*

If the interval I is open but f' is continuous up to and including the end-points, then the conclusion can be extended to include the end-points as well.

Proof. This is just an application of Theorem Theorem 10.2.11 applied to f'. Once we know that f' is increasing on I, the definition of concavity allows us to say that f is concave up on I. Similarly, knowing that f' is decreasing is equivalent to saying that f is concave down. If f' is constant on an interval I, this is exactly what it means for f to be linear on I. ■

Example 10.2.13 Describe the monotonicity and concavity of $f(x) = xe^{-2x}$.

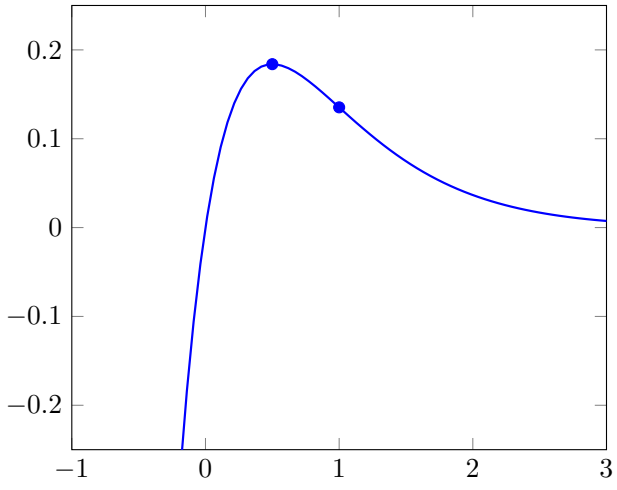
Solution. Start by computing the first and second derivatives. Note that we must use the product rule:

$$\begin{aligned} f(x) &= xe^{-2x}, \\ f'(x) &= 1 \cdot e^{-2x} + x \cdot -2e^{-2x} \\ &= (1-2x)e^{-2x}, \\ f''(x) &= -2 \cdot e^{-2x} + (1-2x) \cdot -2e^{-2x} \\ &= (-2-2+4x)e^{-2x} \\ &= (-4+4x)e^{-2x}. \end{aligned}$$

We can now do sign analysis for $f'(x)$ and $f''(x)$. Because e^{-2x} is a factor for each of the functions, we will use the fact that $e^{-2x} > 0$ for all values of x. The only point where $f'(x) = 0$ is where $1 - 2x = 0$ or $x = \frac{1}{2}$. The resulting sign analysis summary for $f'(x)$ is shown below.

The only point where $f''(x) = 0$ is where $-4 + 4x = 0$ or $x = 1$. The resulting sign analysis summary for $f''(x)$ is shown below.

We now interpret our results. Because f is continuous, we can extend open intervals to include end-points. The function $f(x)$ is increasing on $(-\infty, \frac{1}{2}]$ and decreasing on $[\frac{1}{2}, \infty)$. In addition, $f(x)$ is concave down on $(-\infty, 1]$ and concave up on $[1, \infty)$. A graph of $y = f(x)$ is shown below, with the local maximum at $x = \frac{1}{2}$ and the point of inflection at $x = 1$.



□

10.2.4 Classifying Antiderivatives

The Mean Value Theorem results in another important consequence: all antiderivatives of a particular function differ by constants. In particular, the result applies to intervals where the derivative is defined.

Theorem 10.2.14 *Suppose that $F(x)$ and $G(x)$ are any two differentiable functions defined on some interval I such that for all $x \in I$, $F'(x) = G'(x)$. Then there exists some constant C so that for all $x \in I$, $G(x) = F(x) + C$.*

In particular, if $F(x)$ and $G(x)$ are antiderivatives of the same function $f(x)$, and F and G are differentiable on an interval I, then $G(x) = F(x) + C$ on that interval.

Proof. Define a function $H(x) = G(x) - F(x)$. Because F and G are differentiable at all $x \in I$, $H(x)$ is both continuous and differentiable at all $x \in I$. With $F'(x) = G'(x)$, we have $H'(x) = 0$ for all $x \in I$. Consequently, by Theorem 10.2.11, $H(x)$ is constant on I, or $H(x) = C$ for some constant C. Therefore $G(x) - F(x) = C$ or $G(x) = F(x) + C$. ■

Be aware that the constant only applies to an interval where the antiderivatives are differentiable. The constant can be different over different intervals.

Example 10.2.15 We know that $F(x) = \ln(|x|)$ is an antiderivative of $f(x) = \dfrac{1}{x}$. Now, construct

$$G(x) = \begin{cases} \ln(-2x), & x < 0, \\ \ln(3x), & x > 0. \end{cases}$$

We can differentiate on each interval:

$$G'(x) = \begin{cases} \frac{d}{dx}\Big[\ln(-2x)\Big] = \frac{-2}{-2x} = \frac{1}{x}, & x < 0, \\ \frac{d}{dx}\Big[\ln(3x)\Big] = \frac{3}{3x} = \frac{1}{x}, & x > 0. \end{cases}$$

This shows that $F(x)$ and $G(x)$ are each antiderivatives of $f(x)$.

So what are the constants on the intervals? They can be found from the properties of logarithms:

$$\ln(-2x) = \ln(2) + \ln(-x),$$
$$\ln(3x) = \ln(3) + \ln(x).$$

We see that on the interval $(-\infty, 0)$, $G(x) = F(x) + \ln(2)$, but on the interval $(0, \infty)$, $G(x) = F(x) + \ln(3)$. □

10.2.5 Summary

1. Differentiability: We look for points in the domain of $f(x)$ where $f'(x)$ also exists. The function is nondifferentiable if $f'(x)$ does not exist.

   Examples of causes for nondifferentiability: $f(x)$ not being continuous, left- and right-slopes differ, or the tangent line is vertical.

2. Theorem 10.2.5: Local extremes of $f(x)$ can only occur at critical points, which are values where $f'(x) = 0$ or $f'(x)$ does not exist.

3. Theorem 10.2.7: For a continuous and differentiable function, it will have a horizontal tangent between any two zeros.

4. Theorem 10.2.10: For a continuous and differentiable function, the average rate of change on an interval will be matched by the slope of the tangent line at some intermediate point.

5. The Mean Value Theorem provides the justification of using sign analysis of $f'(x)$ and $f''(x)$ to determine intervals of monotonicity and concavity, respectively, for the function $f(x)$.

6. The Mean Value Theorem also provides the justification that any two antiderivatives of a function $f(x)$ can differ at most by a constant value over an interval on which they are differentiable.

10.2.6 Exercises

10.3 Definite Integrals and Antiderivatives

An accumulation function with rate $f(x)$ is a function A defined as the definite integral of $f(x)$ from a fixed lower limit c to the independent variable as the upper limit,

$$A(x) = \int_c^x f(z)\,dz.$$

The motivation for such a function was that the definite integral computes the total change of a quantity when the rate of change is given. By the splitting property of integration, if $A(x)$ is continuous then

$$\int_a^b f(x)\,dx = A(b) - A(a).$$

The Fundamental Theorem of Calculus showed that if $f(x)$ is continuous, then $A(x)$ is differentiable and $A'(x) = f(x)$. That is, $A(x)$ is an antiderivative of $f(x)$. This motivates another method for computing definite integrals.

In this section, we apply the Fundamental Theorem of Calculus to evaluate definite integrals using any convenient antiderivative. This application is called the Second Part of the Fundamental Theorem of Calculus. We demonstrate these calculations with several examples.

10.3.1 The Fundamental Theorem of Calculus

The Fundamental Theorem of Calculus shows that an accumulation function is an antiderivative of the integrand. The Mean Value Theorem implies that any other antiderivative will differ from the accumulation function as an antiderivative by some constant. The Second Part of the Fundamental Theorem of Calculus will then allow us to calculate definite integrals by calculating the change in any antiderivative.

Theorem 10.3.1 The Fundamental Theorem of Calculus, Part Two (FTC2). *Given any function $f(x)$ that is continuous on an interval I, let $F(x)$ be an antiderivative so that $F'(x) = f(x)$ for all $x \in I$. Then for values $a, b \in I$,*

$$\int_a^b f(x)\,dx = \big[F(x)\big]_a^b = F(b) - F(a).$$

Proof. Because $f(x)$ is continuous, we can define an accumulation function

$$A(x) = \int_a^x f(z)\,dz.$$

Because $A(x)$ and $F(x)$ are both antiderivatives, with $A'(x) = F'(x) = f(x)$ for all $x \in I$, there is some constant C so that $A(x) = F(x) + C$. Because $A(a) = 0$, we have $F(a) + C = 0$ so that $C = -F(a)$. By the splitting property of integrals, we have

$$\int_a^b f(x)\,dx = A(b) = F(b) + C = F(b) - F(a).$$

■

When evaluating definite integrals using the Fundamental Theorem of Calculus, we are *substituting* the evaluation of a definite integral, which is defined as the limit of a Riemann sum, with the change in an antiderivative. To indicate such a substitution, we should refer to the Fundamental Theorem of

Calculus, perhaps using the abbreviation FTC. We can simultaneously make the substitution and note the formula for the antiderivative by using bracket evaluation notation.

Given any function $F(x)$, the notation $\big[F(x)\big]_a^b$ means to evaluate the change of the expression when x goes from a to b:

$$\big[F(x)\big]_a^b = F(b) - F(a).$$

When the formula for $F(x)$ is simple, the brackets can be dropped and replaced by a vertical bar on the right,

$$F(x)\big|_a^b = F(b) - F(a).$$

Example 10.3.2 Evaluate $\displaystyle\int_1^4 x^2\,dx$.

Solution. Let us consider the two steps separately. Then we will see how to represent this more compactly using evaluation notation.

First, we need an antiderivative, computed as an indefinite integral

$$\int x^2\,dx = \frac{1}{3}x^3 + C.$$

This tells us that $F(x) = \frac{1}{3}x^3$ is an antiderivative, as is $F(x) = \frac{1}{3}x^3 + 4$ (or any other constant). We'll use the first, since it is simpler; we only need one antiderivative, not all of them.

Second, the Fundamental Theorem of Calculus allows us to evaluate the definite integral as the change in $F(x)$.

$$\begin{aligned}\int_1^4 x^2\,dx &= F(4) - F(1)\\ &= \frac{1}{3}(4)^3 - \frac{1}{3}(1)^3\\ &= \frac{64}{3} - \frac{1}{3}\\ &= \frac{63}{3} = 21\end{aligned}$$

This can be written more compactly by writing the formula of the antiderivative inside the evaluation notation while simultaneously indicating the use of the Fundamental Theorem of Calculus:

$$\begin{aligned}\int_1^4 x^2\,dx &\overset{\text{FTC}}{=} \big[\tfrac{1}{3}x^3\big]_1^4\\ &= \frac{1}{3}(4)^3 - \frac{1}{3}(1)^3\\ &= \frac{64}{3} - \frac{1}{3} = 21\end{aligned}$$

Notice that we did not need the constant of integration because the Fundamental Theorem of Calculus only requires one antiderivative. We generally choose the most convenient one with a zero constant. □

Evaluation of definite integrals involves recognizing antiderivatives and then evaluating their change.

Example 10.3.3

$$\begin{aligned}\int_1^2 e^{3x}\,dx &\overset{\text{FTC}}{=} \left[\frac{1}{3}e^{3x}\right]_1^2\\ &= \frac{1}{3}e^6 - \frac{1}{3}e^3\\ &= \frac{e^6 - e^3}{3}\end{aligned}$$

□

Example 10.3.4 Find $\displaystyle\int_1^3 x(3x^2+1)^4\,dx$.

Solution. The integrand has a product of x with u^4 where $u = 3x^2+1$. For the chain rule to have been the source of this product, we would need $\frac{du}{dx} = 6x$ rather than x.

$$\begin{aligned}\int_1^3 x(3x^2+1)^4\,dx &= \int_1^3 \frac{1}{6}(6x)(3x^2+1)^4\,dx\\ &\overset{\text{FTC}}{=} \left[\frac{1}{6}\cdot\frac{1}{5}u^5\right]_{x=1}^{x=3}\\ &= \left[\frac{1}{30}(3x^2+1)^5\right]_1^3\\ &= \frac{1}{30}(3(3^2)+1)^5 - \frac{1}{30}(3(1^2)+1)^5\\ &= \frac{28^5}{30} - \frac{4^5}{30}\\ &= \frac{17,210,368}{30} - \frac{1,024}{30} = \frac{17,209,344}{30} = 573,644.8\end{aligned}$$

□

We have to be careful about satisfying the hypotheses. For example, if $f(x)$ is not continuous over the interval of integration, we can not use antiderivatives to calculate the definite integral.

Example 10.3.5 Find $\displaystyle\int \frac{1}{2x-1}\,dx$. How can we use that result for the following definite integrals?

1. $\displaystyle\int_0^2 \frac{1}{2x-1}\,dx$
2. $\displaystyle\int_1^3 \frac{1}{2x-1}\,dx$

Solution. The integrand is of the form u^{-1} where $u = 2x-1$. The derivative of u is $u' = 2$. In order to antidifferentiate the chain rule, we rewrite

$$\int \frac{1}{2x-1}\,dx = \int \frac{1}{2}\frac{2}{2x-1}\,dx = \frac{1}{2}\ln(|2x-1|) + C.$$

To see if the antiderivative can be used in a definite integral, we need to see where $f(x) = \frac{1}{2x-1}$ is continuous. A discontinuity occurs at $x = \frac{1}{2}$. Consequently, a definite integral using the antiderivative can only be used for intervals that do not include $\frac{1}{2}$. Thus, $\displaystyle\int_0^2 \frac{1}{2x-1}\,dx$ can not be computed. On

the other hand,

$$\begin{aligned}\int_1^3 \frac{1}{2x-1}\,dx &\overset{\text{FTC}}{=} \Big[\ln(|2x-1|)\Big]_1^3 \\ &= \ln(|2(3)-1|) - \ln(|2(1)-1|) \\ &= \ln(5) - \ln(1) = \ln(5).\end{aligned}$$

□

10.3.2 Composition with Accumulation Functions

The First Part of the Fundamental Theorem of Calculus tells us the derivative of accumulation functions. Knowing the chain rule allows us to compute derivatives of functions defined by integrals with expressions in the limits of integration.

Example 10.3.6 Compute the following derivatives.

1. $\dfrac{d}{dx}\displaystyle\int_1^x e^{-z^2}\,dz$
2. $\dfrac{d}{dx}\displaystyle\int_x^{x^2} e^{-z^2}\,dz$
3. $\dfrac{d}{dx}\displaystyle\int_0^{\sqrt{x}} \frac{1}{\sqrt{z^4+1}}\,dz$

Solution. When solving problems involving definite integrals, it is often helpful to explicitly remind yourself of the concept of accumulation functions and the fundamental theorem of calculus's conclusion.

1. Define $A(x) = \displaystyle\int_1^x e^{-z^2}\,dz$, which is the accumulation function with integrand $f(z) = e^{-z^2}$. The Fundamental Theorem of Calculus tells us that $A'(x) = f(x) = e^{-x^2}$. The following work would communicate these results:

$$\begin{aligned}A(x) &= \int_1^x e^{-z^2}\,dz \\ A'(x) &\overset{\text{FTC}}{=} e^{-x^2} \\ \frac{d}{dx}\int_1^x e^{-z^2}\,dz &= A'(x) = e^{-x^2}\end{aligned}$$

2. To compute $\dfrac{d}{dx}\displaystyle\int_x^{x^2} e^{-z^2}\,dz$, we first need the accumulation function $A(x) = \displaystyle\int_1^x e^{-z^2}\,dz$ with rate $f(x) = e^{-x^2}$. The integral that defines our function involves a composition by the splitting property,

$$\int_x^{x^2} e^{-z^2}\,dz = A(x^2) - A(x).$$

When we differentiate this, we must use the chain rule knowing

$$\frac{d}{dx}\Big[A(u)\Big] = A'(u)\frac{du}{dx} \overset{\text{FTC}}{=} f(u)\frac{du}{dx}.$$

The following work would communicate these results:

$$\begin{aligned} A(x) &= \int_1^x e^{-z^2}\,dz \\ \frac{d}{dx}\int_x^{x^2} e^{-z^2}\,dz &= \frac{d}{dx}\Big[A(x^2) - A(x)\Big] \\ &\overset{\text{FTC}}{=} f(x^2)\cdot\frac{d}{dx}[x^2] - f(x) \\ &= e^{-(x^2)^2}\cdot 2x - e^{-x^2} \\ &= 2xe^{-x^4} - e^{-x^2} \end{aligned}$$

3. To compute $\dfrac{d}{dx}\left[\displaystyle\int_0^{\sqrt{x}} \frac{1}{\sqrt{z^4+1}}\,dz\right]$, we will need to define an accumulation function and then apply the Fundamental Theorem of Calculus to find the derivative required.

$$\begin{aligned} A(x) &= \int_0^x \frac{1}{\sqrt{z^4+1}}\,dz \\ A'(x) &\overset{\text{FTC}}{=} \frac{1}{\sqrt{x^4+1}} \\ \frac{d}{dx}\left[\int_0^{\sqrt{x}} \frac{1}{\sqrt{z^4+1}}\,dz\right] &= \frac{d}{dx}[A(\sqrt{x})] \\ &= A'(\sqrt{x})\frac{d}{dx}[\sqrt{x}] \\ &= \frac{1}{\sqrt{(\sqrt{x})^4+1}}\cdot(\frac{1}{2}x^{-1/2}) \\ &= \frac{1}{\sqrt{x^2+1}}\cdot\frac{1}{2\sqrt{x}} \\ &= \frac{1}{2\sqrt{x(x^2+1)}} \end{aligned}$$

□

10.3.3 Summary

1. The Fundamental Theorem of Calculus, Part 1, together with the Mean Value Theorem, imply that for any continuous function $f(x)$, an accumulation function and any other antiderivative will differ by a constant value.

2. The Fundamental Theorem of Calculus, Part 2, states that the definite integral of a function that is continuous on the interval of integration can be substituted for the change in *any* antiderivative of the rate. That is, if $F(x)$ is an antiderivative of $f(x)$ and $f(x)$ is continuous on the interval containing a and b,

$$\int_a^b f(x)\,dx \overset{\text{FTC}}{=} \Big[F(x)\Big]_a^b = F(b) - F(a).$$

3. The Fundamental Theorem of Calculus, Part 1, together with the chain rule, allows us to compute the derivative of functions where the limits

of integration are expressions involving the independent variable. Let u and w be expressions involving x and suppose that $f(x)$ is a continuous function.

$$\frac{d}{dx}\Big[\int_u^w f(z)\,dz\Big] \overset{\text{FTC}}{=} f(w)\frac{dw}{dx} - f(u)\frac{du}{dx}$$

10.3.4 Exercises

10.4 L'Hôpital's Rule

Limits involving combinations of continuous functions are generally computed by evaluating the values of the functions. Arithmetic involving infinity will follow elementary rules. Suppose $L > 0$ is a positive number. Then when performing arithmetic, the following rules will hold.

$$\begin{gathered}\infty \pm L = \infty \\ \infty + \infty = \infty \\ -\infty + -\infty = -\infty \\ L \cdot \infty = \infty \\ -L \cdot \infty = -\infty \\ \infty \cdot \infty = \infty \\ \infty \cdot -\infty = -\infty \\ -\infty \cdot -\infty = \infty\end{gathered}$$

However, when calculations would result that attempt to cancel away an infinite quantity (or zero), the limit has an indeterminate form. That is, calculations involving any of the following arithmetic are in an indeterminate form,

$$\frac{\infty}{\infty} \qquad \infty - \infty \qquad \frac{0}{0} \qquad 0 \cdot \infty.$$

The general strategy for evaluating indeterminate limits involves rewriting the limit in a different form, or finding another limit that is known to be equivalent.

L'Hôpital's rule is a theorem that allows us to rewrite a limit which has an indeterminate form $\frac{0}{0}$ or $\frac{\infty}{\infty}$ using derivatives.

Theorem 10.4.1 L'Hôpital's Rule. *Suppose $f(x)$ and $g(x)$ are functions so that*

$$\lim_{x\to a} \frac{f(x)}{g(x)} \to \frac{0}{0} \quad \textit{or} \quad \lim_{x\to a} \frac{f(x)}{g(x)} \to \frac{\infty}{\infty}.$$

If $\lim\limits_{x\to a} \dfrac{f'(x)}{g'(x)}$ exists, then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

Proof. Consider the case where $f(x) \to 0$ and $g(x) \to 0$ and $g(x) \neq 0$ as $x \to a$. If f or g is not continuous at $x = a$, consider the continuous extensions so that $f(a) = 0$ and $g(a) = 0$. Similar to the proof of the Mean Value Theorem, define $h(z) = f(z) - \frac{f(x)}{g(x)} \cdot g(z)$, treating x as constant. Note that $h(z)$ is differentiable and therefore continuous because $f'(x)$ and $g'(x)$ must both exist for the limit of the quotient to exist. In addition, $h(a) = 0$ and $h(x) = 0$.

Rolle's theorem implies that $h'(z) = 0$ for some z between a and x,

$$h'(z) = f'(z) - \frac{f(x)}{g(x)} g'(z) = 0,$$

so that

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

Because z is between a and x, as $x \to a$ we must have $z \to a$.

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{z\to a} \frac{f'(z)}{g'(z)}.$$

■

It is essential that you verify that the hypotheses of L'Hôpital's rule apply before replacing the expressions with their derivatives. Limits and derivatives are not the same. Computing a limit does not always mean we are computing a derivative.

In addition, make note that changing the limit using L'Hôpital's rule is not computing the derivative of a quotient. It is replacing the limit of a quotient with the limit of a new quotient involving two derivatives.

Our first few examples will illustrate that L'Hôpital's rule gives the same result as methods we learned earlier involving factoring.

Example 10.4.2 Evaluate the limit $\lim_{x\to 2} \dfrac{x^2-3x+2}{x^2+x-6}$ using factoring and using L'Hôpital's rule.

Solution. If we try to use the value directly, we find an indeterminate form,

$$\lim_{x\to 2} \frac{x^2-3x+2}{x^2+x-6} \to \frac{2^2-3(2)+2}{2^2+2-6} = \frac{0}{0}.$$

We must find alternative representations of this limit to determine its value.

Using factoring, we rewrite the limit by canceling common factors.

$$\begin{aligned}\lim_{x\to 2} \frac{x^2-3x+2}{x^2+x-6} &= \lim_{x\to 2} \frac{(x-2)(x-1)}{(x+3)(x-2)} \\ &= \lim_{x\to 2} \frac{x-1}{x+3} = \frac{2-1}{2+3} = \frac{1}{5}\end{aligned}$$

Because the original limit had indeterminate form $\frac{0}{0}$, L'Hôpital's rule will apply with $f(x) = x^2-3x+2$ and $g(x) = x^2+x-6$. We replace the limit of $f(x)/g(x)$ with the limit of $f'(x)/g'(x)$, assuming that limit exists.

$$\lim_{x\to 2} \frac{x^2-3x+2}{x^2+x-6} = \lim_{x\to 2} \frac{2x-3}{2x+1} = \frac{2(2)-3}{2(2)+1} = \frac{1}{5}$$

We see that both approaches give the same limit, exactly as predicted by L'Hôpital's rule. □

Sometimes, we need to apply L'Hôpital's rule more than once when the modified limit is still in indeterminate form.

Example 10.4.3 Evaluate the limit $\lim_{x\to\infty} \dfrac{2x^2+x-3}{x^2-x-5}$ using factoring and using L'Hôpital's rule.

Solution. Limits at infinity generally require factoring out the fastest growing terms.

$$\lim_{x\to\infty} \frac{2x^2+x-3}{x^2-x-5} = \lim_{x\to\infty} \frac{x^2(2+\frac{1}{x}-\frac{3}{x^2})}{x^2(1-\frac{1}{x}-\frac{5}{x^2})}$$

In this form, we see that the limit has form $\frac{\infty}{\infty}$. By canceling the common factor x^2, we find

$$\lim_{x\to\infty} \frac{2x^2+x-3}{x^2-x-5} = \lim_{x\to\infty} \frac{2+\frac{1}{x}-\frac{3}{x^2}}{1-\frac{1}{x}-\frac{5}{x^2}} = \frac{2+0-0}{1-0-0} = 2.$$

Because the limit had form $\frac{\infty}{\infty}$, we can use L'Hôpital's rule to rewrite the limit in a new form,

$$\lim_{x\to\infty} \frac{2x^2+x-3}{x^2-x-5} = \lim_{x\to\infty} \frac{4x+1}{2x-1}.$$

This limit is still of form $\frac{\infty}{\infty}$, so we can use L'Hôpital's rule again to get yet another equivalent form,

$$\lim_{x\to\infty} \frac{2x^2+x-3}{x^2-x-5} = \lim_{x\to\infty} \frac{4}{2} = 2.$$

Again, either approach will give the same value. □

One of the advantages of L'Hôpital's rule is that it allows us to evaluate limits where factoring does not help.

Example 10.4.4 Compute $\lim_{x\to 3} \frac{2^x-8}{x^2+x-12}$.

Solution. The first step is always to try evaluating directly.

$$\lim_{x\to 3} \frac{2^x-8}{x^2+x-12} \to \frac{2^3-8}{3^2+3-12} = \frac{0}{0}$$

The limit has indeterminate form $\frac{0}{0}$ so that we can use L'Hôpital's rule. In this case, note that the numerator 2^x-8 does not factor. L'Hôpital's rule is the preferred approach.

A typical solution would be written as follows.

$$\begin{aligned}\lim_{x\to 3} \frac{2^x-8}{x^2+x-12} &\to \frac{2^3-8}{3^2+3-12} = \frac{0}{0}\\ \overset{\text{L'H}}{=} \lim_{x\to 3} \frac{2^x\ln(2)}{2x+1} &= \frac{2^3\ln(2)}{2(3)+1} = \frac{8\ln(2)}{7}\end{aligned}$$

The first line shows that the original limit is an indeterminate form. Writing "L'H" over the equal sign shows that we are using L'Hôpital's rule to replace the original limit with the modified limit involving derivatives. Also, we used the derivative of an exponential, $\frac{d}{dx}[b^x] = b^x\ln(b)$. □

Indeterminate limits that are not fractions of the form $\frac{0}{0}$ or $\frac{\infty}{\infty}$ do not directly apply L'Hôpital's rule. You must first use algebra to rewrite them in a way that they do have the appropriate form.

Example 10.4.5 Evaluate $\lim_{x\to 0^+} x\ln(x)$.

Solution. When $x\to 0^+$, we have $ln(x)\to -\infty$. As written, the limit has the indeterminate form $\lim_{x\to 0^+} x\ln(x) \to 0\cdot -\infty$. This is indeterminate because multiplying ∞ by zero would be a form of trying to cancel the infinite. Instead, we need to rewrite the formula so that it is a fraction.

There are two approaches:

$$x\ln(x) = \frac{\ln(x)}{x^{-1}} = \frac{x}{(\ln(x))^{-1}}.$$

When choosing which approach will be better, you should ask yourself which formula will lead to simpler derivatives. For this problem, we use the first expression, knowing that $x^{-1}\to+\infty$ as $x\to 0^+$.

$$\begin{aligned}\lim_{x\to 0^+} x\ln(x) = \lim_{x\to 0^+} \frac{\ln(x)}{x^{-1}} &\to \frac{-\infty}{\infty}\\ \overset{\text{L'H}}{=} \lim_{x\to 0^+} \frac{x^{-1}}{-1x^{-2}} &= \lim_{x\to 0^+} \frac{x^{-1}x^2}{-x^{-2}x^2}\\ = \lim_{x\to 0^+} -x &= 0.\end{aligned}$$

□

Example 10.4.6 Evaluate $\lim_{x\to\infty} x^2 e^{-3x}$.

Solution. We know limit values for the exponential: $e^{\infty} = \infty$ and $e^{-\infty} = 0$. The given limit will be an indeterminate form $\infty \cdot 0$. So we rewrite it in quotient form and then use L'Hôpital's rule.

$$\begin{aligned}\lim_{x\to\infty} x^2 e^{-3x} &= \lim_{x\to\infty} \frac{x^2}{e^{3x}} \to \frac{\infty}{\infty} \\ &\overset{\text{L'H}}{=} \lim_{x\to\infty} \frac{2x}{3e^{3x}} \to \frac{\infty}{\infty} \\ &\overset{\text{L'H}}{=} \lim_{x\to\infty} \frac{2}{9e^{3x}} \to \frac{2}{\infty} = 0\end{aligned}$$

Notice that we used L'Hôpital's rule twice when the first time resulted in another indeterminate form. □

We end with an example involving powers. When both the base and the exponent are variable, we must interpret a power in terms of composition with the exponential function,

$$u^v = \exp(\ln(u^v)) = \exp(v\ln(u)) = e^{v\ln(u)}.$$

Because the natural exponential function is continuous, we only need to evaluate the limit of $v\ln(u)$ and then evaluate the exponential function at the corresponding limit. This is a consequence of Theorem Theorem 5.3.22.

Example 10.4.7 Evaluate $\lim_{x\to\infty} (1 + \frac{r}{x})^{xt}$, where r and t are constant values.

Solution. The function for which we compute a limit can be rewritten as a composition with the natural exponential function:

$$\begin{aligned}f(x) &= (1 + \frac{r}{x})^{xt} \\ &= \exp\left(\ln\left((1 + rx^{-1})^{xt}\right)\right) \\ &= \exp\left(xt\ln(1 + rx^{-1})\right).\end{aligned}$$

So we start by evaluating the limit of the expression inside the exponential.

$$\lim_{x\to\infty} xt\ln(1 + rx^{-1}) \to \infty \cdot \ln(1 + 0) = \infty \cdot 0$$

This limit has an indeterminate form.

We rewrite the indeterminate limit as a fraction so that we can use L'Hôpital's rule. From our earlier experience, we will leave the logarithm in the numerator.

$$\begin{aligned}\lim_{x\to\infty} xt\ln(1 + rx^{-1}) &= \lim_{x\to\infty} \frac{t\ln(1 + rx^{-1})}{x^{-1}} \to \frac{t\cdot\ln(1)}{0} = \frac{0}{0} \\ &\overset{\text{L'H}}{=} \lim_{x\to\infty} \frac{t\cdot\frac{1}{1+rx^{-1}}\cdot\frac{d}{dx}[1 + rx^{-1}]}{-x^{-2}} \\ &= \lim_{x\to\infty} \frac{t\cdot\frac{1}{1+rx^{-1}}\cdot(-rx^{-2})}{-x^{-2}} \\ &= \lim_{x\to\infty} t\cdot\frac{1}{1 + rx^{-1}}\cdot(r) \to t\cdot\frac{1}{1+0}\cdot r = rt\end{aligned}$$

Using the Limit of a Continuous Composition, we conclude

$$\begin{aligned}\lim_{x\to\infty} \exp\left(xt\ln(1 + rx^{-1})\right) &= \exp\left(\lim_{x\to\infty} xt\ln(1 + rx^{-1})\right) \\ &= \exp(rt) = e^{rt}.\end{aligned}$$

□

Chapter 11

Calculus for Trigonometry

11.1 The Derivatives of Trigonometric Functions

11.1.1 Essential Trigonometric Identities

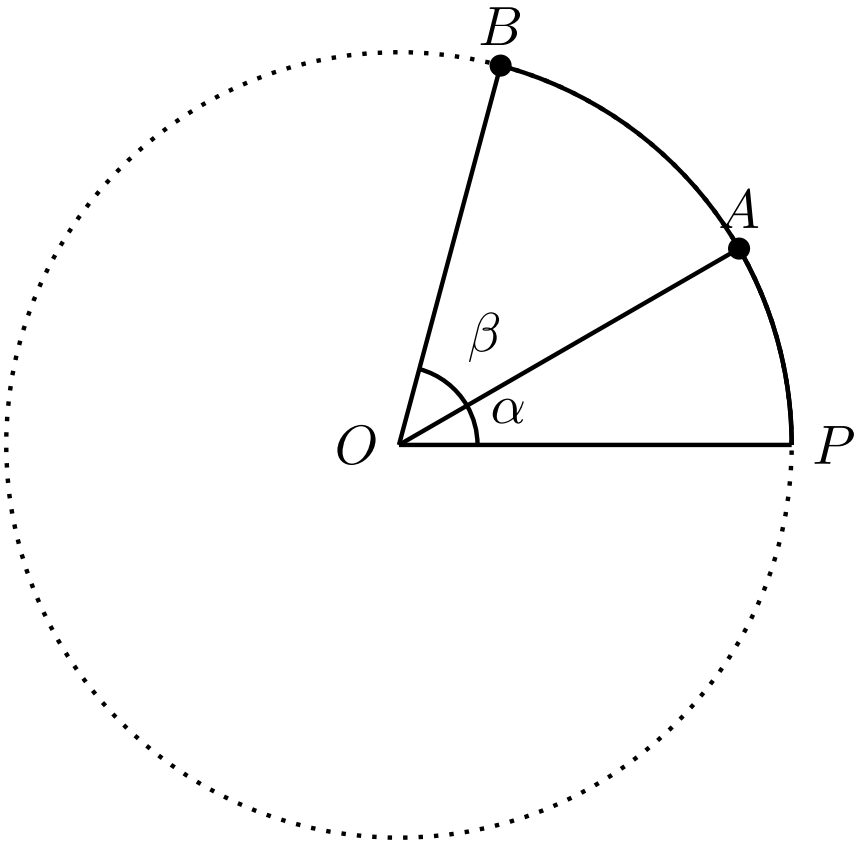
When we found the derivative of elementary exponential functions, we found that we needed to use a rule to rewrite $b^{x+h} = b^x \cdot b^h$. This type of rule is called an identity. Identities provide rules to rewrite a formula in another form without changing the value of the formula. Trigonometric functions are all defined in terms of the elementary sine and cosine functions. Consequently, we need the basic identities of sine and cosine.

We start with the sum identities.

Theorem 11.1.1 Sum Identities of Sine and Cosine. *Given any* $\alpha, \beta \in \mathbb{R}$,

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta),$$
$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta).$$

Proof. The following geometric proof is valid for acute angles, $0 < \alpha, \beta < \frac{\pi}{2}$. Consider the location of the angle α and $\alpha + \beta$ on the unit circle, illustrated as points A and B, respectively. The origin at $(0, 0)$ will be the point O and the point $(1, 0)$ will be the point P. Construct a line segment from B to intersect $\overline{OA}$ at a right angle; call the point of intersection C. Draw a vertical line from C which intersects OP at a point Q. Finally draw a horizontal line through C and a vertical line through B, which intersect at a point D.



By construction, we know $m\angle POA = \alpha$ and $m\angle AOB = \beta$. Because $OB = 1$, we know that $OC = \cos(\beta)$ and $BC = \sin(\beta)$. By geometry, we can prove that triangle BDC is a right triangle with $m\angle DBC = \alpha$. Consequently,

$$BD = BC\cos(\alpha) = \cos(\alpha)\sin(\beta),$$

$$DC = BC\sin(\alpha) = \sin(\alpha)\sin(\beta).$$

Similarly, triangle OQC is a right triangle with $m\angle COQ = \alpha$. Since $OC = \cos(\beta)$, we know

$$OQ = OC\cos(\alpha) = \cos(\alpha)\cos(\beta),$$
$$CQ = OC\sin(\alpha) = \sin(\alpha)\cos(\beta).$$

The x-coordinate of B is $\cos(\alpha+\beta)$ so that

$$\cos(\alpha+\beta) = OQ - CD = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta).$$

The y-coordinate of B is $\sin(\alpha+\beta)$ so that

$$\sin(\alpha+\beta) = CQ + DB = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta).$$

■

Next, we state the symmetries of sine and cosine.

Theorem 11.1.2 Sum Identities of Sine and Cosine. *Sine is an odd function. Cosine is an even function. That is, for any $\alpha \in \mathbb{R}$,*

$$\sin(-\alpha) = -\sin(\alpha),$$
$$\cos(-\alpha) = \cos(\alpha).$$

Proof. An angle $-\alpha$ goes in the opposite direction as the angle α. Consequently, the points on the unit circle have the same horizontal coordinate,

$$\cos(-\alpha) = \cos(\alpha),$$

and opposite vertical coordinates,

$$\sin(-\alpha) = -\sin(\alpha).$$

■

Finally, because the sine and cosine are defined on a unit circle (with radius 1), we have a Pythagorean identity regarding the sum of the squared values.

Theorem 11.1.3 The Pythagorean Identity. *For any $\alpha \in \mathbb{R}$,*

$$\sin^2(\alpha) + \cos^2(\alpha) = 1,$$
$$\tan^2(\alpha) + 1 = \sec^2(\alpha),$$
$$1 + \cot^2(\alpha) = \csc^2(\alpha).$$

Proof. By definition, the point $(x, y) = (\cos(\alpha), \sin(\alpha))$ is on the unit circle $x^2 + y^2 = 1$. By substitution, $x = \cos(\alpha)$ and $y = \sin(\alpha)$, we get the identity $\cos^2(\alpha) + \sin^2(\alpha) = 1$. If we divide both sides of the equation by $\cos^2(\alpha)$, we obtain

$$\frac{\cos^2(\alpha)}{\cos^2(\alpha)} + \frac{\sin^2(\alpha)}{\cos^2(\alpha)} = \frac{1}{\cos^2(\alpha)} \quad \Leftrightarrow \quad 1 + \tan^2(\alpha) = \sec^2(\alpha).$$

The last identity comes from dividing both sides of the equation by $\sin^2(\alpha)$. ■

11.1.2 Differentiation of Sine and Cosine

We start by computing the derivatives for sine and cosine at the origin, $\sin'(0)$ and $\cos'(0)$. Once we know those values, we will be able to find the derivatives

as functions.

Theorem 11.1.4

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

Proof. By definition,

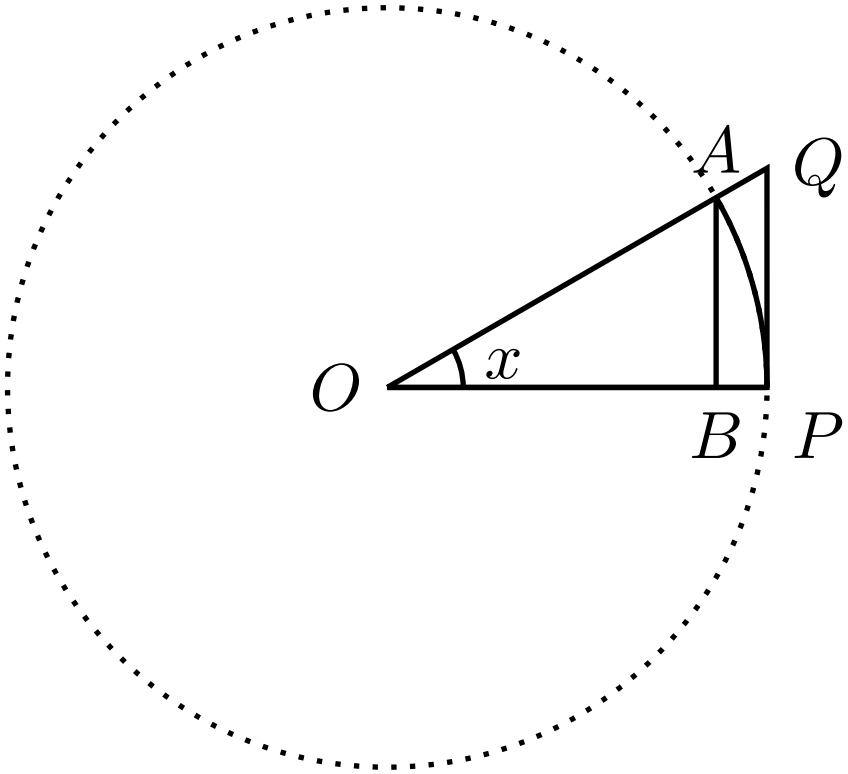
$$\sin'(0) = \lim_{h\to 0} \frac{\sin(0+h) - \sin(0)}{h}.$$

Since $\sin(0) = 0$, we can rewrite

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

since the variable name does not affect the value of the limit.

Consider the figure below with angle $x > 0$. The point A is on the unit circle and has coordinates $(\cos(x), \sin(x))$. Consequently, triangle OBA has area $\frac{1}{2}\sin(x)\cos(x)$. The point Q has coordinates $(1, \tan(x))$, so triangle OPQ has area $\frac{1}{2}\tan(x)$. If we consider the sector of the circle OAP, it has an area that is the corresponding fraction $\frac{x}{2\pi}$ of the area of the unit circle, π, which has the value $\frac{x}{2}$.



Comparing the areas leads to the inequality,

$$\frac{1}{2}\sin(x)\cos(x) < \frac{1}{2}x < \frac{1}{2}\tan(x).$$

Multiplying by 2 and dividing by $\sin(x)$, we have another inequality

$$\cos(x) < \frac{x}{\sin(x)} < \frac{1}{\cos(x)}.$$

Since $\cos(x)$ is continuous at $x = 0$, we know

$$\lim_{x\to 0} \cos(x) = \cos(0) = 1,$$

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

Using the Limit Squeeze Theorem and the Limit of a Reciprocal, we then know

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

■

It is important to note that getting the value for this limit to be the value 1 was a consequence of measuring angles in radians. For any other way that we might measure angles, the fraction of the circles area would be a ratio of the value x to the measurement of the angle to complete a full circle. For example, if we measured angles in degrees, we would have instead found $\sin'_{\text{deg}}(0) = \frac{2\pi}{360}$. Mathematically, one justification for measuring angles in radians is simply in order to guarantee that this $\sin'(0) = 1$.

Theorem 11.1.5

$$\cos'(0) = \lim_{x\to 0}\frac{\cos(x)-1}{x} = 0$$

Proof. By definition,

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

Since $\cos(0) = 1$, we can rewrite

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

Multiplying the numerator and denominator by $\cos(x)+1$, we find

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

By the Pythagorean identity, we know $\cos^2(x) - 1 = -\sin^2(x)$ so that we can rewrite

$$\cos'(0) = \lim_{x\to 0}\frac{\sin(x)}{x}\frac{-\sin(x)}{\cos(x)+1}.$$

Since $\lim_{x\to 0}\frac{\sin(x)}{x} = \sin'(0) = 1$ and

$$\lim_{x\to 0}\frac{-\sin(x)}{\cos(x)+1} = \frac{-\sin(0)}{\cos(0)+1} = \frac{0}{2} = 0,$$

the limit rule for a product guarantees $\cos'(0) = 0$. ■

Knowing the instantaneous rates of change of sine and cosine at $x = 0$ allows us to compute the derivative at any input value. The proofs for these differentiation rules rely on the sum identities for trigonometric functions.

Theorem 11.1.6

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

Proof. Using the definition of the derivative, we write

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

The sum identity for sine allows us to rewrite $\sin(x+h) = \sin(x)\cos(h) +$

$\cos(x)\sin(h)$, so that the derivative can be rewritten

$$\begin{aligned}\sin'(x) &= \lim_{h\to 0}\frac{\sin(x)\cos(h)+\cos(x)\sin(h)-\sin(x)}{h}\\ &= \lim_{h\to 0}\frac{\sin(x)\cdot(\cos(h)-1)+\cos(x)\cdot\sin(h)}{h}\\ &= \lim_{h\to 0}\sin(x)\cdot\frac{\cos(h)-1}{h}+\cos(x)\frac{\sin(h)}{h}.\end{aligned}$$

Because $\sin(x)$ and $\cos(x)$ do not depend on h, they play the role of a constant multiple. By the rules for a limit of a sum and the limit of a constant multiple, we can write

$$\sin'(x) = \sin(x)\cos'(0)+\cos(x)\sin'(0) = 0\cdot\sin(x)+1\cdot\cos(x) = \cos(x).$$

■

Theorem 11.1.7

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

Proof. Using the definition of the derivative, we write

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

The sum identity for cosine allows us to rewrite $\cos(x+h) = \cos(x)\cos(h) - \sin(x)\sin(h)$, so that the derivative can be rewritten

$$\begin{aligned}\cos'(x) &= \lim_{h\to 0}\frac{\cos(x)\cos(h)-\sin(x)\sin(h)-\cos(x)}{h}\\ &= \lim_{h\to 0}\frac{\cos(x)\cdot(\cos(h)-1)-\sin(x)\cdot\sin(h)}{h}\\ &= \lim_{h\to 0}\cos(x)\cdot\frac{\cos(h)-1}{h}-\sin(x)\frac{\sin(h)}{h}.\end{aligned}$$

We can therefore write

$$\cos'(x) = \cos(x)\cos'(0)-\sin(x)\sin'(0) = 0\cdot\cos(x)-1\cdot\sin(x) = -\sin(x).$$

■

11.1.3 Derivatives of Other Trigonometric Functions

All other trigonometric functions are defined in terms of the sine and cosine functions. Knowing the derivatives of sine and cosine allow us to compute the derivative rules for each of the other trigonometric functions.

Theorem 11.1.8

$$\begin{aligned}\frac{d}{dx}[\tan(x)] &= \tan'(x) = \sec^2(x)\\ \frac{d}{dx}[\sec(x)] &= \sec'(x) = \sec(x)\tan(x)\\ \frac{d}{dx}[\cot(x)] &= \cot'(x) = -\csc^2(x)\\ \frac{d}{dx}[\csc(x)] &= \csc'(x) = -\csc(x)\cot(x)\end{aligned}$$

Proof. The proofs for these rules are all based on the definitions of these trigonometric functions in terms of sine and cosine.

$$\tan(x) = \frac{\sin(x)}{\cos(x)} \qquad \sec(x) = \frac{1}{\cos(x)}$$
$$\cot(x) = \frac{\cos(x)}{\sin(x)} \qquad \csc(x) = \frac{1}{\sin(x)}$$

We will look at the derivatives of $\tan(x)$ and $\sec(x)$ and leave the other two proofs to the reader.

Because $\tan(x)$ is defined as a quotient, we compute its derivative using the quotient rule.

$$\begin{aligned}\frac{d}{dx}[\tan(x)] &= \frac{d}{dx}[\frac{\sin(x)}{\cos(x)}] \\ &= \frac{\cos(x)\sin'(x) - \sin(x)\cos'(x)}{\cos^2(x)} \\ &= \frac{\cos(x)\cdot\cos(x) - \sin(x)\cdot(-\sin(x))}{\cos^2(x)} \\ &= \frac{\cos^2(x)+\sin^2(x)}{\cos^2(x)} \\ &= \frac{1}{\cos^2(x)} = \sec^2(x)\end{aligned}$$

Similarly, $\sec(x)$ is defined as a reciprocal, so we use the reciprocal rule of derivatives.

$$\begin{aligned}\frac{d}{dx}[\sec(x)] &= \frac{d}{dx}[\frac{1}{\cos(x)}] \\ &= \frac{-\cos'(x)}{\cos^2(x)} = \frac{-(-\sin(x))}{\cos^2(x)} = \frac{\sin(x)}{\cos^2(x)} \\ &= \frac{1}{\cos(x)}\cdot\frac{\sin(x)}{\cos(x)} = \sec(x)\tan(x)\end{aligned}$$

■

11.1.4 Practice with Derivatives

When we take into account the chain rule, we have the following general derivative rules for trigonometric functions. Notice that cosine, cotangent, and cosecant all have a negative sign. Also note the similarity in formulas between the derivatives for sine and cosine, for tangent and cotangent, and for secant and cosecant. There are really only three differentiation rules, each with a complementary rule for the complementary functions.

General Derivative Rules for Trigonometric Functions.

Let u represent any expression that depends on x.

$$\frac{d}{dx}[\sin(u)] = \cos(u)\frac{du}{dx}$$
$$\frac{d}{dx}[\cos(u)] = -\sin(u)\frac{du}{dx}$$
$$\frac{d}{dx}[\tan(u)] = \sec^2(u)\frac{du}{dx}$$

$$\frac{d}{dx}[\cot(u)] = -\csc^2(u)\frac{du}{dx}$$
$$\frac{d}{dx}[\sec(u)] = \sec(u)\tan(u)\frac{du}{dx}$$
$$\frac{d}{dx}[\csc(u)] = -\csc(u)\cot(u)\frac{du}{dx}$$

The following examples illustrate how these rules can be used with other rules of differentiation.

Example 11.1.9 Find $\frac{d}{dx}[3\sin(x^2)]$.

Solution.

$$\begin{aligned}
\frac{d}{dx}[3\sin(x^2)] &= 3\frac{d}{dx}[\sin(x^2)] && \text{Constant Multiple}\\
&= 3\sin'(x^2)\cdot\frac{d}{dx}[x^2] && \text{Chain Rule, } u = x^2\\
&= 3\cos(x^2)\cdot\frac{d}{dx}[x^2] && \text{Derivative of Sine}\\
&= 3\cos(x^2)\cdot(2x) && \text{Derivative of Power}\\
&= 6x\cos(x^2)
\end{aligned}$$

□

Example 11.1.10 Find $\frac{d}{dx}[\sec(e^{3x})]$.

Solution.

$$\begin{aligned}
\frac{d}{dx}[\sec(e^{3x})] &= \sec'(e^{3x})\cdot\frac{d}{dx}[e^{3x}] && \text{Chain Rule, } u = e^{3x}\\
&= \sec(e^{3x})\tan(e^{3x})\frac{d}{dx}[e^{3x}] && \text{Derivative of Secant}\\
&= \sec(e^{3x})\tan(e^{3x})\cdot e^{3x}\frac{d}{dx}[3x] && \text{Chain Rule, } e^u \text{ with } u = 3x\\
&= 3e^{3x}\sec(e^{3x})\tan(e^{3x})
\end{aligned}$$

□

Example 11.1.11 Find $\frac{d}{dx}[e^{-3x}\sin(5x)]$.

Solution. The function is a product of e^{-3x} and $\sin(5x)$. Using the chain rule on these individual parts, we find

$$\begin{aligned}
\frac{d}{dx}[e^{-3x}] &= e^{-3x}\frac{d}{dx}[-3x]\\
&= -3e^{-3x}
\end{aligned}$$
$$\begin{aligned}
\frac{d}{dx}[\sin(5x)] &= \sin'(5x)\frac{d}{dx}[5x]\\
&= 5\cos(5x)
\end{aligned}$$

Knowing those derivatives, we use the product rule to find the derivative of the overall formula.

$$\frac{d}{dx}[e^{-3x}\sin(5x)] = \frac{d}{dx}[e^{-3x}]\cdot\sin(5x) + e^{-3x}\frac{d}{dx}[\sin(5x)] \quad \text{Product Rule}$$

$$= (-3e^{-3x})\sin(5x) + e^{-3x} \cdot (5\cos(5x))$$
$$= -3e^{-3x}\sin(5x) + 5e^{-3x}\cos(5x)$$

□

11.1.5 The Squeeze Theorem for Limits

Theorem 11.1.12 Limit Squeeze Theorem. *If $f(x)$ is bounded between two functions $\ell(x)$ (lower bound) and $u(x)$ (upper bound) and we know*

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

then this guarantees

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

More formally, if there exists $\delta > 0$ such that $\ell(x) < f(x) < u(x)$ whenever $a < x < a + \delta$ and $\ell(x) \to L$ and $u(x) \to L$ as $x \to a^+$, then

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

A similar statement holds for the lower limit and two-sided limit.

11.2 Derivatives of Inverse Trigonometric Functions

In the previous section, we used implicit differentiation to show that the derivative of an inverse function at a point is the reciprocal of the derivative of the original function at the equivalent inverse point. That is, if f^{-1} (an inverse function) has a point (x, y) on its graph, $y = f^{-1}(x)$, then f has a corresponding point (y, x), $x = f(y)$. The rate of change (derivative) of f^{-1} at (x, y) is the reciprocal of the rate of change of f at (y, x) defined by $f'(y)$:

$$\frac{d}{dx}[f^{-1}(x)] = \frac{1}{f'(y)} = \frac{1}{f'(f^{-1}(x))}.$$

We will use this result repeatedly in this section to find the derivative of the inverse trigonometric functions. The interesting thing about these derivatives is that they will all be algebraic. This is a consequence of the Pythagorean identities that relate trigonometric functions with their derivatives.

Theorem 11.2.1 Derivative of Arcsine.

$$\arcsin'(x) = \frac{d}{dx}[\sin^{-1}(x)] = \frac{1}{\sqrt{1-x^2}}$$

Proof. Starting with the relation $y = \arcsin(x) = \sin^{-1}(x)$, we have the inverse relation $x = \sin(y)$. Because $\sin'(x) = \cos(x)$, the derivative of the inverse is given by

$$\arcsin'(x) = \frac{1}{\sin'(y)} = \frac{1}{\cos(y)}.$$

The Pythagorean identity relates $\sin(y)$ and $\cos(y)$ by

$$\sin^2(y) + \cos^2(y) = 1$$

so that $\cos^2(y) = 1 - \sin^2(y) = 1 - x^2$. Because the domain for the restricted sine is in quadrants 1 and 4, $\cos(y) \geq 0$ and

$$\arcsin'(x) = \frac{1}{\sqrt{1-x^2}}.$$

■

Theorem 11.2.2 Derivative of Arccosine.

$$\arccos'(x) = \frac{d}{dx}[\cos^{-1}(x)] = \frac{-1}{\sqrt{1-x^2}}$$

Proof. Starting with the relation $y = \arccos(x) = \cos^{-1}(x)$, we have the inverse relation $x = \cos(y)$. Because $\cos'(x) = -\sin(x)$, the derivative of the inverse is given by

$$\arccos'(x) = \frac{1}{\cos'(y)} = \frac{-1}{\sin(y)}.$$

The Pythagorean identity and the domain of the restricted cosine (quadrants 1 and 2) implies $\sin(y) = \sqrt{1-x^2}$ such that

$$\arccos'(x) = \frac{-1}{\sqrt{1-x^2}}.$$

■

Notice that the derivatives of the arcsine and arccosine only differ by the change in sign. The arcsine has positive derivative, consistent with its graph

being increasing. In contrast, the arccosine is decreasing and has negative derivative.

However, it is more than just one function is positive and the other is negative. The formulas are identical other than sign because the graphs themselves are the same, except for a reflection and a shift. We know that $\cos(x) = \sin(\frac{\pi}{2} - x)$ which implies that

$$\arccos(x) = \frac{\pi}{2} - \arcsin(x).$$

Differentiation of this equation shows

$$\arccos'(x) = -\arcsin'(x).$$

A similar argument applies to a relationship between the derivatives of the arctangent and arccotangent and of the arcsecant and arccosecant. Consequently, we only need to find the derivatives of the arctangent and arcsecant.

Theorem 11.2.3 Derivative of Arctangent.

$$\arctan'(x) = \frac{d}{dx}[\tan^{-1}(x)] = \frac{1}{x^2+1}$$

Proof. Starting with the relation $y = \arctan(x) = \tan^{-1}(x)$, we have the inverse relation $x = \tan(y)$. Because $\tan'(x) = \sec^2(x)$, the derivative of the inverse is given by

$$\arctan'(x) = \frac{1}{\tan'(y)} = \frac{1}{\sec^2(y)}.$$

The Pythagorean identity relates $\tan(y)$ and $\sec(y)$ by

$$\tan^2(y) + 1 = \sec^2(y)$$

so that $\sec^2(y) = x^2 + 1$. Consequently

$$\arctan'(x) = \frac{1}{x^2+1}.$$

■

Theorem 11.2.4 Derivative of Arcsecant.

$$\operatorname{arcsec}'(x) = \frac{d}{dx}[\sec^{-1}(x)] = \frac{1}{|x|\sqrt{x^2-1}}$$

Proof. Starting with the relation $y = \operatorname{arcsec}(x) = \sec^{-1}(x)$, we have the inverse relation $x = \sec(y)$. Because $\sec'(x) = \sec(x)\tan(x)$, the derivative of the inverse is given by

$$\operatorname{arcsec}'(x) = \frac{1}{\sec'(y)} = \frac{1}{\sec(y)\tan(y)}.$$

The Pythagorean identity relates $\tan(y)$ and $\sec(y)$ by

$$\tan^2(y) + 1 = \sec^2(y)$$

so that $\tan^2(y) = \sec^2(y) - 1 = x^2 - 1$. In the restricted domain of the secant, the tangent and secant have the same sign so that the product will always be positive. Consequently

$$\operatorname{arcsec}'(x) = \frac{1}{|x|\sqrt{x^2-1}}.$$

■

Chapter 12

Other Stuff

12.1 Introduction to Optimization

When we know how to find extreme values of a function, we can use those techniques to answer physical questions involving optimization. Optimization problems involve at least two related physical quantities. One quantity is a control variable, a physical attribute of the system that one can adjust. The other quantity depends on the control variable and measures an aspect of the system that we wish to improve.

In this section, we consider some examples of optimization. The primary challenge for such problems is in clearly defining the system, identifying the control variable and the quantity to optimize. We then apply the calculus techniques for finding extreme values.

12.1.1 Objective Functions

Optimization is the application of finding extreme values to either maximize or minimize some quantity of interest. We usually have a physical incentive for this optimization, such as minimizing energy consumption, maximizing evolutionary fitness, minimizing costs of materials, or maximizing profit. The quantity of interest will depend on some independent variable that we have the ability to control or adjust. We call the mapping from the control variable to the physical quantity being optimized the **objective function**.

Frequently, identifying the appropriate objective function is the more challenging aspect of an optimization problem. Once the function is identified, the task is reduced to identifying local extreme values and behavior at end points. Sometimes the objective function depends on multiple independent variables that are related through some constraint. The constraint typically determines an equation which can be used to rewrite the objective functions as having only have a single independent variable. In addition, we need to determine a meaningful physical domain for the function.

We begin with several examples of creating objective functions for optimization problems. The actual analysis will follow later. Several simple examples come from geometry where we need to construct a shape that has some feature (like a given perimeter, area or volume) and we wish to make some other feature as large as possible. We use these examples not because they are practical but because they illustrate the principles of optimization effectively.

Example 12.1.1 Suppose we want to create a rectangle that has an area of 500 cm^2. Three sides will have one type of trim while the fourth side will have trim that is twice as expensive. What should be the dimensions of the rectange to minimize the cost of the trim?

Solution. Start by identifying the variables.

- h is the horizontal width of the rectangle
- v is the vertical length of the rectangle
- C is the cost of the trim around the rectangle

Once we have identified our variables, we need to find a formula for the cost because that is what we want to minimize. We will assume that the more expensive side is one of the horizontal lengths. Let p be the unit cost (per cm) of the less expensive trim so that $2p$ is the unit cost of the more expensive trim. The total cost of the trim is given by

$$C = (h + 2v) \cdot p + h \cdot (2p) = (3h + 2v) \cdot p.$$

Our objective function $(h, v) \mapsto C$ involves two independent variables. This means we need an additional constraint. Reviewing the problem, we recall that the total area needs to be 500 cm^2. The area is computed by $A = h \cdot v = 500$ so that we can treat v as another dependent variable,

$$v = \frac{500}{h}.$$

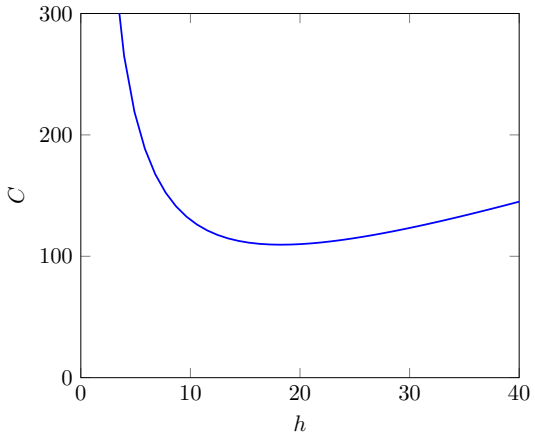
Substituting this formula into our objective function, we can rewrite it involving only a single independent variable h:

$$C = \left(3h + 2 \cdot \frac{500}{h}\right) \cdot p = 3ph + \frac{1000p}{h}.$$

Because p is a constant multiple in this formula, the location of the minimum will not depend on p.

Finally, we need to consider the physical domain for the objective function. The natural domain for the map $h \mapsto C$ is $h \neq 0$. However, negative values for h don't make physical sense. The physical domain for this problem will be $h \in (0, \infty)$. That is, the optimization problem will be answered by finding the global minimum of C on the interval $(0, \infty)$.

A graph of this relation is shown below using $p = 1$. The minimum value occurs somewhere near $h = 20$ with a cost C close to $100p$. We need to use calculus to find the exact value.



□

Example 12.1.2 Suppose you have a flexible pipe of length 10 meters that you will bend to make three sides of a rectangle. How long should you make these sides so that the rectangle has as large an area as possible?

Solution. We start by identifying the variables. It is often helpful to draw a figure. A sample diagram is shown in Figure 12.1.3. We label the two opposite vertical sides by the variable h (height) and the horizontal side by the variable w (width).

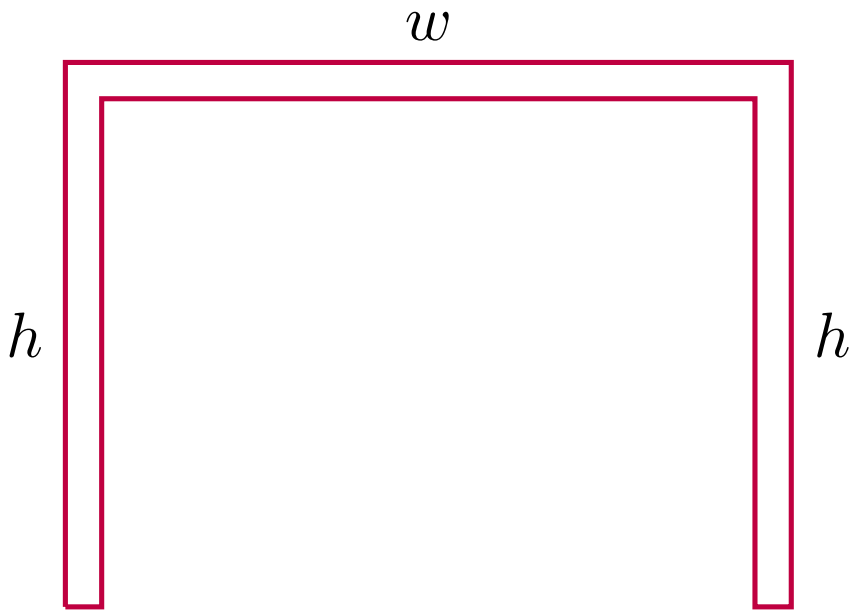


Figure 12.1.3 Three sides of a rectangle are made from a flexible pipe.

We want to make the area as large as possible. This makes the area of the rectangle A the dependent variable. The area of a rectangle is the height times the width, so our objective function is defined by

$$A = h \cdot w.$$

We need to write this as a function of one independent variable.

The constraint for our independent variables h and w is that the total length of pipe used is 10 meters. The pipe is used for two edges of length h and one edge of length w. As an equation, the constraint becomes

$$2h + w = 10.$$

If we solve this equation for w, we find

$$w = 10 - 2h$$

which we can substitute into the objective function,

$$A = h \cdot (10 - 2h) = 10h - 2h^2.$$

The last step is to identify the physical domain for the objective function. A physical measurement of length must be non-negative, so $h \geq 0$. What is the largest value of h that is possible? We need $w \geq 0$ which requires $10 - 2h \geq 0$. This implies $h \leq 5$. The physical domain is therefore $h \in [0, 5]$. Even though the shape would be an empty rectangle (no area), all of the variables are still defined when $h = 0$ or $h = 5$. We include the end points since closed intervals are easier to analyze.

A graph of the objective function is shown in Figure 12.1.4. We can see that the area will be maximized at the vertex of this parabola. Calculus will give us an efficient method to find this point.

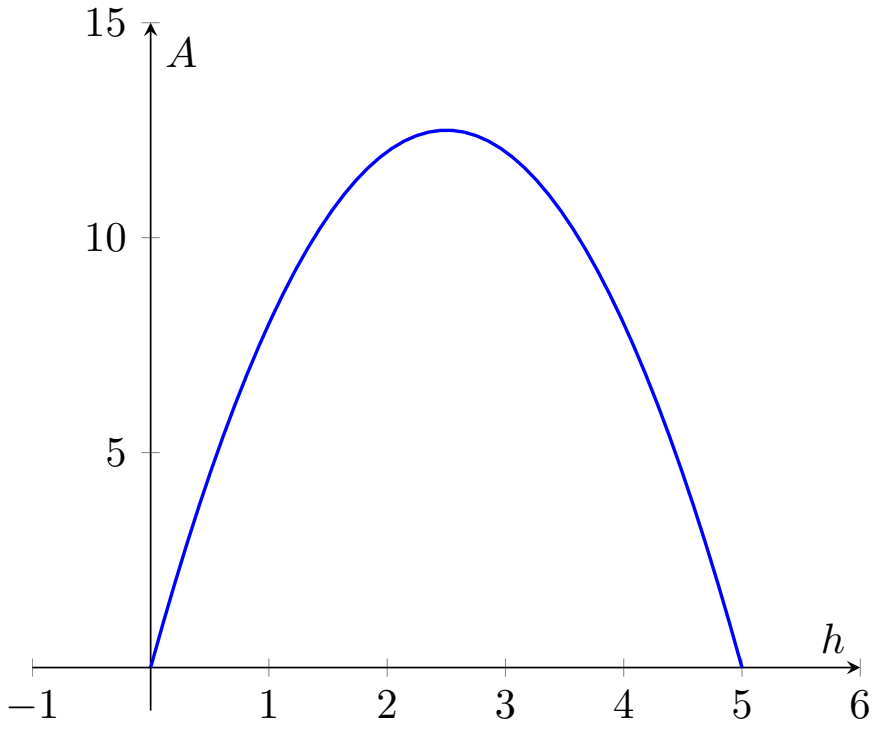


Figure 12.1.4 $A = 10h - 2h^2$ with domain $[0, 5]$

□

A biological example follows. A fundamental hypothesis of biology is that evolution drives organisms to maximize their fitness, which corresponds to the number of surviving offspring. There is often a trade-off between the number of offspring and the probability that the offspring survive. Let f (fecundity) represent the number of offspring an organism produces and let s (survival) represent the probability that an offspring will survive. The then fitness is given by $F = f \cdot s$, the average number of offspring that survive.

Example 12.1.5 Suppose that the survival probability is related to fecundity so that it decreases linearly. If each organism has ten offspring, the survival probability is $s = 0.95$. If each organism has forty offspring, the survival probability drops to $s = 0.8$. How many offspring should the organism have to maximize fitness?

Solution. First, identify the variables. The objective function is the fitness F which depends on both f (fecundity) and s (survival probability) through

$$F = f \cdot s.$$

This objective function has two independent variables, $(f, s) \mapsto F$.

We need to reduce the number of independent variables to a single variable by realizing that f and s will satisfy a linear relation. Because the original question asks for how many offspring should be produced, we will choose f to be the independent variable. The line passes through points $(f, s) = (10, 0.95)$ and $(f, s) = (40, 0.8)$. We can compute the slope

$$\frac{\Delta s}{\Delta f} = \frac{0.8 - 0.95}{40 - 10} = \frac{-0.15}{30} = -0.005.$$

Using the point-slope equation of a line, we find

$$s = 0.95 - 0.005(f - 10) = -0.005f + 1.$$

Using substitution in the objective function gives

$$F = f \cdot (-0.005f + 1) = -0.005f^2 + f.$$

To find the physical domain, we require $f \geq 0$ and $s \geq 0$. The second requirement becomes $-0.005f + 1 \geq 0$, which means that $f \leq 200$. The physical domain is therefore $f \in [0, 200]$. A graph shows that the maximum should occur at the vertex of a parabola.

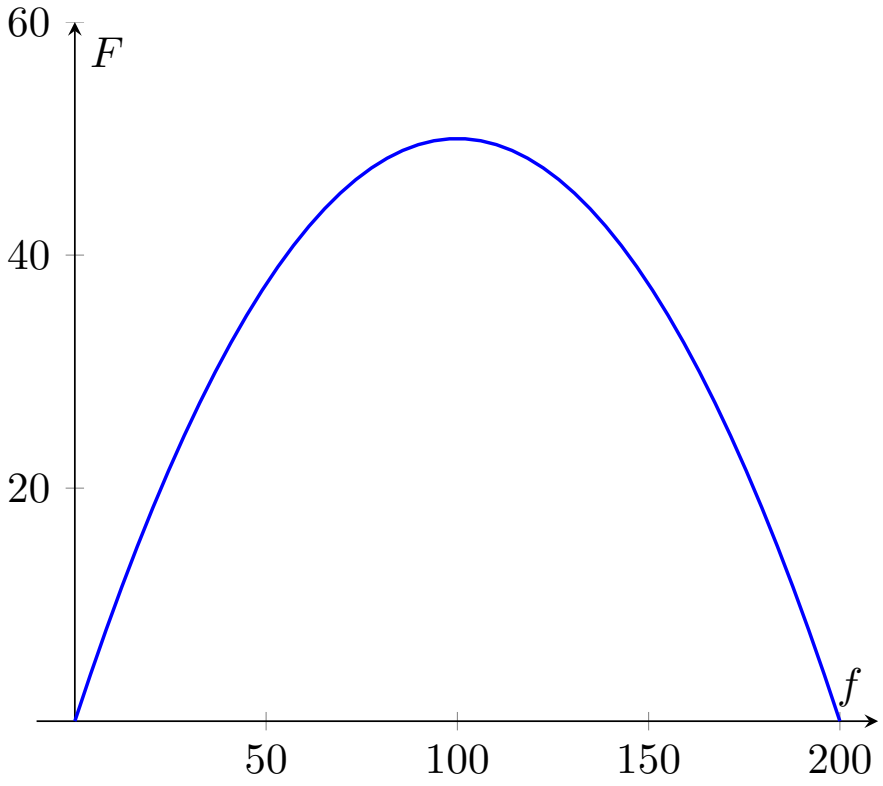


Figure 12.1.6 $F = -0.005f^2 + f$ with domain $[0, 200]$

□

12.1.2 Analysis for Optimization

Now that we have illustrated how to find the objective function for several examples, let us work through the analysis to solve the optimization problems. Two of our examples had objective functions that were quadratic polynomials. We start with those examples.

Example 12.1.7 The bent pipe example resulted in an objective function

$$A = 10h - 2h^2$$

and a physical domain $h \in [0, 5]$. Complete the optimization and find the dimensions that will maximize the area of the resulting rectangle.

Solution. To find the global extreme of the function $A(h) = 10h - 2h^2$, we begin by computing the derivative,

$$A'(h) = 10(1) - 2(2h) = 10 - 4h.$$

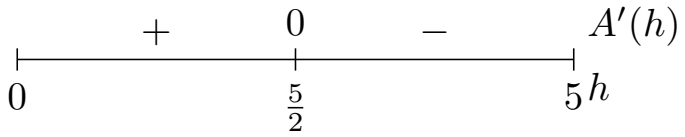
To perform sign analysis of $A'(h)$, we first find the root $A'(h) = 0$:

$$\begin{aligned} 10 - 4h &= 0 \\ \frac{10}{4} &= h \\ h &= \frac{5}{2}. \end{aligned}$$

Our test intervals are $[0, \frac{5}{2})$ and $(\frac{5}{2}, 5]$. Testing the sign at $h = 1$ and $h = 4$ as sample points, we find

$$\begin{aligned} A'(1) &= 10 - 4(1) = 6 > 0, \\ A'(4) &= 10 - 4(4) = -6 < 0. \end{aligned}$$

The results of our sign analysis are summarized on the following number line.



Our sign analysis of $A'(h)$ implies that A has a maximum value at $h = \frac{5}{2}$. Because A is increasing on $[0, \frac{5}{2}]$ and decreasing on $[]\frac{5}{2}, 5]$, we see that this is a global maximum on the domain. The resulting dimensions of the rectangle are $h = \frac{5}{2} = 2.5$ meters and $w = 10 - 2h = 10 - 2(2.5) = 5$ meters. The area of the rectangle will be $A = 12.5$ square meters. □

Example 12.1.8 The fitness example resulted in an objective function

$$F = -0.005f^2 + f$$

and a physical domain $f \in [0, 200]$. Complete the optimization to find the number of offspring that will maximize the fitness.

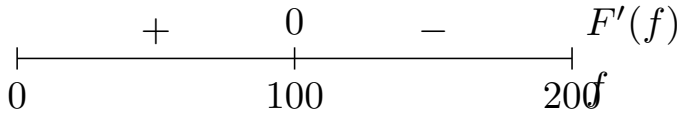
Solution. To find the global maximum of $F(f)$, we first compute the derivative,

$$F'(f) = -0.005(2f) + 1 = -0.01f + 1.$$

The root $F'(f) = 0$ occurs at $f = 100$. Our sign analysis uses test intervals $[0, 100)$ and $(100, 200]$. We compute the sign of $F'(f)$ at sample points $f = 0$ and $f = 200$:

$$\begin{aligned} F'(0) &= -0.01(0) + 1 = 1 > 0, \\ F'(200) &= -0.01(200) + 1 = -1 < 0. \end{aligned}$$

The results of our sign analysis are summarized on the following number line.



The First Derivative Test allows us to conclude that F has a local maximum value at $f = 100$. Because F is increasing on $[0, 100]$ and decreasing on $[100, 200]$, this must also be a global maximum. The fitness will be maximized when each individual reproduces with 100 offspring. □

For our third example, the objective function is not a polynomial. Because we have not yet established a rule for the derivative in this case, we will use

technology to find it.

Example 12.1.9 The cost to put trim on our rectangle was found to be the objective function

$$C(h) = 3ph + \frac{1000p}{h}$$

with a physical domain $h \in (0, \infty)$.

Solution. The SageMath computer algebra system allows us to compute derivatives automatically.

```
# Tell the system about our variables
var('h','p')
# Define our function
C(h) = 3*p*h + 1000*p/h
# Compute the derivative with variable h
show( diff(C(h), h) )
```

```
3*p - 1000*p/h^2
```

We now know

$$C'(h) = 3p - \frac{1000p}{h^2}.$$

Like $C(h)$, this derivative is not defined for $h = 0$ but is otherwise continuous. We find a root by solving $C'(h) = 0$ and finding a common denominator:

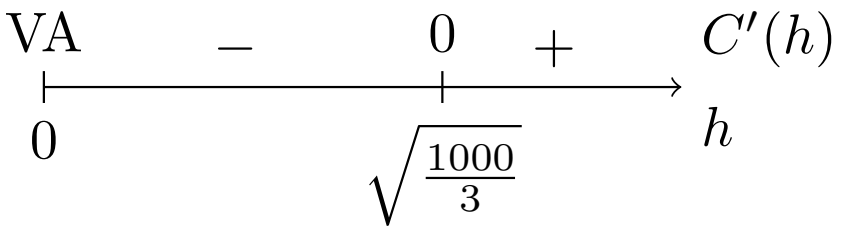
$$\begin{aligned} 3p - \frac{1000p}{h^2} &= 0 \\ \frac{3ph^2}{h^2} - \frac{1000p}{h^2} &= 0 \\ \frac{p(3h^2 - 1000)}{h^2} &= 0 \\ 3h^2 - 1000 &= 0 \\ h^2 &= \frac{1000}{3} \\ h &= \pm\sqrt{\frac{1000}{3}} \end{aligned}$$

Only $h = +\sqrt{\frac{1000}{3}} \approx 18.257$ is in the domain.

We can test the signs of $C'(h)$ using $h = 10$ and $h = 20$.

$$C'(10) = 3p - \frac{1000p}{10^2} = 3p - 10p = -7p < 0$$

$$C'(20) = 3p - \frac{1000p}{20^2} = 3p - \frac{5p}{2} = \frac{p}{2} > 0$$



The First Derivative Test shows that C has a local minimum at $h = \sqrt{\frac{1000}{3}}$. Because C is decreasing on $(0, \sqrt{\frac{1000}{3}}]$ and increasing on $[\sqrt{\frac{1000}{3}}, \infty)$, this minimum is a global minimum.

We finish by interpreting our mathematics. The question was how to find the dimensions of the rectangle. Our analysis gave us a value for $h = \sqrt{\frac{1000}{3}} \approx 18.257\ cm$. We also need v, which was another dependent variable:

$$v = \frac{500}{h} = 500 \cdot \sqrt{frac31000} \approx 27.386\ cm.$$

The minimal cost to trim a rectangle would have a horizontal length of 18.26 cm, one of which has the more expensive trim, and a vertical length of 27.386 cm. □

12.1.3 Summary

- Optimization is the application of finding extreme values to physical problems. The dependent variable is the quantity that should be as large or as small as possible. The independent variable(s) are the quantities we can adjust. The map from the independent variable to the dependent variable is called the **objective function**.
- When more than one independent variable is involved, an extra equation called a constraint allows us to solve for one of the independent variables in terms of the other.
- A physical domain for the objective function represents the values of the independent variable(s) that are physically relevant.

12.1.4 Exercises

1. A rectangular frame will be made with horizontal edges that cost \$0.50 per inch and vertical edges that cost \$0.40 per inch. What are the dimensions of a rectangle that will maximize the enclosed area for a total cost of \$20.00?

2. Suppose that the survival probability for a species is related to fecundity so that it decreases linearly. If each organism has five offspring, the survival probability is $s = 0.9$. If each organism has twenty offspring, the survival probability drops to $s = 0.75$. How many offspring should the organism have to maximize fitness?

3. A population of animals has the property that each individual has fewer offspring per year when the population is bigger. When the population has 200 individuals, the average number of offspring per individual per year is 4.8. When the population has 300 individuals, the average number of offspring per individual per year drops to 4.2. Assuming a linear relation between the per capita number of offspring per year and the population size, what population size corresponds to the largest total number of offspring per year? (The total number of offspring equals the per capita number of offspring times the population size.)

4. A company sells bowling balls. The higher the price the company charges, the fewer balls are sold. When the price is \$50, the company can sell 500 balls per week. When the price is \$60, the company can sell 400 balls per week. Assuming a linear relation between the price and the number of balls sold per week, find the price for which the company earns the most revenue per week. (Weekly revenue equals the price per ball times the number of balls sold per week.)

5. A rectangular container with a square base (top/bottom) is to be manufactured. The top and bottom (squares) are made from a material that

costs \$1.50 per square meter while the other four sides (rectangles) are made from a material that costs \$1.00 per square meter. What should be the dimensions of the container that would maximize the volume and cost \$20 in materials?

6. A rectangular container with a square profile (front/back) is to be manufactured. The top and bottom (rectangles) are made from a material that costs \$1.50 per square meter while the other four sides (two squares and two rectangles) are made from a material that costs \$1.00 per square meter. What should be the dimensions of the container that would maximize the volume and cost \$20 in materials?

7. A beverage can is being designed in the shape of a circular cylinder (volume=$\pi r^2 h$). The top and bottom (circles, area=πr^2) are made from metal that costs \$0.01 per square centimeter while the curved wall of the can (curved rectangle, area=$2\pi rh$) is made from metal that costs \$0.004 per square centimeter. What should be the radius and height of the can that would maximize the volume in the container for a can that costs \$0.25 in materials?

8. A rectangular box with a square base and no top needs to contain a volume of 1000 cubic centimeters. The square base (all sides equal) is made from a material that costs 10 cents per square centimeter. The other four sides are made from a material that costs 6 cents per square centimeter. What dimensions should the box have to minimize the total cost of materials?

9. A beverage can is being designed in the shape of a circular cylinder to hold 360 cubic centimeters (volume=$\pi r^2 h$). The top and bottom (circles, area=πr^2) are made from metal that costs \$0.01 per square centimeter while the curved wall of the can (curved rectangle, area=$2\pi rh$) is made from metal that costs \$0.004 per square centimeter. What should be the radius and height of the can that would minimize the materials cost?

12.2 Extreme Values and Optimization

We have already learned that derivatives can help us identify the location of local extreme values, points that are the highest or lowest values in a neighborhood of that point. It is often the case that we need to find the highest or lowest value that a function ever achieves, not just in its own neighborhood. We call these global extremes.

Applications involving the identification of extreme values are often called optimization problems. The task in optimization is to identify the value of an independent variable in the system that will maximize or minimize some objective. Aside from the calculus in finding the extreme values, creating an appropriate function that will serve as the objective is often the greatest challenge.

12.2.1 Global Extreme Values

The Extreme Value Theorem guarantees that a continuous function restricted to a closed interval will always have global maximum minimum values. Those extremes can only occur at either the end points of the interval or at critical points (points with horizontal or undefined tangents). This guides our strategy.

1. Compute $f'(x)$.
2. Identify critical points: solve $f'(x) = 0$ and identify all points where $f'(x)$ is not defined.
3. Classify the points, including the end points for comparison.

Example 12.2.1 Find the maximum and minimum values of $f(x) = \dfrac{x-1}{x^2+1}$ on the interval $[-2, 2]$.

Solution. Because the denominator of $f(x)$ is never zero, $x^2 + 1 \neq 0$, $f(x)$ is continuous everywhere. Consequently, the Extreme Value Theorem guarantees that f will have a maximum and minimum value on the closed interval $[-2, 2]$.

First, we compute $f'(x)$, which involves the quotient rule.

$$\begin{aligned} f'(x) &= \frac{d}{dx}\left[\frac{x-1}{x^2+1}\right] \\ &= \frac{(x^2+1)(1) - (x-1)(2x)}{(x^2+1)^2} \\ &= \frac{x^2+1-2x^2+2x}{(x^2+1)^2} \\ &= \frac{-x^2+2x+1}{(x^2+1)^2} \end{aligned}$$

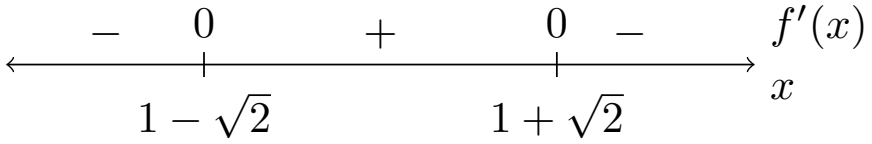
Next, we find critical points. Again, the denominator is nonzero, so $f'(x)$ will be defined and continuous for every value x. We only need to find solutions to $f'(x) = 0$, which are solutions to $-x^2 + 2x + 1 = 0$. The quadratic formula gives

$$x = \frac{-2 \pm \sqrt{4 - 4(-1)(1)}}{2(-1)} = \frac{-2 \pm 2\sqrt{2}}{-2}.$$

The critical values are $x = 1 - \sqrt{2} \approx 0.414$ and $1 + \sqrt{2} \approx 2.414$.

If we were to test these critical values as turning points, we would look at the signs of $f'(x)$ on the intervals formed by the critical points. The sign

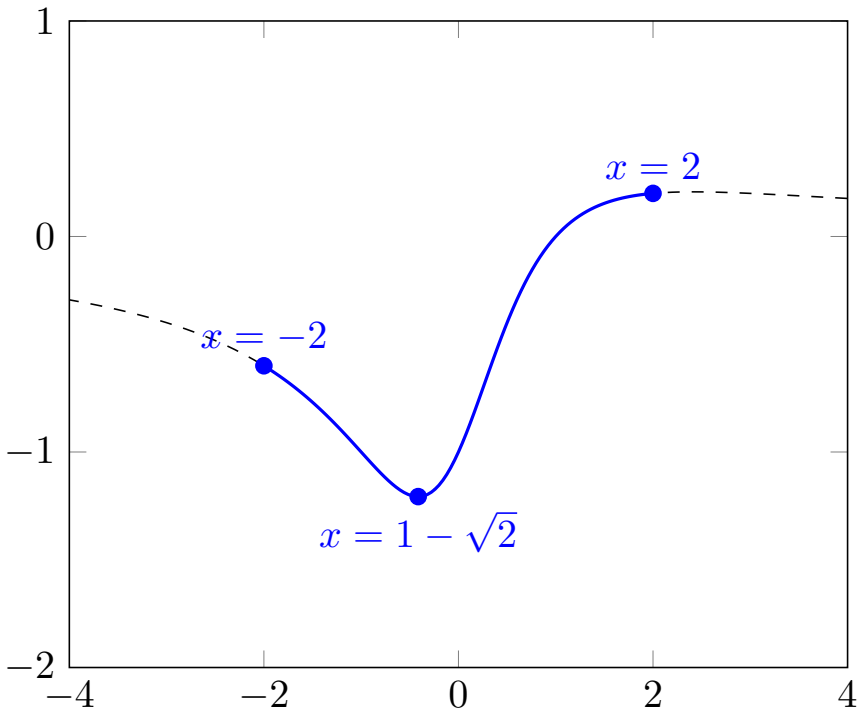
analysis is summarized with the number line below, showing that $x = 1 - \sqrt{2}$ is a local minimum and $x = 1 + \sqrt{2}$ is a local maximum.



To find the extreme values on the interval, we really just need to compare the values of $f(x)$ at the end points of the interval with the critical points that are inside the interval. Because $x = 1+\sqrt{2}$ is outside $[-2, 2]$, we do not include that point.

$$f(-2) = \frac{-2-1}{(-2)^2+1} = \frac{-3}{5} = -0.6$$
$$f(2) = \frac{2-1}{(-2)^2+1} = \frac{1}{5} = 0.2$$
$$f(1-\sqrt{2}) = \frac{(1-\sqrt{2})-1}{(1-\sqrt{2})^2+1} = \frac{-\sqrt{2}}{4-2\sqrt{2}} \approx -1.207$$

We finish by interpreting our results. The maximum value of f on the interval $[-2, 2]$ is $\frac{1}{5}$, occurring at $x = 2$. The minimum value of f on the interval is $\frac{\sqrt{2}}{4-2\sqrt{2}} \approx -1.207$, occurring at $x = 1 - \sqrt{2} \approx 0.414$. A graph of the function showing these extremes is given below.



□

When the function is not continuous or the interval is not a closed interval, the function is not guaranteed to have global extreme values. When an end point for a continuous function is not included, the function achieves every value up to the limit at that end point. This means that it is possible that the function does not actually have an extreme value. For every value the function does achieve, there may be another value that is more extreme.

Example 12.2.2 The function $f(x) = x$ restricted to $(0, 1)$ is continuous. The values (range) is obviously $(0, 1)$. But f does not have a maximum or minimum value. The upper limit $y = 1$ is never achieved because $x = 1$ is not in the restricted domain. But for any value $y < 1$, there will always be another value that is larger. Similarly, $y = 0$ is an unacheived lower limit and f does not have a minimum value. □

When looking for extreme values for functions that have discontinuities or where the domain is restricted to an open interval, we do the same things as for continuous functions: find critical points and compare values of the function. However, we need to include any relevant limits in the comparison. If a limiting value is more extreme than any of the achieved extreme values, then the function does not achieve that extreme value.

Example 12.2.3 Find the extreme values of the function $f(x) = x^3+x^2-2x+2$ on the interval $(-2, 2)$.

Solution. The function is continuous. If the interval was closed, $[-2, 2]$, the Extreme Value Theorem would guarantee that it had a maximum and minimum value. By excluding the end points, we might no longer achieve one or both of those values.

First, find $f'(x) = 3x^2 + 2x - 2$. Use this to find critical points, where $f'(x) = 0$. The quadratic formula is needed:

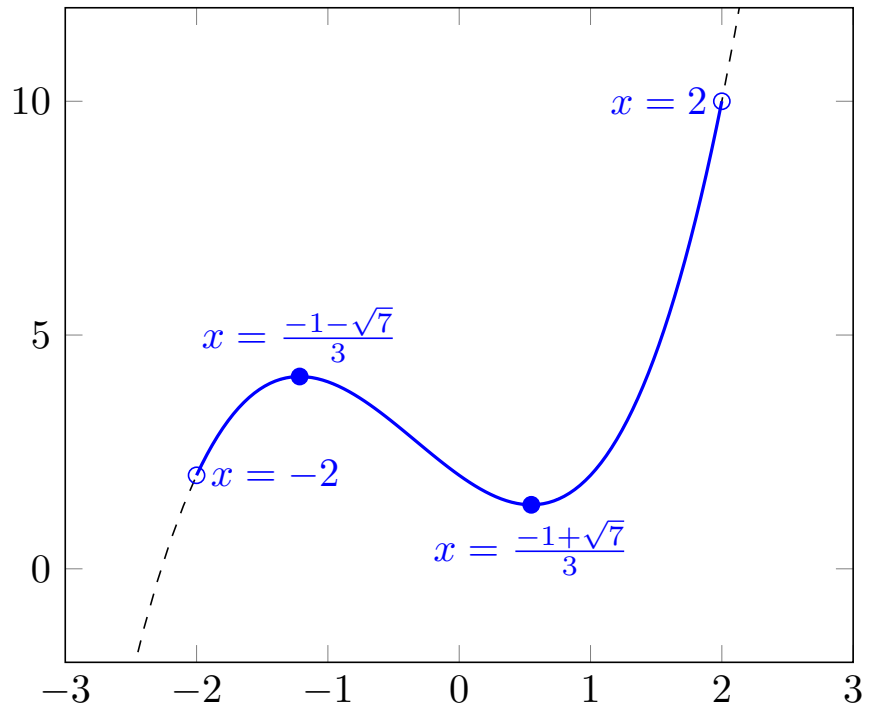
$$\begin{aligned} x &= \frac{-2 \pm \sqrt{4 - 4(3)(-2)}}{2(3)} \\ &= \frac{-2 \pm \sqrt{28}}{6} \\ &= \frac{-1 \pm \sqrt{7}}{3}. \end{aligned}$$

The critical values are $x = \frac{-1-\sqrt{7}}{3} \approx -1.215$ and $x = \frac{-1+\sqrt{7}}{3} \approx 0.5486$. Both critical values are in the interval.

We now compare the values of the function at the critical values and at the end points.

$$\begin{aligned} &f(\frac{-1 - \sqrt{7}}{3}) \approx 4.113 \\ &f(\frac{-1 + \sqrt{7}}{3}) \approx 1.369 \\ &f(-2) = 2 \\ &f(2) = 10 \end{aligned}$$

The minimum value for f on the interval $[-2, 2]$ is approximately 1.369 at $x = \frac{-1+\sqrt{7}}{3}$; the maximum value is 10 at $x = 2$. However, when working with the open interval $(-2, 2)$, the maximum value is a limit value that is not achieved. So f does not have a maximum value on $(-2, 2)$, though it does have the same minimum value.



□

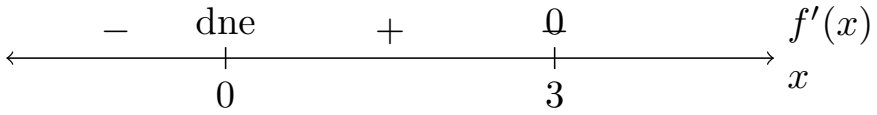
Example 12.2.4 Find the extreme values of the function $f(x) = \frac{10x-15}{x^2}$.

Solution. The function $f(x) = \frac{10x-15}{x^2}$ has a discontinuity at $x = 0$ because division by zero is undefined. This corresponds to a vertical asymptote at $x = 0$. Because the question does not give an interval, we must be considering the entire domain, $(-\infty, 0) \cup (0, \infty)$. Extreme values are not guaranteed.

We begin by finding critical values. The derivative requires the quotient rule.

$$\begin{aligned} f'(x) &= \frac{d}{dx}[\frac{10x-15}{x^2}] \\ &= \frac{x^2(10) - (10x-15)(2x)}{x^4} \\ &= \frac{10x^2 - 20x^2 + 30x}{x^4} \\ &= \frac{x(-10x+30)}{x^4} = \frac{-10x+30}{x^3} \end{aligned}$$

The critical value is the solution to $-10x + 30 = 0$, or $x = 3$. When doing sign analysis, we also need to use the discontinuity $x = 0$ to create the intervals that are tested.



The sign analysis on $f'(x)$ informs us about the vertical asymptote. The function decreases immediately to the left of $x = 0$, letting us know

$$\lim_{x \to 0^-} f(x) = -\infty.$$

Similarly, the function is increasing immediately to the right of $x = 0$, showing

that

$$\lim_{x\to 0^+} f(x) = -\infty.$$

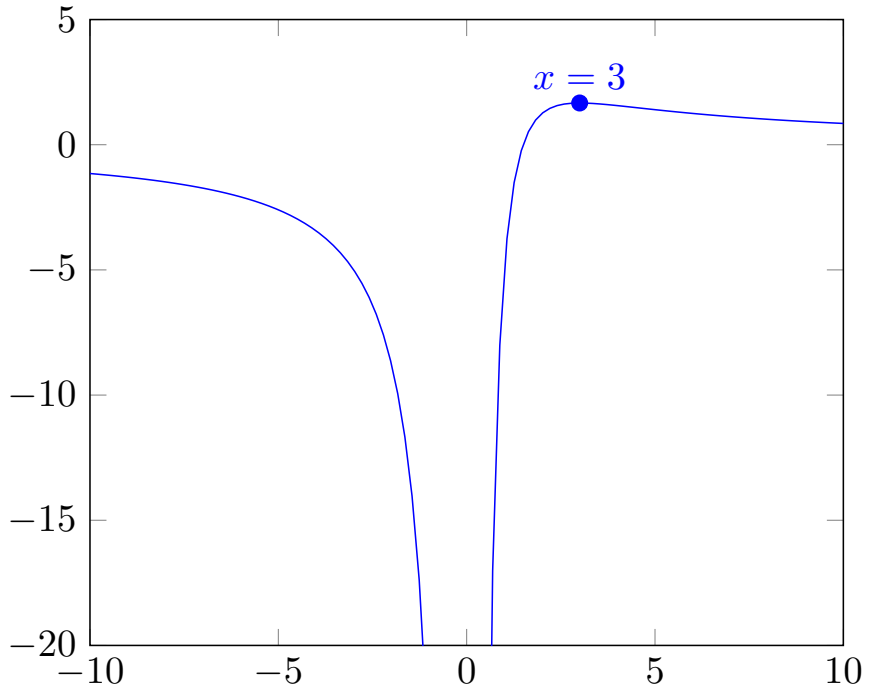
This is enough to guarantee that $f(x)$ does not have a minimum value because it is unbounded in the negative direction.

The critical point $x = 3$ is a turning point corresponding to a local maximum. The value of the function is $f(3) = \frac{15}{9} = \frac{5}{3}$. To see if this is a global maximum, we need to compare it with the ends of the intervals, as $x \to \pm\infty$. These limits are both zero.

$$\lim_{x\to -\infty} f(x) = \lim_{x\to -\infty} \frac{10x-15}{x^2} = \lim_{x\to -\infty} \frac{10}{x} - \frac{15}{x^2} \to \frac{10}{-\infty} - \frac{15}{\infty} = 0$$
$$\lim_{x\to \infty} f(x) = \lim_{x\to -\infty} \frac{10x-15}{x^2} = \lim_{x\to -\infty} \frac{10}{x} - \frac{15}{x^2} \to \frac{10}{\infty} - \frac{15}{\infty} = 0$$

These limits show that $y = 0$ is a horizontal asymptote for $f(x)$ as $x \to \pm\infty$.

Interpreting our results, we see that $f(x)$ has a maximum value of $\frac{5}{3}$ at $x = 3$ and no minimum value due to the infinite limit at $x = 0$.



□

12.2.2 Optimization

Optimization is the application of finding extreme values to either maximize or minimize some quantity of interest. In general, we will have a system where there is some variable that we have freedom to vary and some quantity that is a function of that variable that we want to be at a maximum or minimum value. The variable that we vary is the independent variable. The quantity that we optimize is called the **objective function**.

For example, consider a crystal goblet. When a pure note is sounded, the goblet will resonate with a strength that depends on the frequency of the note. If we wanted to shatter the goblet, we would want to find the frequency with which the goblet resonated the most. In this example, the frequency of the

note being played is the independent variable and the strength of resonance would be the objective function.

Example 12.2.5 The most elementary example of resonance is for a forced simple harmonic oscillator. The independent variable is the forcing frequency ω. The objective function is the amplification factor of the resonant response A. The system also has parameters related to the oscillator itself: ω_0, which represents natural frequency of the oscillator in the absence of friction, and α, which represents a rate at which the oscillator's motion would decay in the absence of a stimulus. The amplification factor is defined by the equation

$$A(\omega) = \frac{1}{\sqrt{(\omega^2 - \omega_0^2)^2 + 4\alpha^2\omega^2}} = \left((\omega^2 - \omega_0^2)^2 + 4\alpha^2\omega^2\right)^{-1/2}.$$

Find the frequency that is amplified the most.

Solution. Physically, the driving frequency must be a non-negative value, so $\omega \geq 0$. So we will look for extreme values on the domain $[0, \infty)$. The denominator involves the sum of two squares which can not be simultaneously equal to zero. Consequently, $A(\omega)$ is a continuous function defined for all values of ω. We need to find the critical values by solving $A'(\omega) = 0$. Computing A' involves repeated use of the chain rule.

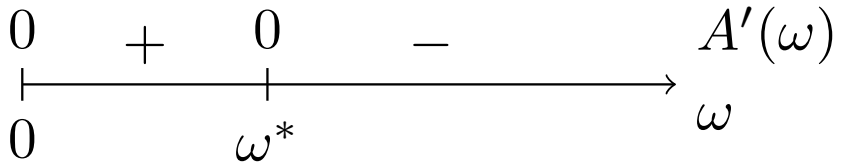
$$\begin{aligned} A'(\omega) &= \frac{d}{d\omega}[\left((\omega^2 - \omega_0^2)^2 + 4\alpha^2\omega^2\right)^{-1/2}] \\ &= -\frac{1}{2}\left((\omega^2 - \omega_0^2)^2 + 4\alpha^2\omega^2\right)^{-3/2} \cdot \frac{d}{d\omega}[(\omega^2 - \omega_0^2)^2 + 4\alpha^2\omega^2] \\ &= -\frac{1}{2}\left((\omega^2 - \omega_0^2)^2 + 4\alpha^2\omega^2\right)^{-3/2} \cdot \left(2(\omega^2 - \omega_0^2)(2\omega) + 8\alpha^2\omega\right) \\ &= -\frac{\omega(\omega^2 - \omega_0^2 + 2\alpha^2)}{\left((\omega^2 - \omega_0^2)^2 + 4\alpha^2\omega^2\right)^{3/2}} \end{aligned}$$

Critical values are solutions to the equation $2\omega(\omega^2 - \omega_0^2 + 2\alpha^2) = 0$. Solutions occur when $\omega = 0$ and $\omega^2 = \omega_0^2 - 2\alpha^2$, which only occurs when $\omega_0^2 > 2\alpha^2$.

When $\omega_0^2 > 2\alpha^2$, the sign of $A'(\omega)$ changes sign at the critical value at

$$\omega = \omega^* \equiv \sqrt{\omega_0^2 - 2\alpha^2}$$

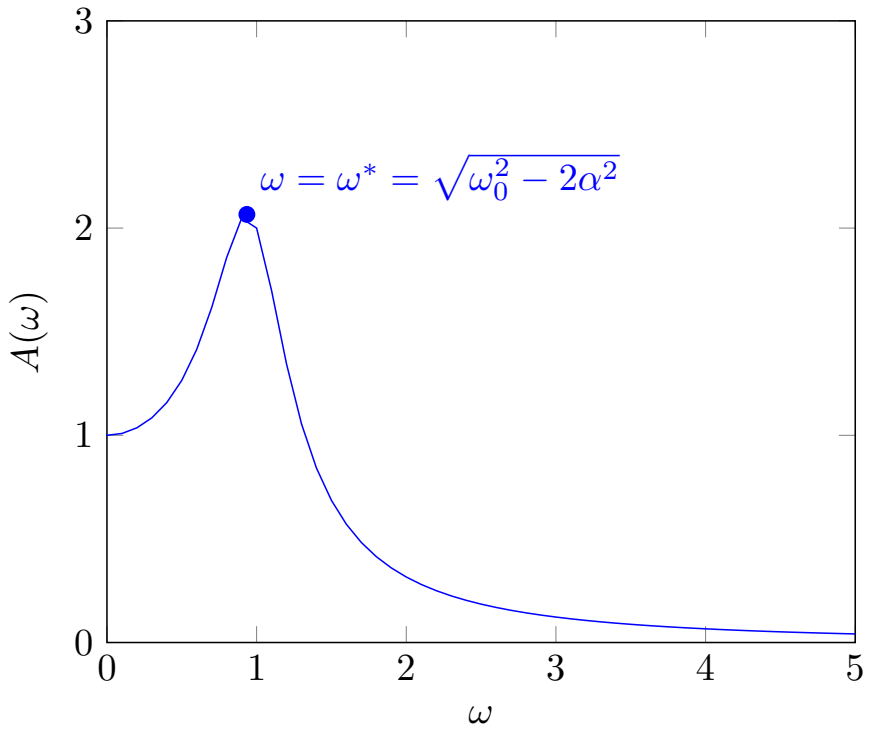
since the factor $\omega^2 - (\omega_0^2 - 2\alpha^2)$ changes sign. The denominator will always be positive. This allows us to determine the sign analysis of $A'(\omega)$.



The sign analysis of $A'(\omega)$ shows that $A(\omega)$ is increasing for ω in the interval $[0, \omega^*]$ and decreasing for ω in the interval $[\omega^*, \infty)$. Consequently, A achieves a maximum value when $\omega = \omega^*$. That maximum value is

$$A(\omega^*) = \frac{1}{\sqrt{4\alpha^2(\omega_0^2 - \alpha^2)}}.$$

A typical resonance response curve is shown below for this case ($\omega_0 = 1$ and $\alpha = 0.25$).



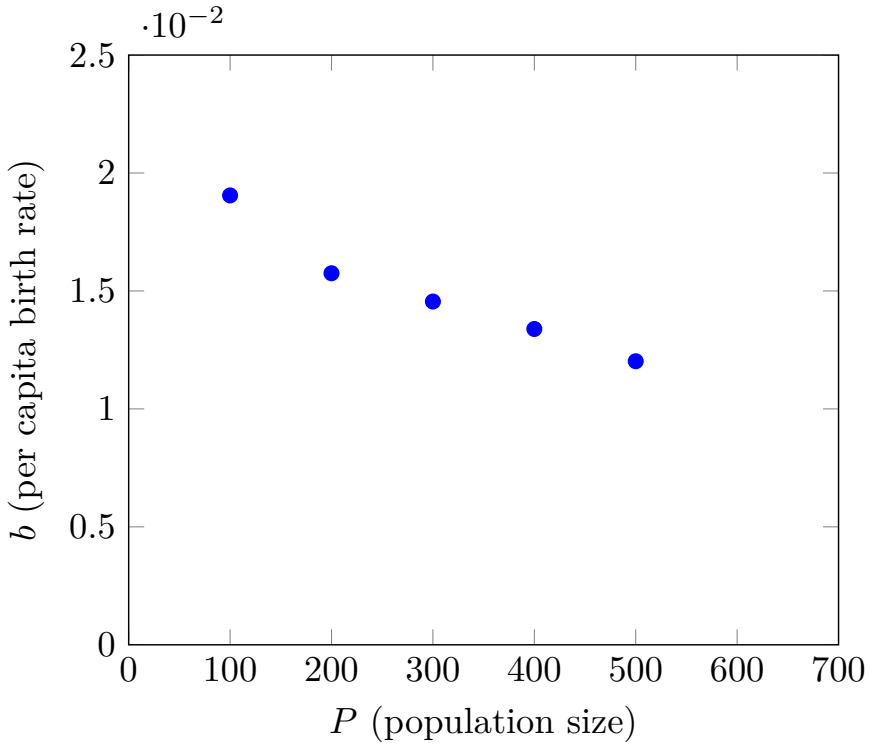
In the case that $\omega_0^2 < 2\alpha^2$ and the only critical value is $\omega = 0$, the factor in the numerator of $A'(\omega)$ given by $\omega^2 - \omega_0^2 + 2\alpha^2$ will always be positive. This implies that $A'(\omega) < 0$ for all $\omega > 0$, meaning that $A(\omega)$ is a decreasing function whose maximum must be at the end-point $\omega = 0$. □

Not every application involves unspecified parameters.

Example 12.2.6 The number of births in a population during a given time period is equal to the per capita birth rate times the population size. Suppose that the per capita birth rate was found to also depend on the population size. Average per capita birth rates for certain controlled population sizes were experimentally obtained and shown in the table below. Find a model for the per capita birth rate as a function of population size and use that model to predict the maximum population birth rate.

Population	Per Capita Birth Rate
100	0.0190
200	0.0158
300	0.0146
400	0.0134
500	0.0120

Solution. We start by looking at the graph of the data. Let P be the population size and let b be the per capita birth rate. We are interested in the function $P \mapsto b$, so the graph plots points (P, b).



The relationship between these points is decreasing. A linear model looks like it could be appropriate. Using linear regression on our data, we find a model

$$b(P) = -1.6423 \times 10^{-5} P + 0.019878.$$

The total birth rate is defined as $B(P) = b(P) \cdot P$, which according to our model is given by

$$B = 0.019878P - 1.6423 \times 10^{-5} P^2.$$

This is our objective function, the quantity we want as large as possible.

To find the maximum value of B, we compute $B' = 0.019878 - 3.2846 \times 10^{-5} P$ and solve $B' = 0$ to find the critical point.

$$0.019878 - 3.2846 \times 10^{-5} P = 0$$
$$0.019878 = 3.2846 \times 10^{-5} P$$
$$P = \frac{0.019878}{3.2846 \times 10^{-5}} = 605.2$$

Because $B'' = -3.2846 \times 10^{-5}$ is negative, we know that B is a concave down function and the critical value corresponds to a maximum value. The maximum total birth rate is

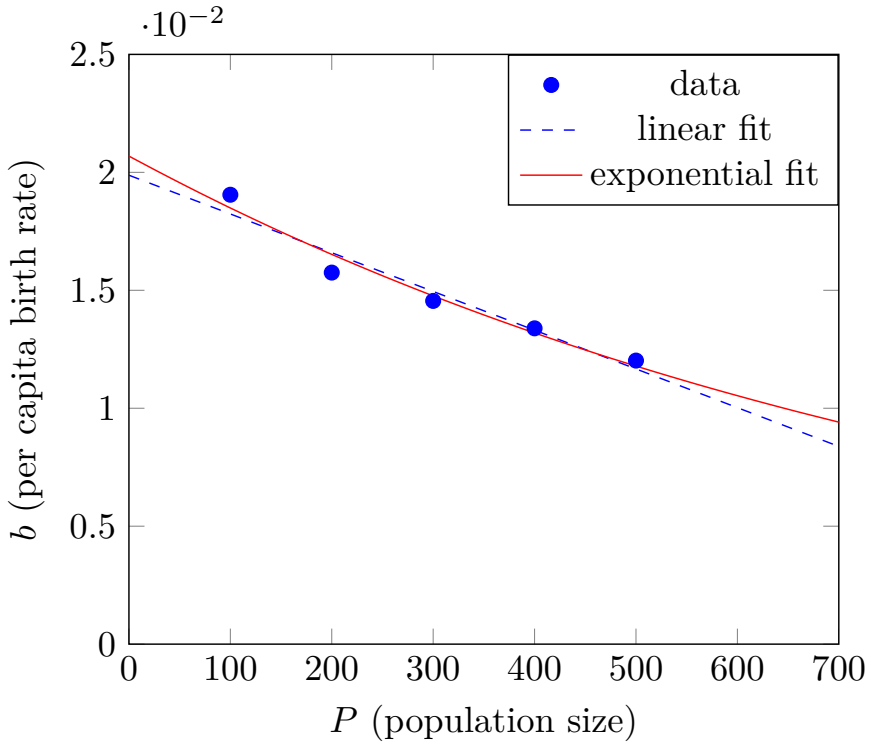
$$B(605.2) = 6.015,$$

occurring for $P = 605.2$. A population should be an integer value, so we would expect the maximum birth rate to occur when $P = 605$.

However, our prediction is dependent on the model that we chose. What if we used a different equation for our original per capita birth rate? The original data points have the appearance of being concave up. So maybe we should use an exponential regression curve. This gives

$$b(P) = 0.02069e^{-0.001125P}.$$

The graph below shows the two models for per capita birth rate with the data.



Using the exponential model for the per capita birth rate, we obtain a modified objective function

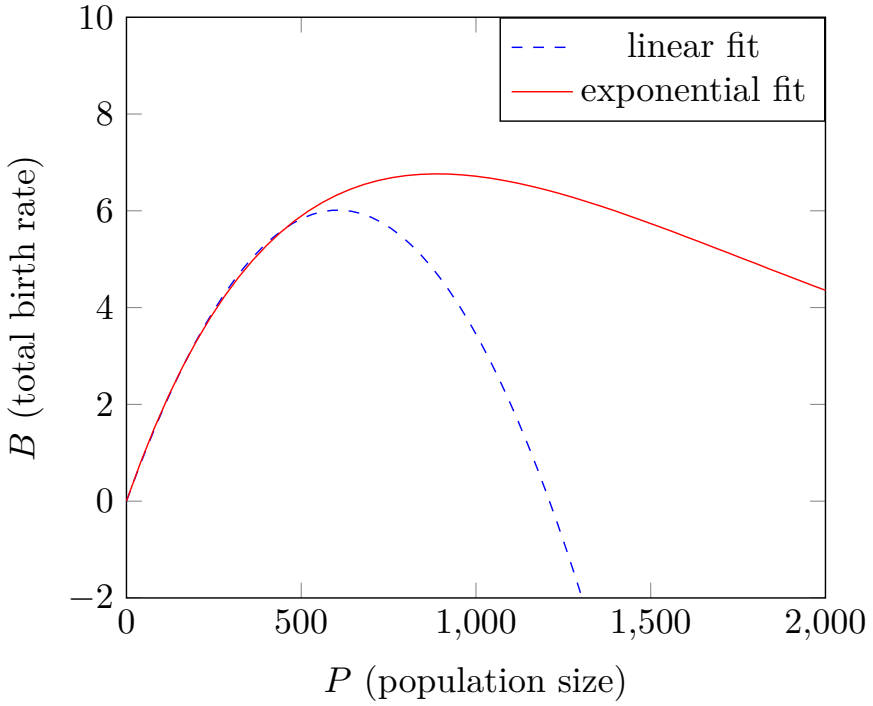
$$B(P) = 0.02069Pe^{-0.001125P}.$$

The derivative is

$$B'(P) = 0.02069e^{-0.001125P} + 0.02069P(-0.001125e^{-0.001125P}) = 0.02069e^{-0.001125P}(1 - 0.001125P).$$

The critical value must solve $1 - 0.001125P$ since the exponential factor is always positive. This gives $P = \frac{1}{0.001125} \approx 888.9$. Sign analysis of $B'(P)$ shows that this critical value corresponds to a maximum value, $B(888.9) = 6.763$.

Notice how two different models can yield significantly different predictions. The two models are illustrated together in the figure below. Notice that the original, quadratic model has the non-physical predication of a negative number of births for sufficiently large population sizes.


10
8
6
4
2
0
−2
0
500
1,000
1,500
2,000
B (total birth rate)
P (population size)
linear fit
exponential fit

□

12.3 Integrals and the Method of Substitution

Every rule for differentiation has a corresponding rule for integrals or antidifferentiation. This section focuses on the integration rule that corresponds to the chain rule.

Recall that the chain rule states that if $F(x)$ is a function with a derivative $F'(x)$ and u is any expression (or function), then

$$\frac{d}{dx}[F(u)] = F'(u)\frac{du}{dx}.$$

The corresponding antidifferentiation rule says that if we have a function $f(x)$ with an antiderivative $F(x)$ ($F'(x) = f(x)$), then

$$\int f(u)\frac{du}{dx}dx = F(u) + C.$$

Usually, the integrand does not appear so obviously in the form of the chain rule. The method of substitution provides a formalized method to guide the process. The method relies on transforming the integral from an integrand in terms of the independent variable, say x, as a new integral with an integrand in terms of the chain variable u. For the transformation to be valid, we must account for the chain rule factor $\frac{du}{dx} = u'$. We use the substitution rule for differentials

$$du = u' \cdot dx \qquad \Leftrightarrow \qquad dx = \frac{du}{u'}.$$

12.3.1 Substitution and Antiderivatives

To apply the method of substitution, we start with an integral whose integrand is a function the independent variable (x) which appears to involve a composition (suggesting a chain rule). Define u to be the formula in the composition and compute u'. We then substitute $dx = \dfrac{du}{u'}$ in the integral and attempt to rewrite the entire integrand in terms of only u. We then find antiderivatives in terms of u and express the result in terms of the orignal variable.

Example 12.3.1 Use the method of substitution to find $\displaystyle\int e^{3x}\,dx$.

Solution. The integrand e^{3x} involves composition with $u = 3x$. This is our substitution variable. Because $u' = 3$, we have $du = 3dx$ so that $dx = \dfrac{du}{3}$. We rewrite the integral in terms of the substitution variable u. After antidifferentiation using the variable u, we back-substitute our original formula for $u = 3x$. The work is shown below.

$$\begin{aligned}\int e^{3x}\,dx \qquad & u = 3x \\ & du = 3\,dx \\ &= \int e^u \cdot \frac{du}{3} \\ &= \int \frac{1}{3}e^u du \\ &= \frac{1}{3}e^u + C \\ &= \frac{1}{3}e^{3x} + C\end{aligned}$$

□

Example 12.3.2 Use the method of substitution to find $\int \sqrt{3x+5}\,dx$.

Solution. The integrand $\sqrt{3x+5} = (3x+5)^{1/2}$ involves composition with $u = 3x+5$. Because $u' = 3$, we have $du = 3dx$ and $dx = \dfrac{du}{3}$. We rewrite the integral in terms of u, find an antiderivative, and then back-substitute to find a formula in terms of x.

$$\begin{aligned}\int \sqrt{3x+5}\,dx \qquad & \begin{matrix} u = 3x+5 \\ du = 3\,dx \end{matrix} \\ &= \int \sqrt{u}\cdot\frac{du}{3} \\ &= \int \frac{1}{3}u^{1/2}du \\ &= \frac{1}{3}\cdot\frac{2}{3}u^{3/2} + C \\ &= \frac{2}{9}(3x+5)^{3/2} + C\end{aligned}$$

□

Example 12.3.3 Use the method of substitution to find $\int x\sin(x^2)\,dx$.

Solution. The integrand $x\sin(x^2)$ is a product with the composition involving $u = x^2$. We hope that the product is a result of the chain rule. Because $u' = \frac{du}{dx} = 2x$, we have $du = 2x\,dx$ or $dx = \dfrac{du}{2x}$. We rewrite the integral

$$\begin{aligned}\int x\sin(x^2)\,dx \qquad & \begin{matrix} u = x^2 \\ du = 2x\,dx \end{matrix} \\ &= \int x\sin(u)\cdot\frac{du}{2x} \\ &= \int \frac{x}{2x}\sin(u)du \\ &= \int \frac{1}{2}\sin(u)du \\ &= -\frac{1}{2}\cos(u) + C \\ &= -\frac{1}{2}\cos(x^2) + C\end{aligned}$$

This problem relied on the factor x and the formula for $u' = 2x$ having x cancel so that the transformed integral involves only the substitution variable u. □

Example 12.3.4 Use the method of substitution to find $\int \tan(x)\,dx$.

Solution. The integrand $\tan(x)$ can be rewritten as a quotient, or as a product involving a negative power,

$$\tan(x) = \frac{\sin(x)}{\cos(x)} = \sin(x)\cdot(\cos(x))^{-1}.$$

Once we have the negative power, we see the composition variable $u = \cos(x)$. Because $u' = \frac{du}{dx} = -\sin(x)$, we have $du = -\sin(x)\,dx$ or $dx = \dfrac{-du}{\sin(x)}$. We

rewrite the integral

$$\begin{aligned}\int \tan(x)\,dx &= \int \sin(x)(\cos(x))^{-1}\,dx \qquad \begin{matrix} u=\cos(x) \\ du=-\sin(x)\,dx \end{matrix} \\ &= \int \sin(x)u^{-1}\frac{-du}{\sin(x)} \\ &= \int -u^{-1}\,du \\ &= -\ln(|u|)+C \\ &= -\ln(|\cos(x)|)+C\end{aligned}$$

□

Sometimes, after substitution, the integrand still involves the original variable. If the formula can be rewritten using only the substitution variable, then we may still be able to find an antiderivative.

Example 12.3.5 Use the method of substitution to find $\int x\sqrt{1-x}\,dx$.

Solution. The integrand $x\sqrt{1-x} = x(1-x)^{1/2}$ is a product with the composition involving $u = 1-x$. Because $u' = \frac{du}{dx} = -1$, we have $du = -dx$ or $dx = -du$. We rewrite the integral

$$\begin{aligned}\int x\sqrt{1-x}\,dx &= \int xu^{1/2}\cdot -du \\ &= \int -xu^{1/2}du\end{aligned}$$

If we start with the substitution equation $u = 1-x$ and solve for x, we find $x = 1-u$ and can use this substitution in the integral.

$$\begin{aligned}\int x\sqrt{1-x}\,dx &= \int -xu^{1/2}du \\ &= \int -(1-u)u^{1/2}du\end{aligned}$$

As currently written in a product, the antiderivative can not be found. However, if we multiply this out we can find antiderivatives using the power rule.

$$\begin{aligned}\int x\sqrt{1-x}\,dx &= \int -(1-u)u^{1/2}du \\ &= \int -u^{1/2}+u^{3/2}du \\ &= -\frac{2}{3}u^{3/2}+\frac{2}{5}u^{5/2}+C \\ &= -\frac{2}{3}(1-x)^{3/2}+\frac{2}{5}(1-x)^{5/2}+C\end{aligned}$$

□

The method of substitution does not work (or at least does not help) if the transformed integral is no closer to finding an antiderivative than the original.

Example 12.3.6 Use the method of substitution to rewrite $\int e^{-x^2}\,dx$.

Solution. The integrand e^{-x^2} has a composition involving $u = x^2$ and $u' =$

$2x$. For $x > 0$ we have the back-substitution

$$u = x^2 \quad \Leftrightarrow \quad x = \sqrt{u}.$$

The method of substitution allows us to rewrite this integral.

$$\begin{aligned} \int e^{-x^2}\,dx \qquad & \begin{array}{c} u = x^2 \\ du = 2x\,dx \end{array} \\ &= \int e^{-u}\frac{du}{2x} \\ &= \int \frac{1}{2\sqrt{u}} e^{-u}\,du \end{aligned}$$

While these integrals are equivalent for $x > 0$, the new integral is no easier to evaluate than the original. It happens that this integral does not have an elementary antiderivative formula. □

12.3.2 Substitution and Definite Integrals

When using definite integrals, the Fundamental Theorem of Calculus allows us to compute a definite integral as the change in an antiderivative. If the method of substitution is used, our antiderivative will be a function of the substitution variable u which is a function of the independent variable. Rather than rewrite the antiderivative in terms of the original variable and then compute the change of the antiderivative, we can compute the change in the antiderivative in terms of the variable u.

Suppose that $F(x)$ is an antiderivative of $f(x)$. Now, suppose that u is a function of x so that $u(a) = c$ and $u(b) = d$. If we have an integral involving composition and the chain rule, we find

$$\begin{aligned} \int_a^b f(u(x))u'(x)dx &\overset{\text{FTC}}{=} [F(u(x))]_a^b \\ &= F(u(b)) - F(u(a)) = F(d) - F(c). \end{aligned}$$

This is identical to the integral we would get for the related definite integral

$$\int_c^d f(u)\,du \overset{\text{FTC}}{=} \Big[F(u)\Big]_c^d = F(d) - F(c).$$

Consequently, using the method of substitution on a definite integral can be performed by changing the limits of integration to the values of the substitution variable.

Example 12.3.7 Compute $\displaystyle\int_1^3 (2x+1)^4\,dx$.

Solution. The substitution variable is $u = 2x+1$. When $x = 1$, $u = 2(1)+1 = 3$, and when $x = 3$, $u = 2(3)+1 = 7$. The substitution step involves $u' = \frac{du}{dx} = 2$ so that $du = 2dx$ or $dx = \dfrac{du}{2}$. In order to keep track of whether the limit of integration refers to x or u, we need to clearly indicate this when both variables are involved.

$$\begin{aligned} \int_1^3 (2x+1)^4\,dx \qquad & \begin{array}{ll} u = 2x-1 & x = 1 \Rightarrow u = 3 \\ du = 2\,dx & x = 3 \Rightarrow u = 7 \end{array} \\ &= \int_{x=1}^{x=3} u^4\,\frac{du}{2} \end{aligned}$$

$$
\begin{aligned}
&= \int_3^7 \frac{1}{2}u^4\,du \\
&\overset{\text{FTC}}{=} \left[\frac{1}{10}u^5\right]_3^7 \\
&= \frac{1}{10}(7^5) - \frac{1}{10}(3^5) \\
&= \frac{16564}{10} = \frac{8282}{5}
\end{aligned}
$$

□

Sometimes the substitution variable is a decreasing function of the independent variable. This will cause the apparent order of the limits to reverse. Be careful that the limits of integration remain in the same starting and ending position as the original.

Example 12.3.8 Compute $\displaystyle\int_3^4 \frac{x\,dx}{25-x^2}$.

Solution. The composition may not be apparent until we think of division as multiplication by a negative power:

$$\frac{x}{25-x^2} = x(25-x^2)^{-1}.$$

This suggests a substitution $u = 25 - x^2$.

$$
\begin{aligned}
\int_3^4 \frac{x\,dx}{25-x^2} \qquad & u = 25-x^2 \quad x = 3 \Rightarrow u = 16 \\
& du = -2x\,dx \quad x = 4 \Rightarrow u = 9 \\
&= \int_{x=3}^{x=4} \frac{x}{u}\frac{du}{-2x} \\
&= \int_{16}^{9} -\frac{1}{2}\frac{du}{u} \\
&\overset{\text{FTC}}{=} \left[-\frac{1}{2}\ln(|u|)\right]_{16}^{9} \\
&= -\frac{1}{2}\ln(9) - -\frac{1}{2}\ln(16) = \frac{1}{2}(\ln(16)-\ln(9)) \\
&= \ln\left(\sqrt{\frac{16}{9}}\right) = \ln(\frac{4}{3})
\end{aligned}
$$

□

12.4 Limits Involving Infinity

We have focused on limits of functions that correspond to points. That is, we have looked at functions that approach a specific value in the output as the input variable approaches a certain value in or at the edge of the domain. In this section, we will consider examples where limits inform us about the behavior of a function as the input or output grow without bound.

12.4.1 Vertical Asymptotes and Infinite Discontinuities

An **asymptote** is a curve (most commonly a line) that a graph approaches. The two most important asymptotes are vertical asymptotes and horizontal asymptotes. In order to classify each of these, we need to introduce a new type of limit statement.

The mathematical statement

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

means that the value of $f(x)$ essentially increases without bound for any sequence of values from the domain $x_n \downarrow a$. More precisely, for any value M (no matter how large), the sequence of values $f(x_n)$ must eventually exceed M, $f(x_n) > M$ for all n, eventually.

Definition 12.4.1 Infinite Limit. The mathematical statement

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

formally represents the following statement: Given any M, there exists a value $\delta > 0$ such that $f(x) > M$ for every $x \in (a, a + \delta)$.

The mathematical statement

$$\lim_{x \to a+} f(x) = -\infty$$

formally represents the following statement: Given any M, there exists a value $\delta > 0$ such that $f(x) < M$ for every $x \in (a, a + \delta)$.

Similar definitions for left limits involve an interval to the left of a, $(a-\delta, a)$. Two-sided limits require left- and right-limits agree. Otherwise, the two-sided limit does not exist. $\Diamond$

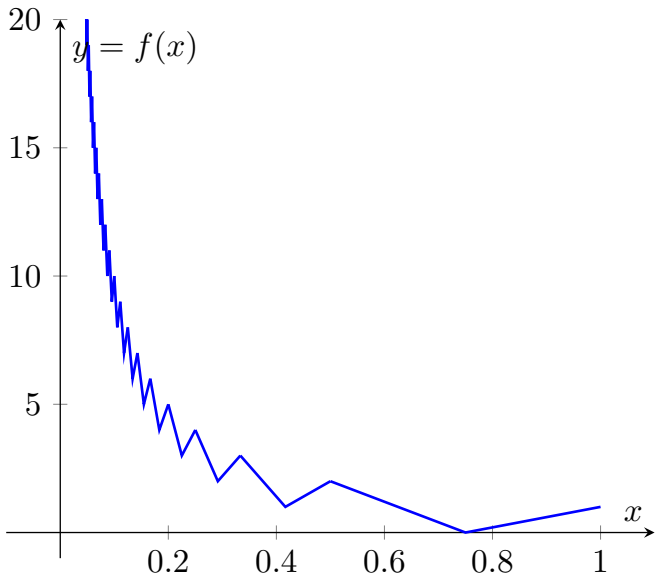
The definition of an infinite limit allows for the possibility that the function might rise and fall so long as overall the function ultimately is rising above every number imaginable.

Example 12.4.2 Consider the function that is formed by joining line segments that alternately go up and down over shorter and shorter intervals given in the graph below. The peaks in the graph are given by the sequence of points defined by

$$\{(\frac{1}{n}, n) : n = 1, 2, 3, \ldots\}$$

while the minimum points are defined halfway between these points by

$$\{(\frac{1}{2}(\frac{1}{n} + \frac{1}{n+1}), n - 1) : n = 1, 2, 3, \ldots\}.$$



If we considered values of x approaching 0 from the right, $x \to 0+$, the values of $f(x)$ might alternately go up and down. Overall, the value of $f(x)$ increases without bound because the graph will eventually surpass every positive real number. Consequently, this function has a limit

$$\lim_{x \to 0+} f(x) = +\infty.$$

□

For algebraic functions, infinite limits occur when the formula involves division such that the numerator has a non-zero limit and the denominator gets smaller and smaller. Dividing a number by an infinitely small value results in an infinitely large value. However, the denominator needs to approach zero monotonically as repeatedly alternating between positive and negative will make the limit not exist. The limit is either $+\infty$ or $-\infty$ depending on which signs are involved.

Theorem 12.4.3 Infinite Limits from Division by Zero. *Given $f(x)$ defined as a quotient $f(x) = \frac{p(x)}{q(x)}$ such that $p(x) \to L$ and $q(x) \to 0$ as $x \to a+$. Then $f(x)$ is unbounded as $x \to a+$ with limits determined by the signs of $p(x)$ and $q(x)$ as follows.*

- *If $p(x) \to L > 0$ and $q(x) \to 0+$, then $\lim_{x \to a+} f(x) = +\infty$.*
- *If $p(x) \to L > 0$ and $q(x) \to 0-$, then $\lim_{x \to a+} f(x) = -\infty$.*
- *If $p(x) \to L < 0$ and $q(x) \to 0+$, then $\lim_{x \to a+} f(x) = -\infty$.*
- *If $p(x) \to L < 0$ and $q(x) \to 0-$, then $\lim_{x \to a+} f(x) = +\infty$.*
- *If $q(x)$ changes sign infinitely many times as $x \to a+$, then the limit does not exist.*

We apply the theorem for rational functions by identifying points where the formula involves division by zero, identifying all removable discontinuities, and then determining the sign of the function immediately to the left and right of each infinite discontinuity. Each infinite discontinuity corresponds to a **vertical asymptote**.

Example 12.4.4 Classify all of the discontinuities of $f(x) = \dfrac{x^3 - 9x}{x^4 - x^3 - 6x^2}$.

Solution. Discontinuities occur when the denominator $q(x) = x^4 - x^3 - 6x^2$ equals zero. We solve for these points by factoring the denominator.

$$\begin{aligned} q(x) &= x^4 - x^3 - 6x^2 \\ &= x^2(x^2 - x - 6) \\ &= x^2(x-3)(x+2) \end{aligned}$$

So there are discontinuities at $x = 0$, $x = 3$ and at $x = -2$.

We see if the discontinuities are removable by factoring the numerator $p(x) = x^3 - 9x$ and seeing which factors might cancel.

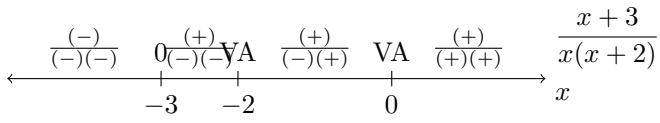
$$\begin{aligned} f(x) &= \frac{x^3 - 9x}{x^4 - x^3 - 6x^2} \\ &= \frac{x(x^2 - 9)}{x^2(x-3)(x+2)} \\ &= \frac{x(x-3)(x+3)}{x^2(x-3)(x+2)} \\ &= \frac{(x+3)}{x(x+2)}, \quad x \neq 3. \end{aligned}$$

The discontinuity at $x = 3$ is removable. The nonremovable discontinuities at $x = 0$ and $x = -2$ will be infinite discontinuities corresponding to vertical asymptotes.

We finish classifying the removable discontinuity by evaluating the limit. This limit will be the output value of the simplified (and continuous) formula:

$$\lim_{x\to 3} f(x) = \lim_{x\to 3} \frac{x+3}{x(x+2)} = \frac{6}{3(5)} = \frac{2}{5}.$$

The infinite discontinuities are analyzed by determining if the unbounded growth is positive or negative. This is usually different on each side, so we check the signs. Because we already factored $f(x)$, we can use the factors to quickly determine a sign analysis summary.



We will now interpret the signs as we evaluate the limits at the discontinuities. First consider $x \to -2$. If we attempt to evaluate the limit directly, we find

$$\lim_{x\to -2} \frac{x+3}{x(x+2)} \to \frac{1}{0},$$

and this division by zero is precisely the hallmark of infinite limits. Our sign analysis summary shows that the denominator has a positive sign $(-)(-) = (+)$ for $x \to -2-$ and has a negative sign $(-)(+) = (-)$ for $x \to -2+$. Consequently, our one-sided limits give

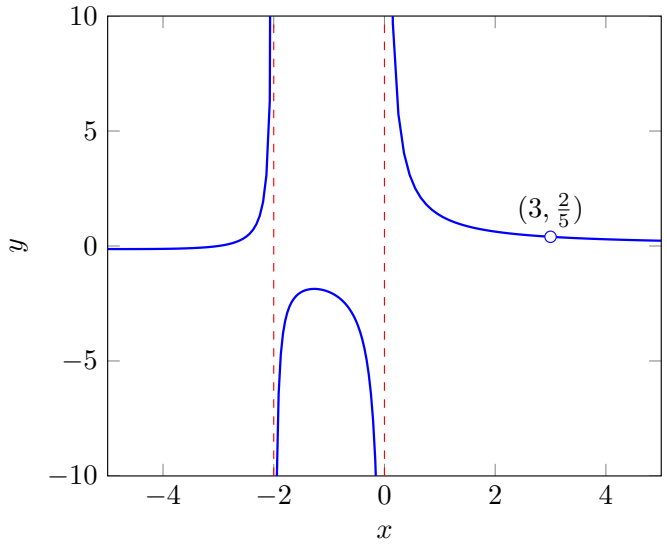
$$\begin{aligned} \lim_{x\to -2-} f(x) &= \lim_{x\to -2-} \frac{x+3}{x(x+2)} \to \frac{1}{0+} = +\infty \\ \lim_{x\to -2+} f(x) &= \lim_{x\to -2+} \frac{x+3}{x(x+2)} \to \frac{1}{0-} = -\infty \end{aligned}$$

Using the $0+$ and $0-$ is a notation that reminds us which sign the denominator has as it approaches zero. We can then use arguments about sign to determine if the resulting infinity is positive or negative. Because these limits are opposite, the two-sided limit does not exist.

The work associated with the limits at $x \to 0$ is summarized below.

$$\lim_{x\to 0-} f(x) = \lim_{x\to 0-} \frac{x+3}{x(x+2)} \to \frac{3}{0-} = -\infty$$
$$\lim_{x\to 0+} f(x) = \lim_{x\to 0+} \frac{x+3}{x(x+2)} \to \frac{1}{0+} = +\infty$$
$$\lim_{x\to 0} f(x) \text{ does not exist.}$$

The graph $y = f(x)$ is given below. Make note how the infinite limits correspond to the vertical asymptotes $x = -2$ and $x = 0$. Be sure to connect in your mind how the sign of the infinite limit corresponds to the direction in which the graph of the function approaches the asymptote.



□

The following example does a similar analysis, but keeps comments to a minimum to demonstrate what work might be normally expected.

Example 12.4.5 Classify the discontinuities of $f(x) = \dfrac{x^2-4}{x^4-7x^3+10x^2}$.

Solution. Factor $f(x)$:

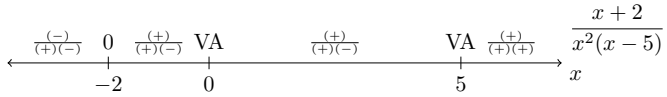
$$\begin{aligned} f(x) &= \frac{x^2-4}{x^4-7x^3+10x^2} \\ &= \frac{(x+2)(x-2)}{x^2(x^2-7x+10)} \\ &= \frac{(x+2)(x-2)}{x^2(x-2)(x-5)} \\ &= \frac{x+2}{x^2(x-5)}, \quad x \neq 2 \end{aligned}$$

There is a removable discontinuity at $x = 2$ with limit

$$\lim_{x\to 2} f(x) = \lim_{x\to 2} \frac{x+2}{x^2(x-5)} = \frac{4}{4(-3)} = -\frac{1}{3}.$$

There are infinite discontinuities at $x = 0$ and $x = 5$ corresponding to vertical asymptotes.

Sign analysis:



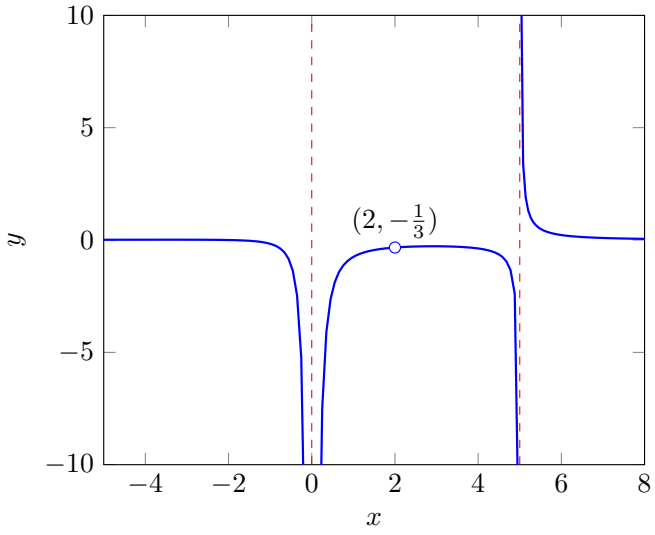
The limits associated with the vertical asymptote $x = 0$:

$$\lim_{x\to 0-} f(x) = \lim_{x\to 0-} \frac{x+2}{x^2(x-5)} = \frac{2}{0-} = -\infty,$$
$$\lim_{x\to 0+} f(x) = \lim_{x\to 0+} \frac{x+2}{x^2(x-5)} = \frac{2}{0-} = -\infty,$$
$$\lim_{x\to 0} f(x) = -\infty.$$

The limits associated with the vertical asymptote $x = 5$:

$$\lim_{x\to 5-} f(x) = \lim_{x\to 5-} \frac{x+2}{x^2(x-5)} = \frac{7}{0-} = -\infty,$$
$$\lim_{x\to 5+} f(x) = \lim_{x\to 5+} \frac{x+2}{x^2(x-5)} = \frac{7}{0+} = +\infty,$$
$$\lim_{x\to 5} f(x) \text{ does not exist.}$$

A graph illustrates the results below.



□

12.4.2 Horizontal Asymptotes and Limits at Infinity

A function has a horizontal asymptote if the function behaves more and more like a constant value for large input values. Horizontal asymptotes often have applications relating to the idea of **saturation**. For example, when food is scarce, the total amount of food an individual eats during a day will be proportional to the amount of food available. However, there comes a point where increasing the amount of food available does not lead to continuing increase in the amount of food eaten per individual. Consumption saturates.

A common misconception by students is that a function does not cross a

horizontal asymptote. This likely results from students applying something they heard about vertical asymptotes and generalizing it to all asymptotes. A function does not cross a vertical asymptote only because functions must obey the vertical line test. If the graph crossed a vertical asymptote, it would need to bend back to approach the asymptote from the other side; that process violates the definition of a function. Horizontal asymptotes can be crossed multiple times (even infinitely many times).

When a function has a horizontal asymptote, we are considering the behavior of the function as the input $x \to +\infty$ or $-\infty$. The value of the limit is the y-value of the horizontal asymptote.

Definition 12.4.6 Limits at Infinity. The mathematical statement

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

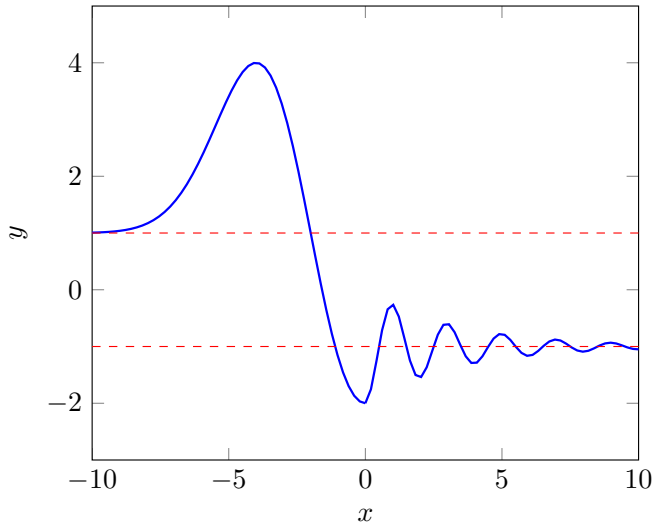
for a real number L means $|f(x_n) - L| \to 0$ for every unbounded increasing sequence $x_n \uparrow \infty$. Formally, this corresponds to the statement: For every $\epsilon > 0$, there exists $M > 0$ so that $|f(x) - L| < \epsilon$ for every $x > M$. ♢

Example 12.4.7 Consider the function illustrated in the graph below. Notice that the function goes above and below the value $y = -1$ but that the size of the difference is shrinking in size as $x \to +\infty$. Consequently, we would say $y = -1$ is a horizontal asymptote and

$$\lim_{x\to\infty} f(x) = -1.$$

In the other direction, notice that the function approaches another horizontal asymptote $y = 1$ as $x \to -\infty$, corresponding to a limit

$$\lim_{x\to-\infty} f(x) = -1.$$



□

For functions defined algebraically, we find limits at infinity by identifying terms that go to zero. These are often identified as being the multiplicative inverse of terms that are unbounded. If $p(x) \to \infty$, then $1/p(x) \to 0$. For algebraic formulas, we can use limit arithmetic involving infinity to compute determinate limits.

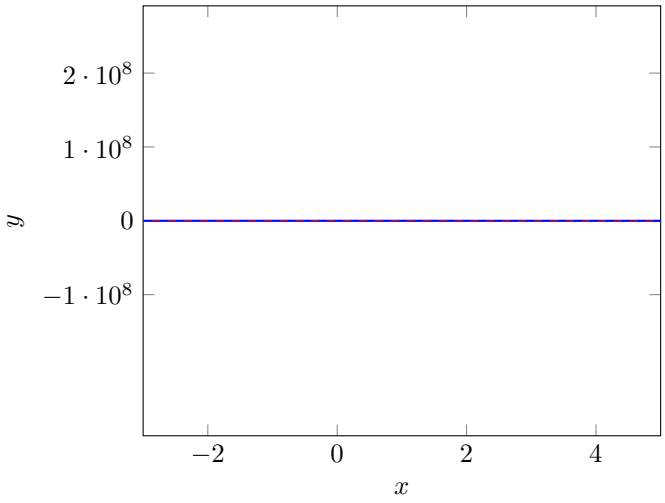
$$\infty^p = \infty, \quad \text{for } p > 0$$
$$\infty^{-p} = \frac{1}{\infty^p} = 0, \quad \text{for } p > 0$$
$$b^{\infty} = \infty, \quad \text{for } b > 1$$
$$b^{-\infty} = 0, \quad \text{for } b > 1$$
$$b^{\infty} = 0, \quad \text{for } 0 < b < 1$$
$$b^{-\infty} = \infty, \quad \text{for } 0 < b < 1$$

Example 12.4.8 Determine the limits at infinity for $f(x) = 4 - 3e^{-2x}$.

Solution. The base e is a number $e > 1$. So we will use $e^{\infty} = \infty$ and $e^{-\infty} = 0$.

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} 4 - 3e^{-2x} = 4 - 3e^{-2(\infty)} = 4 - 3e^{-\infty} = 4 - 0 = 4$$
$$\lim_{x\to-\infty} f(x) = \lim_{x\to-\infty} 4 - 3e^{-2x} = 4 - 3e^{-2(-\infty)} = 4 - 3e^{+\infty} = 4 - \infty = -\infty$$

So $y = 4$ is a horizontal asymptote of $f(x)$ as $x \to +\infty$. There is no horizontal asymptote as $x \to -\infty$ since $f(x) \to -\infty$. A graph is shown below.



□

A limit that appears to have infinities cancel in any way (or zeros cancel in division) is **indeterminate** because the arithmetic of limits does not apply when infinities might cancel. To compute such a limit, must rewrite the formula to eliminate the indeterminate form. When a limit involves infinity, we factor out the term that grows to infinity the fastest and seek to simplify.

Example 12.4.9 Determine the limits at infinity for $f(x) = \dfrac{x^2 - 3x + 1}{4x^2 + 5x + 7}$.

Solution. The numerator and the denominator involve the term x^2 and we know $x^2 \to +\infty$ as $x \to \pm\infty$. This will lead to an indeterminate form ∞/∞. So we factor out x^2 (the fastest growing power) from the numerator and denominator and simplify.

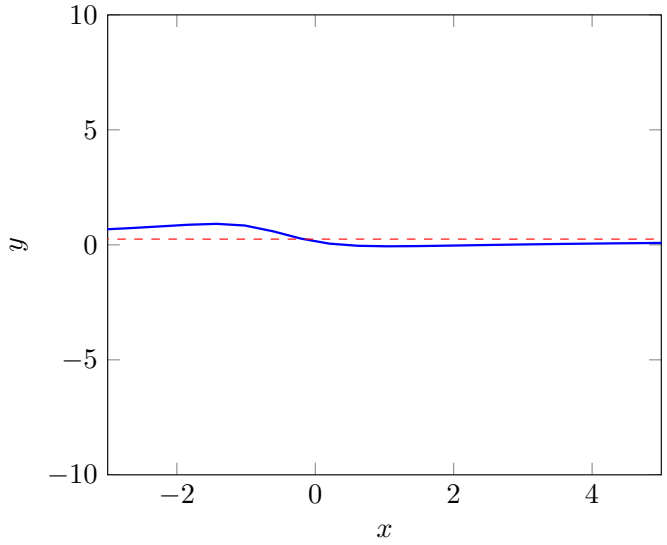
$$f(x) = \frac{x^2 - 3x + 1}{4x^2 + 5x + 7}$$

$$= \frac{x^2(1 - \frac{3}{x} + \frac{1}{x^2})}{x^2(4 + \frac{5}{x} + \frac{7}{x^2})}$$
$$= \frac{1 - \frac{3}{x} + \frac{1}{x^2}}{4 + \frac{5}{x} + \frac{7}{x^2}}$$

This new representation involves terms that go to zero.

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} \frac{1 - \frac{3}{x} + \frac{1}{x^2}}{4 + \frac{5}{x} + \frac{7}{x^2}} = \frac{1 - 0 + 0}{4 + 0 + 0} = \frac{1}{4}$$
$$\lim_{x\to-\infty} f(x) = \lim_{x\to-\infty} \frac{1 - \frac{3}{x} + \frac{1}{x^2}}{4 + \frac{5}{x} + \frac{7}{x^2}} = \frac{1 - 0 + 0}{4 + 0 + 0} = \frac{1}{4}$$

So $y = \frac{1}{4}$ is a horizontal asymptote of $f(x)$ on both sides.



□

12.5 Continuous Functions

12.5.1 Continuity

Recall our definition of continuity for a function at a single point.

Definition 12.5.1 Continuity at a Point. A function f is **continuous** at a if

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

◊

The single equation captures the full definition because for the equation to be true, the limit must exist and the value of the function must exist. Also, recall that the function is **right-continuous** if the limit comes from the right ($x \to a^+$) and **left-continuous** if the limit comes from the left ($x \to a^-$).

These ideas allow us to define what we mean by saying that a function is continuous on an interval.

Definition 12.5.2 Continuity on an Interval. A function f is **continuous on an interval** (a, b) if f is continuous at every point $x \in (a, b)$. We can include an endpoint if the limit statement is true coming from within the interval. That is, we include a if

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

and we include b if

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

◊

12.5.2 Definite Integrals and Average Value

When we studied the definite integral, we learned that continuity implies integrability. However, a discontinuous function might still be integrable. For example, the definite integral of a piecewise continuous function with a finite number of jump discontinuities can be computed using the splitting property. The total definite integral would be equal to the sum of the definite integrals on each of the subintervals.

Continuity does guarantee something stronger than integrability. It guarantees that the function attains its average value over an interval. To make this precise, we first need to define the average value.

Definition 12.5.3 Average Value of a Function. The **average value** of a function f on an interval $[a, b]$, denoted $\langle f \rangle_{[a,b]}$, is defined as

$$\langle f \rangle_{[a,b]} = \frac{1}{b-a} \int_a^b f(x)\, dx,$$

so long as f is integrable on $[a, b]$. ◊

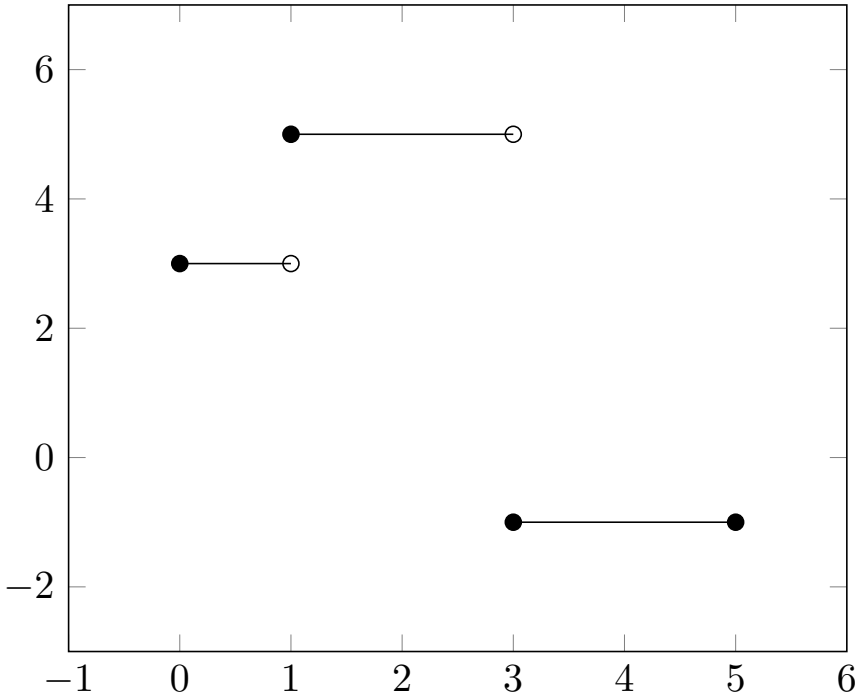
The average value is defined as the value of a constant function that has the same definite integral over the interval:

$$\int_a^b \langle f \rangle_{[a,b]}\, dx = \langle f \rangle_{[a,b]} \cdot (b-a) = \int_a^b f(x)\, dx..$$

Example 12.5.4 The figure below illustrates a simple function $f(x)$ defined

on the interval $[0, 5]$,

$$f(x) = \begin{cases} 3, & 0 \leq x < 1, \\ 5, & 1 \leq x < 3, \\ -1, & 3 \leq x \leq 5. \end{cases}$$

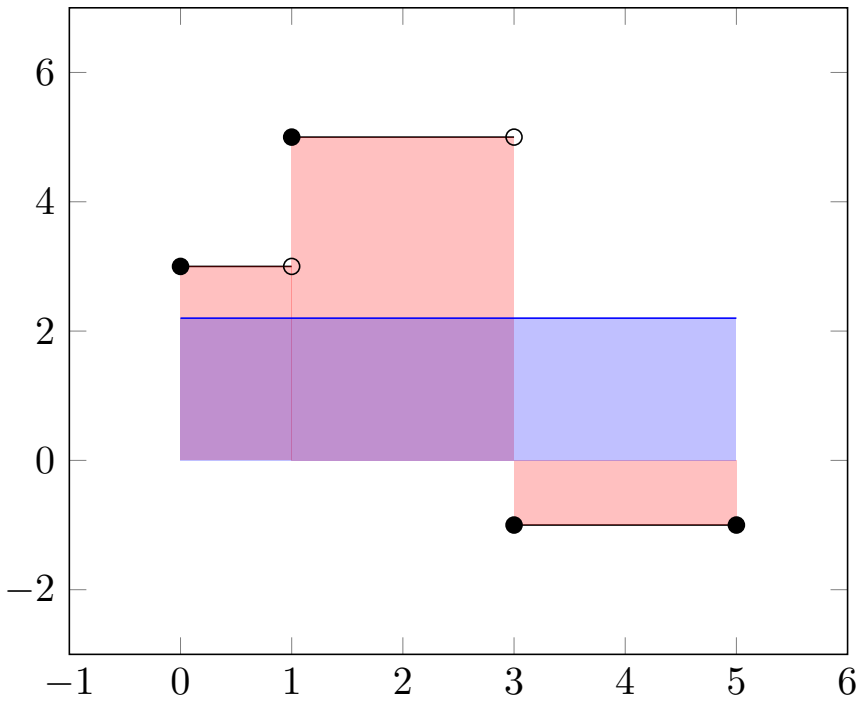


The definite integral equals the sum of the signed areas,

$$\int_0^5 f(x)\,dx = 3 \cdot 1 + 5 \cdot 2 + -1 \cdot 2 = 11.$$

The average value is equal to this definite integral divided by the width of the interval,

$$\langle f \rangle_{[0,5]} = \frac{1}{5} \int_0^5 f(x)\,dx = \frac{11}{5}.$$



□

Theorem 12.5.5 Mean Value Theorem for Integrals. *Given a function f that is continuous on $[a,b]$, there must exist a value $c \in (a,b)$ such that*

$$f(c) = \langle f \rangle_{[a,b]} = \frac{1}{b-a} \int_a^b f(x)\,dx,$$

or equivalently, $\displaystyle\int_a^b f(x)\,dx = f(c) \cdot (b-a)$.

Proof. Because f is continuous on $[a,b]$, the Extreme Value Theorem guarantees that f attains a minimum value $f(x_{\min})$ and a maximum value $f(x_{\max})$ so that $f(x_{\min}) \le f(x) \le f(x_{\max})$ for all $x \in [a,b]$.

The average value $\langle f \rangle_{[a,b]}$ must be between the minimum and maximum values. The (((Unresolved xref, reference "thm-integral-inequality"; check spelling or use "provisional" attribute)))Integral Bounds theorem guarantees

$$f(x_{\min})(b-a) \le \int_a^b f(x)\,dx \le f(x_{\max})(b-a)$$

which then implies

$$f(x_{\min}) \le \langle f \rangle_{[a,b]} \le f(x_{\max}).$$

By the Intermediate Value Theorem with the interval with end points $x_{\min}$ and $x_{\max}$ (we don't know which is on the left/right), there must be some value c between these points, and so $c \in (a,b)$, for which

$$f(c) = \langle f \rangle_{[a,b]}.$$

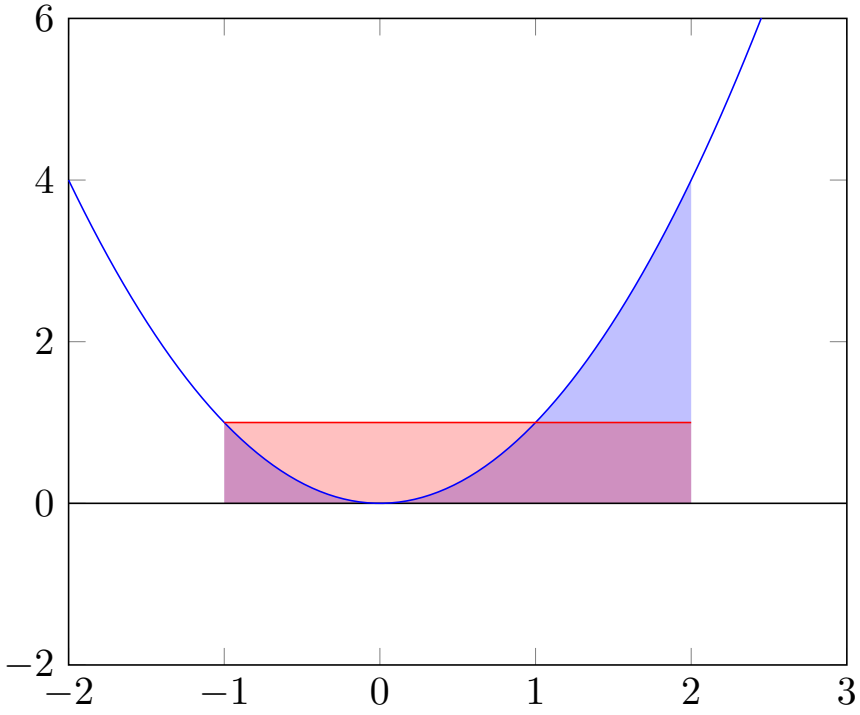
■

In the previous example, f was not continuous and we can see that the graph $y = f(x)$ did not intersect the constant value $\langle f \rangle_{[0,5]}$. The Mean Value Theorem for Integrals guarantees that when the function is continuous, the constant function using the average value must intersect the graph $y = f(x)$.

Example 12.5.6 The function $f(x) = x^2$ is continuous everywhere. The average value on the interval $[-1, 2]$ can be found using the (((Unresolved xref, reference "thm-elementary-definite-integrals"; check spelling or use "provisional" attribute)))elementary accumulation formula for a quadratic rate and the splitting property.

$$\begin{aligned}\langle f\rangle_{[-1,2]} &= \frac{1}{2--1}\int_{-1}^{2} x^2\,dx \\ &= \frac{1}{3}\Big(\int_0^2 x^2\,dx - \int_0^{-1} x^2\,dx\Big) \\ &= \frac{1}{3}\Big(\frac{1}{3}(2^3) - \frac{1}{3}(-1)^3\Big) = \frac{1}{3}\Big(\frac{8}{3}+\frac{1}{3}\Big) = 1\end{aligned}$$

A figure showing the graphs $y = f(x) = x^2$ and $y = \langle f\rangle_{[-1,2]} = 1$ is shown below. The Mean Value Theorem predicted the existence of a point $c \in (-1, 2)$ where $f(c) = \langle f\rangle_{[-1,2]} = 1$, which we can see occurs at $c = 1$.



□

The Mean Value Theorem for Integrals also provides the justification for the Monotonicity Test for Accumulation Functions.

Theorem 12.5.7 Monotonicity Test for Accumulation Functions. *Suppose that $A(x)$ is an accumulation function with corresponding rate function $f(x)$, and suppose that $f(x)$ is continuous on $[a, b]$.*

- *If $f(x) > 0$ for all $x \in (a, b)$, then $A(x)$ is increasing on $[a, b]$.*
- *If $f(x) < 0$ for all $x \in (a, b)$, then $A(x)$ is decreasing on $[a, b]$.*
- *If $f(x) = 0$ for all $x \in (a, b)$, then $A(x)$ is constant on $[a, b]$.*

Proof. Consider any two points $c_1, c_2 \in [a, b]$ with $c_1 < c_2$. Because $A(x)$ is an

accumulation function, by the splitting property of definite integrals,

$$A(c_2) - A(c_1) = \int_{c_1}^{c_2} f(x)\,dx.$$

On the other hand, because f is continuous, the Mean Value Theorem guarantees the existence of a point $c \in (c_1, c_2)$ such that

$$A(c_2) - A(c_1) = \int_{c_1}^{c_2} f(x)\,dx = f(c) \cdot (c_2 - c_1).$$

Now assume that $f(x) > 0$ for all $x \in (a, b)$. Then $f(c) > 0$ and $c_2 - c_1 > 0$, guaranteeing that $A(c_2) - A(c_1) > 0$. That is, $A(c_2) > A(c_1)$. This is what is needed to show that A is increasing.

Next assume that $f(x) < 0$ for all $x \in (a, b)$. Then $f(c) < 0$ while $c_2 - c_1 > 0$, guaranteeing that $A(c_2) - A(c_1) < 0$. That is, $A(c_2) < A(c_1)$, which shows that A is decreasing.

Finally assume that $f(x) = 0$ for all $x \in (a, b)$. Then $f(c) = 0$, implying that $A(c_2) - A(c_1) = 0$. That is, $A(c_2) = A(c_1)$, which shows that A is constant. ■

12.6 Applications Involving Densities

12.6.1 Overview

Having developed the theory of definite integrals and functions defined as the accumulation on increments, we turn our attention to applications of these ideas. One of the most common mathematical applications is the calculation of area of regions bounded by curves. Physically, this is closely related to the calculation of total mass and center of mass. The same calculations are used in statistics to calculate probabilities and averages.

The general setup for many applications involving definite integrals is to think of the total quantity as a sum of the parts. If a region is cut into separate pieces, for example, then the area of the total region should be the sum of the areas of each region measured separately. Quantities that have this property are called **extensive**. A definite integral can be used to compute extensive quantities if we can consider the total as a sum of small increments over a partition of an independent variable.

Theorem 12.6.1 Using Definite Integrals to Compute Extensive Quantities. *Suppose an extensive quantity Q can be subdivided into increments corresponding to a uniform partition of an independent variable x over an interval $[a, b]$. If there is a function $f(x)$ so that on each subinterval $[x_{k-1}, x_k]$ there is some point x_k^* so that*

$$f(x_k^*)\Delta x = \Delta Q_k,$$

then

$$Q = \int_a^b f(x)\,dx.$$

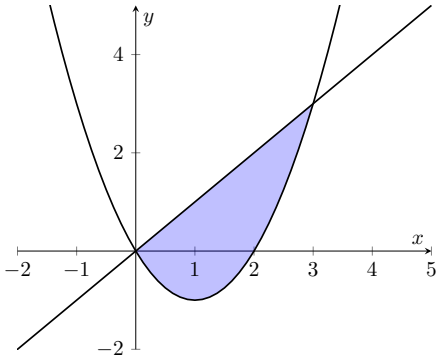
In this context, the rate of accumulation function f used to compute the quantity Q as an integral is often called the **density** of Q with respect to x.

This section will explore the application of definite integrals to compute various extensive quantities. This will require imagining a partition, identifying the independent variable and its corresponding interval for integration, determining the appropriate function used as the density, and setting up the definite integral. Because our emphasis will be on identifying the appropriate definite integral, we will use technology to compute the the resulting value.

12.6.2 Area of Regions in the Plane

When we developed the definite integral of a function f on an interval $[a, b]$, we noted that the integral represented the total signed area over the interval. It was signed because the accumulation of negative values when the function was below the axis was subtracting area. Because area is an extensive quantity, it is an ideal example of a quantity that can be computed using integration.

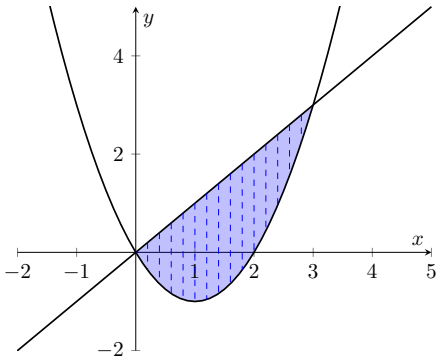
Example 12.6.2 Find the area of the region bounded between $y = x^2 - 2x$ and $y = x$.



Solution. Start by identifying a convenient variable to partition. The region is determined by where the graphs $y = x^2 - 2x$ and $y = x$ intersect. Using substitution and solving, the intersection occurs at the solution to $x^2 - 2x = x$.

$$\begin{aligned} x^2 - 2x &= x \\ x^2 - 3x &= 0 \\ x(x-3) &= 0 \end{aligned}$$

The region is completely contained between $x = 0$ and $x = 3$. We choose our independent variable to be x over the interval $[0, 3]$. Now, imagine a partition of the interval and consider the increments of area over each subinterval.



As the increments Δx are smaller and smaller, the increment of area ΔA will be closely approximated by the width Δx times the vertical distance between $y = x$ and $y = x^2 - 2x$,

$$\Delta A \approx \big(x - (x^2 - 2x)\big)\Delta x.$$

The total area is the sum of the increments, so we can use a definite integral. The height $h(x) = x - (x^2 - 2x) = 3x - x^2$ between the curves acts as the density of area,

$$A = \int_0^3 \big(x - (x^2 - 2x)\big)\, dx = \int_0^3 3x - x^2\, dx.$$

The density is a simple polynomial so that we can compute the value using the elementary accumulation formulas.

$$\begin{aligned} \int_0^3 3x - x^2\, dx &= 3\int_0^3 x\, dx - \int_0^3 x^2\, dx \\ &= 3\big(\frac{1}{2}(3)^2\big) - \big(\frac{1}{3}(3)^3\big) \end{aligned}$$

$$= \frac{27}{2} - 9$$
$$= \frac{9}{2}$$

Except for such simple problems, we can use technology to compute or approximate the value of the integrals. The SageMath engine attempts to compute integrals exactly using the `integrate` command, which uses the following syntax.

```
integrate(3*x-x^2, [x,0,3])
```

```
9/2
```

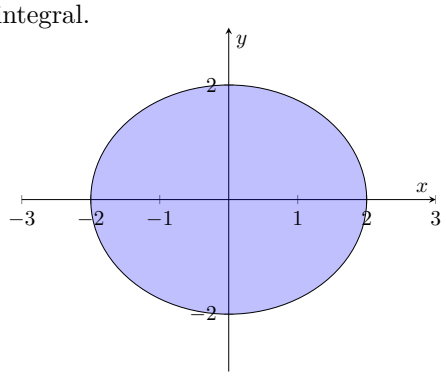
When SageMath is unable to do the exact calculation, we can still do a numerical approximation using the `numerical_integral` command.

```
numerical_integral(3*x-x^2, 0, 3)
```

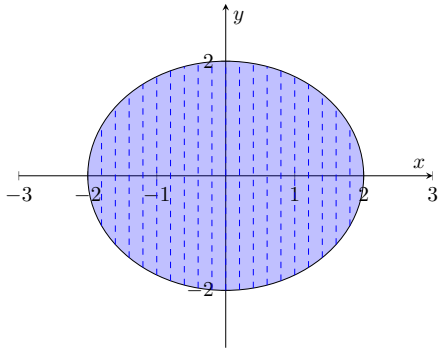
```
(4.5, 4.9960036108132044e-14)
```

The result gives an approximate answer along with an estimated error bound. In this case, we find $A = 4.5 \pm 4.996 \times 10^{-14}$. □

Example 12.6.3 Express the area of a circle with center $(0, 0)$ and radius $r = 2$ as a definite integral.



Solution. The circle is clearly between the lines $x = -2$ and $x = 2$ so that we can imagine a partition with variable x on the interval $[-2, 2]$. The increments of area ΔA correspond to thin vertical strips with width Δx (from the partition) and a height computed as the distance from the top of the circle to the bottom of the circle.



The equation of the circle is

$$x^2 + y^2 = 4.$$

To find the height of the increments, we need to know the two y-values for each x-value,

$$y = \pm\sqrt{4 - x^2}.$$

The height is the difference between the values,

$$h(x) = (\sqrt{4 - x^2}) - (-\sqrt{4 - x^2}) = 2\sqrt{4 - x^2}.$$

Consequently, the area of the circle is defined by the integral

$$A = \int_{-2}^{2} h(x)\,dx = \int_{-2}^{2} 2\sqrt{4 - x^2}.$$

Computational tools can compute this value, which is consistent with the known formula

$$A = \pi r^2 = \pi(2)^2 = 4\pi.$$

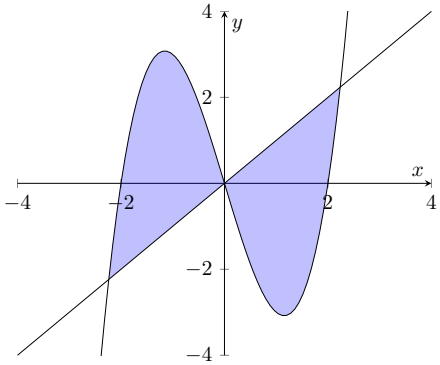
```
integrate(2*sqrt(4-x^2), [x,-2,2])
```

```
4*pi
```

□

When curves cross multiple times, we may need to compute the area of individual regions.

Example 12.6.4 Find the area bounded by the graphs $y = x$ and $y = x^3 - 4x$.



Solution. We start by identifying the points of intersection of the curves by solving the equation $x^3 - 4x = x$.

$$x^3 - 4x = x$$
$$x^3 - 5x = 0$$
$$x(x^2 - 5) = 0$$

One solution is $x = 0$, corresponding to the intersection at $(x, y) = (0, 0)$. Two other solutions come from $x^2 = 5$ at $x = \pm\sqrt{5}$. The total area consists of two regions, the first with $x \in [-\sqrt{5}, 0]$ and the second with $x \in [0, \sqrt{5}]$.

On the first interval $[-\sqrt{5}, 0]$, the height of increments is given by

$$h(x) = (x^3 - 4x) - x = x^3 - 5x$$

because the cubic polynomial is the top curve. So the area over the interval $[-\sqrt{5}, 0]$ is computed by

$$A_1 = \int_{-\sqrt{5}}^{0} x^3 - 5x\,dx.$$

In order to use the elementary accumulation formulas, the integral needs to start at $x = 0$, so we reverse the order of integration and change the sign.

$$\begin{aligned} A_1 &= \int_{-\sqrt{5}}^{0} x^3 - 5x\,dx \\ &= -\int_{0}^{-\sqrt{5}} x^3 - 5x\,dx \\ &= -\int_{0}^{-\sqrt{5}} x^3\,dx + 5\int_{0}^{-\sqrt{5}} x\,dx \\ &= -\left(\frac{1}{4}(-\sqrt{5})^4\right) + 5\left(\frac{1}{2}(-\sqrt{5})^2\right) \\ &= -\frac{25}{4} + \frac{25}{2} \\ &= \frac{25}{4} \end{aligned}$$

On the second interval $[0, \sqrt{5}]$, the height of increments is given by

$$h(x) = x - (x^3 - 4x) = 5x - x^3$$

because now the cubic polynomial is the bottom curve. The corresponding area is

$$A_2 = \int_{0}^{\sqrt{5}} 5x - x^3\,dx$$

which has a value

$$\begin{aligned} A_2 &= \int_{0}^{\sqrt{5}} 5x - x^3\,dx \\ &= 5\int_{0}^{\sqrt{5}} x\,dx - \int_{0}^{\sqrt{5}} x^3\,dx \\ &= 5\left(\frac{1}{2}(\sqrt{5})^2\right) - \left(\frac{1}{4}(\sqrt{5})^4\right) \\ &= \frac{25}{2} - \frac{25}{4} \\ &= \frac{25}{4} \end{aligned}$$

As we should expect from symmetry, the two areas are equal $A_1 = A_2$.

The total area of the region is

$$A = A_1 + A_2 = \frac{25}{4} + \frac{25}{4} = \frac{25}{2}.$$

If we were to think of the distance between the two curves in terms of the absolute value, the integral could be computed over a single interval,

$$A = \int_{-\sqrt{5}}^{\sqrt{5}} \left|x - (x^3 - 4x)\right|\,dx = \int_{-\sqrt{5}}^{\sqrt{5}} \left|5x - x^3\right|\,dx.$$

Using SageMath, the absolute value prevents an exact integral.

```
A1 = integrate(x^3-5*x, [x,-sqrt(5),0])
A2 = integrate(5*x-x^3, [x,0,sqrt(5)])
show(A1)
show(A2)
show(A1+A2)
```

```
25/4
25/4
25/2
```

```
integrate(abs(5*x-x^3), [x,-sqrt(5),sqrt(5)])
```

```
integrate(abs(5*x-x^3), [x,-sqrt(5),sqrt(5)])
```

```
numerical_integral(abs(5*x-x^3), -sqrt(5),sqrt(5))
```

```
(12.5, 1.3855583347321954e-13)
```

□

12.6.3 Density and Mass

12.6.4 Summary

-

12.6.5 Exercises

1.

12.7 Functions Defined by Their Rates

When we learned about definite integrals, we learned that the definite integral $\int_a^b f(x)\,dx$ computes the total change in a quantity that depends on x when x changes from a to b and where $f(x)$ represents the rate of change of that quantity with respect to x. We have worked from an intuitive idea of rate of change with respect to time using concepts like velocity being the rate of change of position or flow rates (as in gallons per minute) as being the rate of change of volume with respect to time. We are now preparing to learn more specifically what we mean by rate of change, namely introducing the concept of the **derivative**.

12.7.1 Describing Function Behavior

One of the consequences of the properties of definite integrals is that we can describe certain behaviors of a quantity in terms of the properties of the rate of change. For sequences, we learned that the properties of a sequence are determined from the increments. That is, a sequence is

- increasing when its increments are positive,
- decreasing when its increments are negative,
- concave up when its increments are increasing,
- concave down when its increments are decreasing.

(See Theorem 13.1.7 and Definition 13.1.8.) We learned analogous properties of functions defined by an accumulation of changes defined by a rate of change. That is, for a quantity Q with a rate of change R:

- Q is **increasing** when its rate of change is R is positive,
- Q is **decreasing** when its rate of change R is negative.

(See Theorem 3.4.5.) We define concavity for functions analogously.

Definition 12.7.1 Concavity of Functions. A quantity Q that is a function of an independent variable x with a corresponding rate of change R that is itself a function of x has **concavity** that is determined by whether the rate is increasing or decreasing. Suppose that $I = (a, b)$ is an interval.

- Q is **concave up** on I if R is increasing on I.
- Q is **concave down** on I if R is decreasing on I.

$\Diamond$

Concavity is closely related to the concept of acceleration (by which we also include the idea of deceleration). A constant rate of change leads to a linear relation. On a graph, this is a straight line. Concavity refers to a rate of change that is itself changing. This is acceleration. In physics, acceleration is caused by a force, so we can think of concavity as the effect on an object in the presence of a force.

Suppose that a quantity has a positive rate of change (increasing) and is also concave up (an increasing rate of change). This would be like a car moving forward (positive rate) with a rocket pushing it forward (positive acceleration).

The result would be that the car continues to go faster (increasing rate), covering ever increasing distances per unit time. A graph of the position would be rising (increasing) and bending up (concave up).

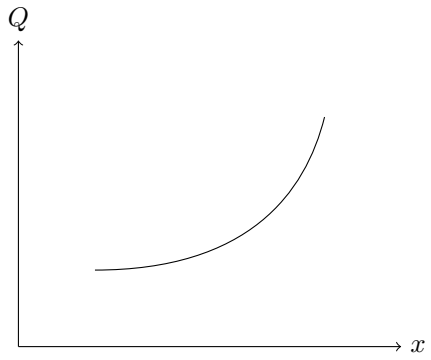


Figure 12.7.2 Q is an increasing and concave up function of x. The rate is positive and increasing.

Next, suppose that a quantity has a positive rate of change (increasing) but is concave down (a decreasing rate of change). This would be like a car moving forward (positive rate) but with a rocket in reverse (negative acceleration). The car would still be moving forward, but the rocket is slowing it down. A graph of position in this case would be rising (increasing) but bending down (concave down).

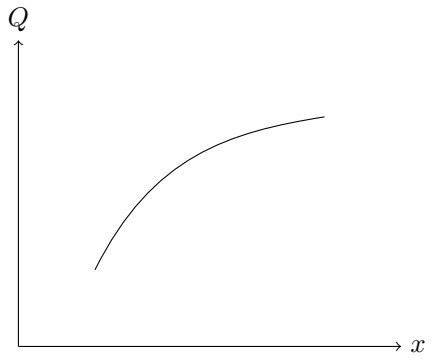


Figure 12.7.3 Q is an increasing and concave down function of x. The rate is positive but decreasing.

If the rocket continues to exert a negative force, there will be a moment when all of the forward momentum is gone and then the car begins to go backwards. Consequently, we learn that a quantity that is concave down can switch from increasing to decreasing. Graphically, this is exactly what a parabola that opens down does. However, if the rocket is gradually reduced, we might be able to slow the car down without ever changing direction.

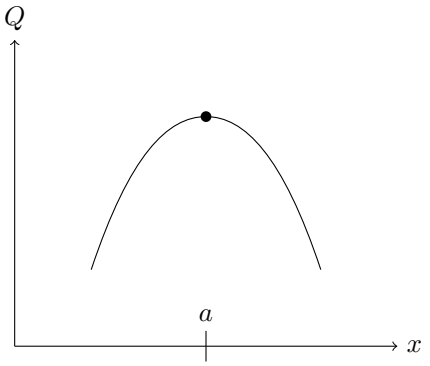


Figure 12.7.4 Q is a concave down function of x that is increasing for $x < a$ and decreasing for $x > a$.

Similar behaviors might be described for negative rates of change. This would correspond to a car that is already going backwards. Being concave up (increasing rate of change) corresponds to a positive acceleration (rocket force), which in this case is opposite the motion and would serve to slow the car down maybe to the point of reversing direction. Graphically, this corresponds to a dropping graph that is bending up (moving toward flat). Being concave down (decreasing rate of change) corresponds to negative acceleration (rocket force) which now is the same direction as the motion. This would cause the car to speed up (in the negative direction). Graphically, being concave down corresponds to a graph that is bending down and growing ever steeper.

Example 12.7.5 Suppose V measures the volume of water (liters) in a container and that V is a function of time t (minutes) such that the rate of change (liters per minute) is also a function of time defined by

$$\frac{dV}{dt} = R(t) = 10 - 0.25t.$$

Describe the behavior of V and sketch a representative graph.

Solution. The rate of change of the volume in the container, $R(t)$ determines the behavior of the volume. Because R has a negative slope $m = -0.25$, the rate is decreasing. This tells us that the volume is a concave down function. Solving the inequalities $R(t) > 0$ and $R(t) < 0$ will allow us to see when the rate is positive or negative, which will imply when the volume is increasing or decreasing, respectively.

The inequalities are solved by solving the equation $R(t) = 0$ and then testing the inequalities in the resulting intervals.

$$\begin{aligned} 10 - 0.25t &= 0 \\ -0.25t &= -10 \\ t &= \frac{-10}{-0.25} = 40 \end{aligned}$$

Testing the sign of $R(t)$ when $t < 40$, we find, for example $R(20) = 10 - 0.25(20) = 5$, that the rate is positive. Consequently, V is increasing when $t < 40$. On the other hand, testing the sign of $R(t)$ when $t > 40$, such as $R(60) = 0 - 0.25(60) = -5$, we find that the rate is negative so that the volume is decreasing when $t > 40$.

The graph of $R(t)$, shown below, is consistent with these analyses. The graph is decreasing (corresponding to the negative slope), above the axis for $t < 40$ and below the axis for $t > 40$.

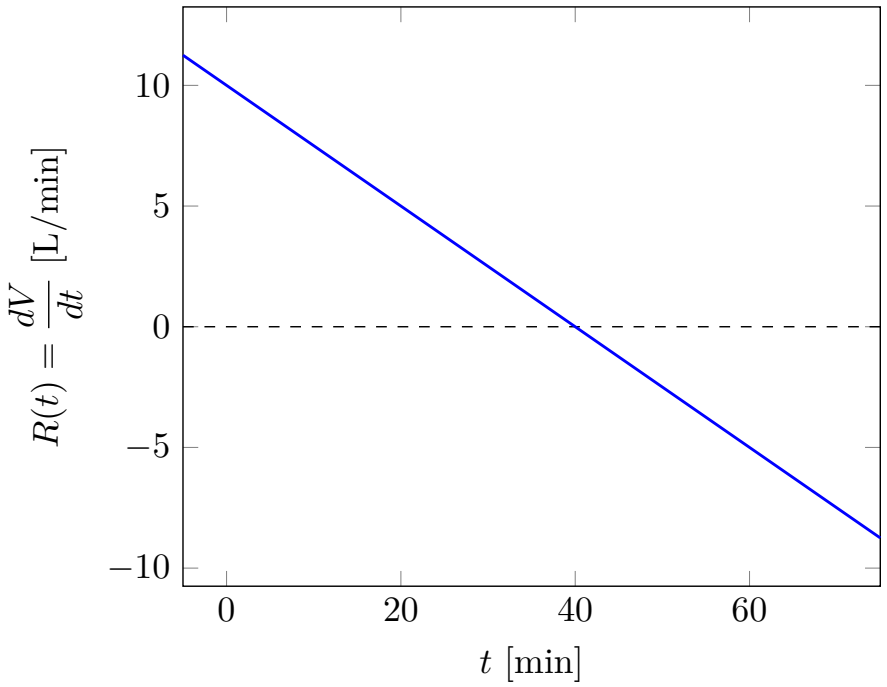


Figure 12.7.6 A graph of the rate of change of volume of water as a function of time.

The graph of the volume therefore needs to be concave down, increasing for $t < 40$ and decreasing for $t > 40$. We do not know the starting volume (it wasn't given), so we might measure the change of volume from the initial value. That is, $V = 0$ on the graph will mean the volume is the same as the initial volume, rather than meaning there is no water. Using our knowledge of (((Unresolved xref, reference "thm-elementary-definite-integrals"; check spelling or use "provisional" attribute)))definite integrals of elementary algebraic formulas and the (((Unresolved xref, reference "thm-definite-integral-linearity"; check spelling or use "provisional" attribute)))linearity of definite integrals, we can find an explicit formula for our volume:

$$\begin{aligned} V(t) &= \int_0^t R(z)\,dz = \int_0^t 10 - 0.25z\,dz \\ &= \int_0^t 10\,dz - 0.25\int_0^t z\,dz = 10t - 0.25\frac{t^2}{2}. \end{aligned}$$

The graph of this functions is shown below.

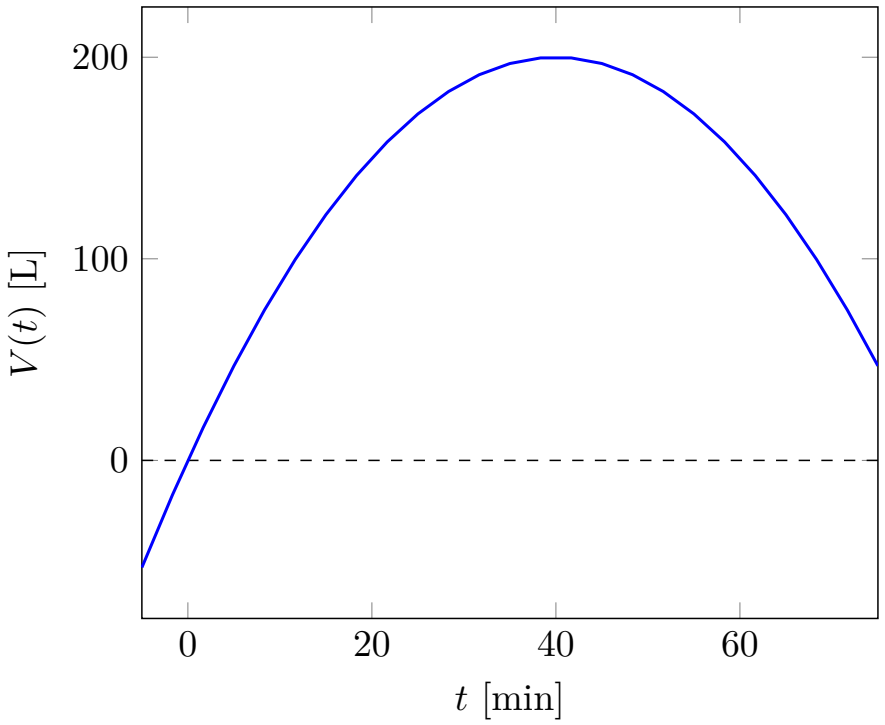


Figure 12.7.7 A graph of the change in volume of water as a function of time.

□

Definite integrals allow us to compute the change in a quantity when we know the rate of change. We are now turning our attention to the reverse question. If we know how to describe the quantity itself as a function of time, how do we find its corresponding rate of change? That rate of change is called the **derivative**.

12.7.2 Introduction to Differential Equations

The rate of change or derivative of a quantity Q with respect to an independent variable x is itself another variable in the state of the system. It is often the case for physical and biological systems that there is a relationship between Q and its derivative $\frac{dQ}{dx}$. When expressed as an equation, such a relationship is called a **differential equation**. A function for which the differential equation is satisfied is called a **solution**.

Example 12.7.8 A population P is a function of time t. The derivative $\frac{dP}{dt}$ is the rate of change of the population, which consists of the birth rate (total births per unit time) and death rate (total deaths per unit time). When each of these rates is proportional to the population size, we have Malthusian growth and can write the differential equation

$$\frac{dP}{dt} = b \cdot P - d \cdot P$$

where b is the per capita birth rate and d is the per capita death rate. If we write $r = b - d$ to define a per capita net growth rate r, then the differential equation simplifies to

$$\frac{dP}{dt} = rP.$$

□

Example 12.7.9 In physics, Newton's second law of motion is usually written as its equation $F = ma$, which says that the total force acting on a body is always equal to the mass of that body times the acceleration of the body. Because acceleration is the rate of change of velocity,

$$a = \frac{dv}{dt},$$

Newton's law is actually a differential equation if we can compute the force.

Consider a falling object with mass m. Gravity is a downward force acting on the body with gravitational force $F_g = -mg$ where g is the constant acceleration due to gravity. In the absence of other forces (no air resistance), the differential equation from Newton's law would be written

$$F_g = ma \quad \Leftrightarrow \quad -mg = m\frac{dv}{dt} \quad \Leftrightarrow \quad \frac{dv}{dt} = -g.$$

If there is air resistance, the resulting force is itself usually a function of the velocity of the object passing through the air, $F_a = f(v)$. Experimentally, it has been found that many objects follow a square law, that the air resistance is proportional to the square of the velocity. Because air resistance is always in opposition to motion, F_a must have the opposite sign as v. So we have

$$F_a = -\gamma v|v| = \begin{cases} -\gamma v^2, & v \geq 0 \\ +\gamma v^2, & v < 0. \end{cases}$$

Consequently, the differential equation of a falling object with air resistance is

$$F = ma \quad \Leftrightarrow \quad F_a + F_g = m\frac{dv}{dt} \quad \Leftrightarrow \quad -\gamma v|v| - mg = m\frac{dv}{dt}.$$

□

Knowing a differential equation can often allow us to understand much of the behavior of the system of interest simply by determining when the rate is predicted to be positive and negative. Reasoning through concavity can be a little more difficult. We will focus on the case where the differential equation only involves the quantity y and its derivative $\frac{dy}{dx}$ but does not involve the independent variable x. Such an equation is called an **autonomous differential equation**.

Theorem 12.7.10 *Suppose that we can write an autonomous differential equation in the form*

$$\frac{dy}{dx} = f(y),$$

so that the rate of change is explicitly a function of the dependent quantity itself. Then the behavior of y as a function of x depends on the sign of $f(y)$.

- *y is increasing if $f(y)$ is positive*
- *y is decreasing if $f(y)$ is negative*
- *y is constant if $f(y)$ is zero*

Note: There are some technical requirements on f that determine whether the differential equation has a good solution, but for typical algebraic functions everything is okay. A course on differential equations would include some discussion of these additional conditions.

Values where an autonomous rate function $f(y) = 0$ are called **equilibrium**

solutions because the value of y will never change if it has that value. That is, y will be a constant function that satisfies the differential equation.

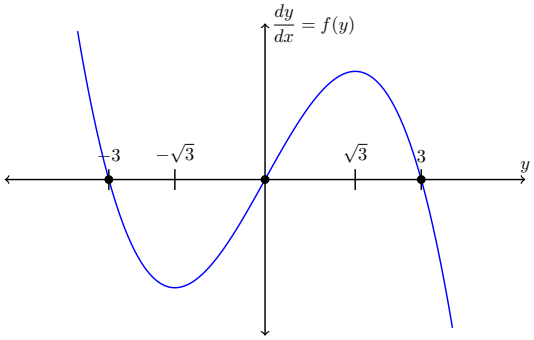
We often summarize the behavior of an autonomous differential equation using a **phase line**. A phase line is a number line representing the dependent quantity y. We mark the equilibrium solutions, where $f(y) = 0$, on the line. Between these points, we draw arrows representing whether y is increasing or decreasing between those points. If we also know where the rate $f(y)$ reaches extreme values, we mark the location of those extreme rate points on the number line as well to represent inflection points.

The behavior of the dependent quantity then depends on where its initial value starts. If the initial value is at an equilibrium, then the dependent variable has the same constant value for all values of the independent variable. Otherwise, the function will be either increasing or decreasing, moving toward or away from equilibrium solutions.

Example 12.7.11 Consider an autonomous differential equation

$$\frac{dy}{dx} = f(y)$$

where the rate function $f(y)$ is illustrated below. Create a phase line and use it to sketch expected shapes for solutions representing initial values in the different regions identified in the phase line.

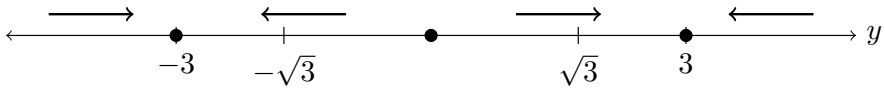


Solution. The graph of $\frac{dy}{dx} = f(y)$ has zeros at $y = 0$ and $y = \pm 3$. These correspond to equilibrium solutions of the differential equation. That is, $y(x) = 3$ (constant function) is one of the possible solutions. If the initial value is $y(0) = 3$, then $y(x) = 3$ for all other values of x as well. Similarly for $y(0) = 0$ or $y(0) = -3$. These will be the points shown on the phase line.

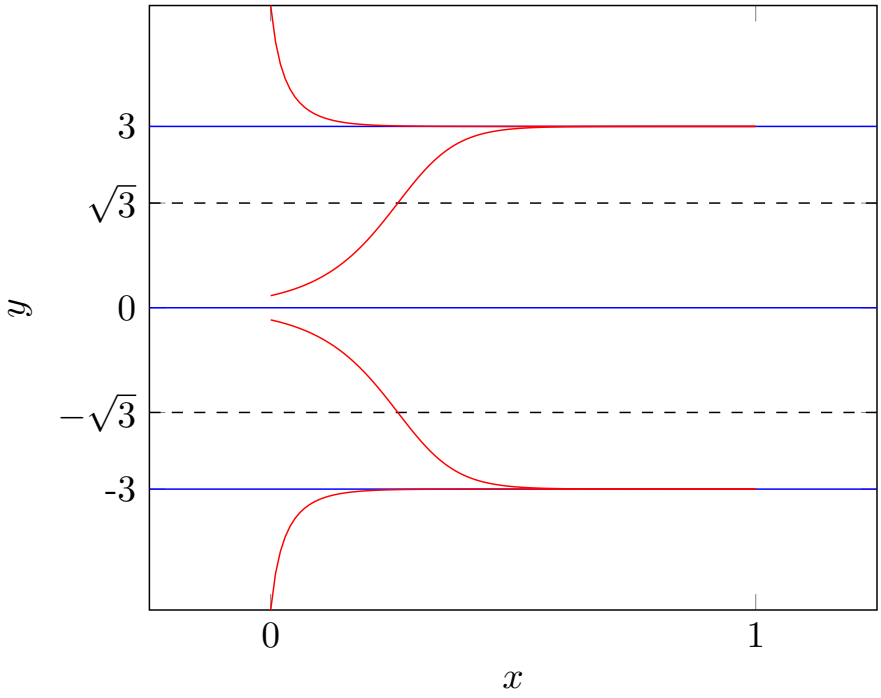
The three equilibrium points divide the phase line into four intervals: $(-\infty, -3)$, $(-3, 0)$, $(0, 3)$ and $(3, \infty)$. The graph of $\frac{dy}{dx} = f(y)$ allows us to look at the sign of $\frac{dy}{dx}$ for each of those regions. We see that $\frac{dy}{dx} > 0$ for $y \in (-\infty, -3)$, so $y(x)$ will be an increasing function if the initial value starts in this interval. We summarize the behavior in each interval with the following table.

Interval	$(-\infty, -3)$	$(-3, 0)$	$(0, 3)$	$(3, \infty)$
Sign of $\frac{dy}{dx} = f(y)$	+ (pos)	− (neg)	+ (pos)	− (neg)
Behavior of y	increasing	decreasing	increasing	decreasing

The phase line is a graphical summary of the table. Marked points on the line represent the equilibrium solutions. Arrows above the line represent the behavior as direction of motion. We will also include the locations of extreme rates, namely $y = \pm\sqrt{3}$, on the phase line as tick marks (not points) to indicate the locations of inflection points.



When we translate the information in the phase line to a sketch of the graph of solutions, we think of the phase line as the y-axis. The information about whether the function is increasing or decreasing is translated into the graph. Equilibrium solutions are also horizontal asymptotes for the solutions in the adjacent intervals. So we need to level off whenever the solution gets close to an equilibrium.



□

Chapter 13

Sequences as Models

13.1 Introduction to Discrete Calculus

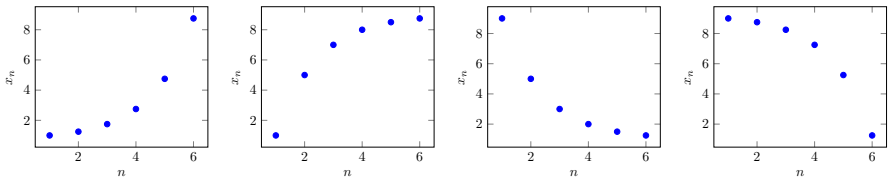
13.1.1 Overview

Calculus studies functions through their rates of change. Our goal is to understand relationships between concepts and not just rules of computation. The better we understand, the easier time we will have in drawing upon those concepts to answer questions. If we can use simpler concepts that can motivate the ideas of calculus, then those ideas will make more sense. Sequences provide one possible framework through which we can motivate the ideas of calculus.

In this section, we introduce the ideas that will be explored throughout the chapter. We learn to describe the behavior of sequences in terms of monotonicity and concavity in terms of increments of change. The ideas of monotonicity will allow us to answer questions relating to extreme values. We then outline how the rest of the chapter will proceed.

13.1.2 Behavior of Sequences

One of the results of calculus is the ability to describe the behavior of functions. We begin our discussion of the analysis of sequences with this in mind. We motivate the discussion on the behavior of sequences with some graphs.



(a) increasing, concave up **(b)** increasing, concave down **(c)** decreasing, concave up **(d)** decreasing, concave down

Figure 13.1.1 Graphs of four short subsequences that illustrate the concepts of monotonicity and concavity.

In mathematics, **monotonicity** refers to the direction of change in a sequence or function. In the figure above, the two sequences that are **increasing** show that the values of the sequence are rising. The sequences that are **decreasing** have values that fall. **Concavity** in the graphs corresponds to how the sequence appears to be bending. Imagine the sequences as points on a bowl shape. If the curve is bending up, the sequence is concave up. If the curve is bending down, the sequence is concave down. A sequence can change behavior multiple times, as shown in the figure below. We seek for methods of analyzing a sequence that will allow us to describe the monotonicity and concavity of a sequence.

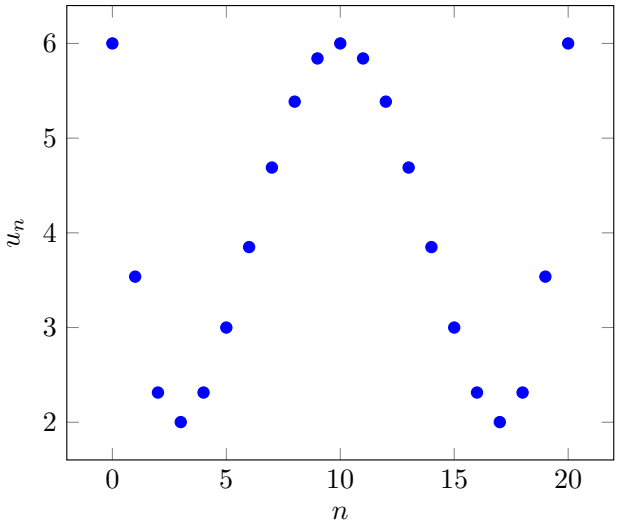


Figure 13.1.2 An example of a sequence that illustrates all four basic behaviors.

13.1.3 Monotonicity: Increasing and Decreasing

Our goal is to establish a way to describe these behaviors without relying solely on graphs. We begin by focusing on monotonicity.

Definition 13.1.3 Monotonicity of Sequences. For a sequence x with index k, we say that x is **increasing** on the interval $\{m, \ldots, n\}$ if for every two values in the interval, $i, j \in \{m, \ldots, n\}$ with $i < j$, we have $x_j > x_i$. That is, values later in the sequence are always greater than values earlier in the sequence.

We say that x is **decreasing** on the interval $\{m, \ldots, n\}$ if for every two values in the interval, $i, j \in \{m, \ldots, n\}$ with $i < j$, we have $x_j < x_i$. That is, values later in the sequence are always less than values earlier in the sequence. ◊

Our definition of monotonicity is on an interval of integers for the index. We talk about increasing and decreasing on intervals for several reasons. First, monotonicity is about comparisons of values. We should not think that the sequence is increasing or decreasing at a particular point. Instead, we are looking at how the sequence changes going from one index value to another. Second, when we later discuss the monotonicity of functions, we will describe monotonicity in terms of intervals from the domain. Consequently, we want to establish that pattern of thinking now.

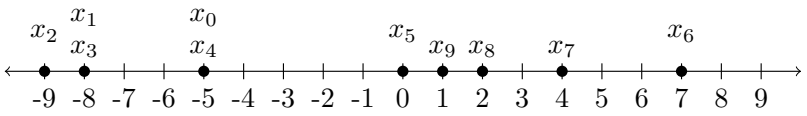
Inequalities are transitive. That is, if $a < b$ and $b < c$, then we know $a < c$. For sequences, this means that we don't really need to look at all possible pairs of index values in the interval. We only need to look at the consecutive terms in the sequence and see if they are increasing or decreasing.

Example 13.1.4 Consider the finite sequence

$$x = (x_k)_{k=0}^{9} = (-5, -8, -9, -8, -5, 0, 7, 4, 2, 1).$$

Describe the monotonicity of x.

Solution. We look at whether the sequence values are moving to the left or to the right on the number line. You could probably just visualize it in your mind, but a number line showing the values is illustrated below.



We can see that x is decreasing on the index interval $\{0, 1, 2\}$ because the values of the sequence move left on the number line:

$$x_1 < x_0, \quad x_2 < x_1.$$

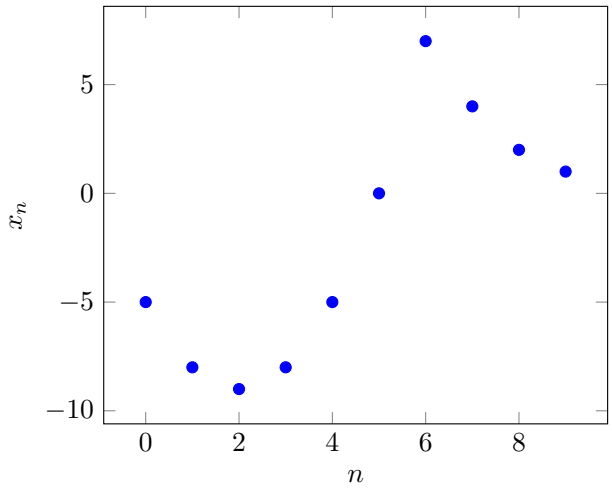
Next, we see that x is increasing on the interval $\{2, \ldots, 6\}$ as the sequence moves to the right:

$$x_3 > x_2, \quad x_4 > x_3, \quad \ldots \quad x_6 > x_5.$$

Finally, x is decreasing on the interval $\{6, \ldots, 9\}$:

$$x_7 < x_6, \quad x_8 < x_7, \quad x_9 > x_8.$$

We compare this analysis with a graph of the sequence, shown below. When the sequence is decreasing, the graph shows points of lowering heights. When the sequence is increasing, the graph shows points of rising heights.



□

When analyzing a sequence to determine monotonicity, we focus on inequalities involving consecutive terms. If we use $k = n - 1$ and $k = n$ to represent consecutive index values, then the inequalities are:

- $x_n > x_{n-1}$ for increasing
- $x_n < x_{n-1}$ for decreasing.

By moving all terms to one side, the inequalities become:

- $x_n - x_{n-1} > 0$ for increasing
- $x_n - x_{n-1} < 0$ for decreasing.

This means that monotonicity can be determined by looking at the signs of the differences between terms. We call those differences the **increments**.

Definition 13.1.5 Given a sequence $x = (x_k)_{k=m}^{n}$, the **increments** form a new sequence, $\nabla x = (\nabla x_k)_{k=m+1}^{n}$, calculated by the **backward difference**,

$$\nabla x_k = x_k - x_{k-1}.$$

◊

When computing the increments, we had to make a choice whether to do the backward difference or the forward difference,

$$\Delta x_k = x_{k+1} - x_k.$$

The values of the backward differences and the forward differences are exactly the same. They differ only in terms of index values. I have chosen to use backward differences in order for our later work with accumulation sequences and summation to work out more cleanly.

Example 13.1.6 For the sequence $x = (x_k)_{k=3}^{10} = (1, 2, 4, 7, 11, 16, 22, 29)$, find the increments defined by the backward difference.

Solution. We can find the values by writing the sequence as a row of values. Then below the gaps between the values, we can write the differences.

x_n	1	2	4	7	11	16	22	29
∇x_n		1	2	3	4	5	6	7

When the increments are defined by the backward difference, the index starts one value later than the original sequence. Thus, the increments are the sequence

$$\nabla x = (\nabla x_k)_{k=4}^{10} = (1, 2, 3, 4, 5, 6, 7).$$

Because all of the increments are positive, x is increasing on the full interval $\{3, \ldots, 10\}$. □

We can therefore analyze monotonicity of a sequence if we know the signs of its increments.

Theorem 13.1.7 Increment Test for Sequence Monotonicity. *Given a sequence x, we look at the sequence of increments defined by the backward difference. Suppose m and n are integers with $m \leq n$.*

- *If the sequence of increments is positive, $\nabla x_k > 0$ for every $k = m, \ldots, n$, then x is increasing on the interval $\{m-1, \ldots, n\}$.*
- *If the sequence of increments is negative, $\nabla x_k < 0$ for every $k = m, \ldots, n$, then x is decreasing on the interval $\{m-1, \ldots, n\}$.*
- *If the sequence of increments is zero, $\nabla x_k = 0$ for every $k = m, \ldots, n$, then x is constant on the interval $\{m-1, \ldots, n\}$.*

The interval of monotonicity always begins one value before the first index in the interval of increments because the first increment is $\nabla x_m = x_m - x_{m-1}$.

13.1.4 Concavity

We now turn our attention to the **concavity** of a sequence. Concavity is based on whether the sequence of increments is increasing or decreasing. When the increments are constant, we have an arithmetic sequence which follows a straight line. To be concave up, we need the graph to bend up. If an increment is positive, then the next increment needs to be a larger positive value. If an increment is negative, then the next increment needs to be a smaller magnitude negative value or a positive value. Either way, we need the increments to increase:

$$\nabla x_n > \nabla x_{n-1}.$$

Similarly, for a sequence to be concave down, the increments need to decrease:

$$\nabla x_n < \nabla x_{n-1}.$$

Definition 13.1.8 Concavity of a Sequence. Suppose x is a sequence with increments ∇x_k and m and n are index values with $m < n$.

- If the increment sequence ∇x is *increasing* on $\{m, \ldots, n\}$, then the sequence x is *concave up* on $\{m-1, \ldots, n\}$.
- If the increment sequence ∇x is *decreasing* on $\{m, \ldots, n\}$, then the sequence x is *concave down* on $\{m-1, \ldots, n\}$.
- If the increment sequence ∇x is *constant* on $\{m, \ldots, n\}$, then the sequence x has no concavity and is *linear* (i.e., straight) on $\{m-1, \ldots, n\}$.

$\Diamond$

Because the increments themselves form a sequence, we can look at the signs of the increments of the increments to analyze concavity. Computing the backward difference of a backward difference is called the **second backward difference**.

Definition 13.1.9 Suppose x is a sequence with increments ∇x_k. The **second backward difference** of x, written $\nabla^2 x_k$, measures the backward difference of the increments,

$$\nabla^2 x_k = \nabla(\nabla x_k) = \nabla x_k - \nabla x_{k-1}.$$

The first index of $\nabla^2 x_k$ is two greater than the first index of x. (A second forward difference $\delta^2 x$ is defined similarly but will not be used.) $\Diamond$

Having defined the second backward difference, we can state the test that allows us to analyze the concavity of a sequence.

Theorem 13.1.10 Second Difference Test for Sequence Concavity. *Given a sequence x and integers m and n with $m < n$.*

- *If the second backward difference is positive, $\nabla^2 x_k > 0$ for every $k = m, \ldots, n$, then x is concave up on the interval $\{m-2, \ldots, n\}$.*
- *If the second backward difference is negative, $\nabla^2 x_k < 0$ for every $k = m, \ldots, n$, then x is concave down on the interval $\{m-2, \ldots, n\}$.*
- *If the second backward difference is zero, $\nabla^2 x_k = 0$ for every $k = m, \ldots, n$, then x has no concavity and is linear on the interval $\{m-2, \ldots, n\}$.*

The interval of concavity always begins two values before the first index in the interval of second backward difference.

Example 13.1.11 The values for the subsequence shown in Figure 13.1.2 are shown in the table below, rounded to the nearest ten-thousandth. Determine the intervals of monotonicity and concavity.

n	u_n
0	6
1	3.5376
2	2.3136
3	2.0016
4	2.3136
5	3
6	3.8496
7	4.6896
8	5.3856
9	5.8416
10	6

n	u_n
11	5.8416
12	5.3856
13	4.6896
14	3.8496
15	3
16	2.3136
17	2.0016
18	2.3136
19	3.5376
20	6

Solution. We can augment the table with additional columns for the first and second backward differences. Then we can look at the signs of those increments to determine intervals for monotonicity and concavity. With as many terms as we are working with, we would definitely want to use a computer to assist here.

n	u_n	∇u_n	$\nabla^2 u_n$
0	6		
1	3.5376	-2.4624	
2	2.3136	-1.224	1.2384
3	2.0016	-0.312	0.912
4	2.3136	0.312	0.624
5	3	0.6864	0.3744
6	3.8496	0.8496	0.1632
7	4.6896	0.84	-0.0096
8	5.3856	0.696	-0.144
9	5.8416	0.456	-0.24
10	6	0.1584	-0.2976

n	u_n	∇u_n	$\nabla^2 u_n$
11	5.8416	-0.1584	-0.3168
12	5.3856	-0.456	-0.2976
13	4.6896	-0.696	-0.24
14	3.8496	-0.84	-0.144
15	3	-0.8496	-0.0096
16	2.3136	-0.6864	0.1632
17	2.0016	-0.312	0.3744
18	2.3136	0.312	0.624
19	3.5376	1.224	0.912
20	6	2.4624	1.2384

We begin by looking at the signs of the increments based on the first backward difference. Again, it is more compact to use a table to summarize our results.

Sign of ∇u_n	Interval	Behavior of u_n	Interval
negative	$\{1, 2, 3\}$	decreasing	$\{0, \ldots, 3\}$
positive	$\{4, \ldots, 10\}$	increasing	$\{3, \ldots, 10\}$
negative	$\{11, \ldots, 17\}$	decreasing	$\{10, \ldots, 17\}$
positive	$\{18, 19, 20\}$	increasing	$\{17, \ldots, 20\}$

We perform a similar analysis on the second backward differences to describe the concavity of the sequence.

Sign of $\nabla^2 u_n$	Interval	Behavior of u_n	Interval
positive	$\{2, \ldots, 6\}$	concave up	$\{0, \ldots, 6\}$
negative	$\{7, \ldots, 15\}$	concave down	$\{5, \ldots, 15\}$
positive	$\{16, \ldots, 20\}$	concave up	$\{14, \ldots, 20\}$

□

When doing this analysis, if you discover that an increment equals zero, remember that is neither positive nor negative. Do not include the index in the interval for the signs of increments. An increment of zero indicates that the sequence was not changing. If the second difference is zero, this indicates

the increments are not changing and the sequence is linear.

13.1.5 Extreme Values

Having discussed monotonicity, we turn our attention to the extreme values of a sequence. An extreme value refers to a maximum or minimum value in the sequence. Of course, when we have a list of the sequence values, we can scan through the list of values to find the highest or lowest value. We want some methods of analysis that will not require checking all of the values.

On an interval where a sequence is monotone, an extreme value will never occur within that interval. Extreme values can only occur at the edge of such an interval. To find a maximum value, we look for where a sequence transitions from increasing to decreasing. To find a minimum value, we look for where it transitions from decreasing to increasing. These turning points identify **local extreme values**.

Definition 13.1.12 Local Extreme Values for Sequences. A sequence x has a **local maximum** at index value k if there are index values $m < k < n$ so that

$$x_k \geq x_i,$$

for all $i \in \{m, \ldots, n\}$.

A sequence x has a **local minimum** at index value k if there are index values $m < k < n$ so that

$$x_k \leq x_i,$$

for all $i \in \{m, \ldots, n\}$. ◊

A sequence can have multiple local extremes. The example shown in Figure 13.1.2 has two local minima and one local maximum. The two minima happen to have the same value. However, even if one was higher than the other, they would both still be minima because they would be where the sequence transitioned from decreasing to increasing.

A **global extreme value** describes a value in the sequence that is either the largest value of all (the **global maximum**) or the lowest value of all (the **global minimum**). To describe the global extreme values, we would compare all of the local extremes as well as check if the sequence increased or decreased without bound. To describe that in more depth, we need to wait until we talk about limits.

13.1.6 Patterns in the Increments

Backward differences are not only useful for describing the monotonicity and concavity of a sequence. There are many sequences where patterns in the increments can be used to create non-recursive recurrence relations. A recurrence relation describes how to go from the previous value in a sequence to the next value. It is recursive if the relation is the same for every index. A non-recursive recurrence relation will have a relation that depends on the index. Sometimes, we can still identify a pattern by looking at the higher-order differences.

An arithmetic sequence is a sequence where the backward difference is a constant sequence. More complicated sequences arise where it is not the backward difference but the second difference or higher that is constant. If we can identify a higher-order backward difference that is constant, then we can use that pattern to predict additional terms.

Example 13.1.13 Consider the sequence $w = (0, -23, -40, -45, -32, 5, 72, 175, \ldots)$. Identify a pattern and find the next two values in the sequence.

Solution. This sequence is not arithmetic, nor is it geometric. To look for a pattern, we proceed to generate the backward differences.

w_n	0	-23	-40	-45	-32	5	72	175
∇w_n		-23	-17	-5	13	37	67	103
$\nabla^2 w_n$			6	12	18	24	30	36
$\nabla^3 w_n$				6	6	6	6	6

The pattern of backward differences shows that the third-order backward difference is a constant value, $\nabla^3 w_n = 6$. We can use this pattern to find the next few values of the sequence. We do this by extending the table, adding the increments one at a time. The last term in $\nabla^2 w_n$ shown was $\nabla^2 w_8 = 36$. So the next term will be $\nabla^2 w_9 = 36 + 6 = 42$. We can use that to find $\nabla w_9 = \nabla w_8 + 42 = 103 + 42 = 145$. That allows us to obtain $w_9 = w_8 + 145 = 175 + 145 = 320$. Repeating the process allows us to find $w_{10} = 513$, as illustrated in the extended table below.

w_n	0	-23	-40	-45	-32	5	72	175	320	513
∇w_n		-23	-17	-5	13	37	67	103	145	193
$\nabla^2 w_n$			6	12	18	24	30	36	42	48
$\nabla^3 w_n$				6	6	6	6	6	6	6

□

13.1.7 Where Do We Go From Here?

In this section, we have introduced the ideas of the characterizing a sequence in terms of monotonicity and concavity. We considered examples of sequences with given values to find the backward differences by hand. In the next section, we will consider how to find an explicit formula for the backward differences when we know an explicit formula for the sequence itself. This will allow us to analyze the monotonicity and concavity of a sequence without needing all of the values being tabulated. Knowing monotonicity will allow us to identify extreme values, again without computing all of the values. The ideas of backward differences are analogous to the concepts of derivatives in calculus.

At the end of this section, we considered an example where we used a pattern in the backward differences to generate additional values for the sequence. In a future section, we will consider the idea of starting with a known sequence at the level of increments and using that sequence to generate a corresponding sequence. We call the generated sequence an accumulation sequence, named this because the new value is an accumulation of increments. An important topic raised in that section is how to show that two sequences are the same. We often have multiple methods of describing a sequence, such as an explicit definition and a recursive definition. We want to know that the sequences really agree because we can't (and wouldn't want to) compute and compare infinitely many values. The ideas of accumulation are analogous to the concepts of integration in calculus.

One of the most important applications of accumulation is summation. Summation formulas will continue to generalize our last example in this section. Where accumulation allows us to take a known sequence of increments and generate the original sequence, summation formulas will allow us to take the formula of an increment sequence and compute an explicit formula for the original sequence. Summation formulas will be critical in generating the explicit formulas of integration later.

Finally, we will end the chapter by introducing limits of sequences. Limits are the key tool that generates all of calculus. A limit in a sequence corresponds

to having the values in a sequence converge to some value. We will then be ready to begin a new chapter that begins our look at calculus.

13.1.8 Summary

- Monotonicity of a sequence describes where the sequence is increasing or decreasing. We say that a sequence is increasing or decreasing *on intervals of integers*. The sequence is **increasing** if the sequence values move to the right on the number line. The sequence is **decreasing** if the sequence values move to the left on the number line.
- We calculate the **increments** of the sequence using the **backward difference** to analyze monotonicity,

$$\nabla x_k = x_k - x_{k-1}.$$

  - If the increments are positive, $\nabla x_k > 0$ for all k on an interval $\{m, \ldots, n\}$, then the values of the sequence x_k are *increasing* on the interval $\{m-1, \ldots, n\}$.
  - If the increments are negative, $\nabla x_k < 0$ for all k on an interval $\{m, \ldots, n\}$, then the values of the sequence x_k are *decreasing* on the interval $\{m-1, \ldots, n\}$.
- Concavity of a sequence describes where the increments are increasing or decreasing. A sequence whose increments are increasing is **concave up**. A sequence whose increments are decreasing is **concave down**.
- We can analyze concavity by computing the increments of the increments using the **second backward difference**,

$$\nabla^2 x_k = \nabla x_k - \nabla x_{k-1}.$$

  - If the second increments are positive, $\nabla^2 x_k > 0$ for all k on an interval $\{m, \ldots, n\}$, then the values of the sequence x_k are *concave up* on the interval $\{m-2, \ldots, n\}$.
  - If the second increments are negative, $\nabla^2 x_k < 0$ for all k on an interval $\{m, \ldots, n\}$, then the values of the sequence x_k are *concave down* on the interval $\{m-2, \ldots, n\}$.
- We can also use the increments and higher-order increments to find patterns in some sequences. Following these patterns can be used to predict later terms in the sequence with index-dependent recurrence relations.

13.1.9 Exercises

For the following finite sequences, find the intervals of monotonicity and concavity. Also, identify the index and values of the local maximum and minimum points. Graph the sequences and compare with your results.

1. $a = (a_n)_{n=1}^{10} = (35, 55, 70, 80, 85, 85, 80, 70, 55, 35)$

2. $b = (b_i)_{i=0}^{9} = (-8, -5, -4, -5, -8, -9, -8, -5, -1, 4)$

3. $c = (c_t)_{t=-3}^{6} = (5, 8, 10, 10, 10, 9, 8, 7, 8, 10)$

For the following sequences, identify a pattern using backward differences. Use the pattern to predict the next two values of the sequence.

4. $u = (u_k)_{n=0}^{\infty} = (-8, -13, -16, -17, -16, -13, -8, \ldots)$

5. $v = (v_j)_{j=1}^{\infty} = (6, 0, -6, -10, -10, -4, 10, \ldots)$

6. $w = (w_n)_{n=0}^{\infty} = (-3, -6, -5, 1, 12, 27, 44, 60, 71, \ldots)$

13.2 Recursive Sequences and Projection Functions

Overview. When looking for patterns in sequences, we usually explore two possibilities. One approach is to look at the values of individual terms and see if there is an explicit formula relating the index with the formula. There were several examples of this in the previous section. Another approach is to look for a pattern in how terms are generated from earlier terms. For example, the sequence $(7, 10, 13, 16, \ldots)$ is easy to recognize that each term is found by adding 3 to the previous term.

In this section, we consider recursively defined sequences. Arithmetic and geometric sequences are two familiar examples of sequences with recursive definitions. We review some basic ideas about functions. We will learn about projection functions used in such recursive definitions. We visualize the role of projection functions as maps between sequence values and through cobweb diagrams.

13.2.1 Arithmetic and Geometric Sequences

We often think of sequences in terms of a pattern for how to find the values. When a sequence can be defined so that the next value can found knowing only the previous value, we say the sequence has a **recursive definition**. The simplest pattern-based sequences follow simple recursive patterns.

An **arithmetic sequence** is a sequence whose terms change by a fixed increment or difference. For example, consider the sequence introduced above,

$$x = (7, 10, 13, 16, \ldots).$$

The pattern for this sequence was that we add 3 to each term in the sequence to find the next term. The value 3 is called the **increment** or **difference** of the sequence. It is called the increment because there is a pattern of adding the same value to values in the sequence:

$$\begin{aligned} x_2 &= x_1 + 3 = 7 + 3, \\ x_3 &= x_2 + 3 = 10 + 3, \\ x_4 &= x_3 + 3 = 13 + 3, \\ &\vdots \end{aligned}$$

It is called the difference because each of those equations can be rewritten as a difference of values:

$$\begin{aligned} x_2 - x_1 &= 3, \\ x_3 - x_2 &= 3, \\ x_4 - x_3 &= 3, \\ &\vdots \end{aligned}$$

A **geometric sequence** is a sequence whose terms change by a fixed multiple or ratio. An example of a geometric sequence is given by

$$u = (2, 6, 18, 54, \ldots).$$

Each term is found by multiplying the previous term by 3, which is the **multiple** or **ratio** of the sequence. We call the value 3 the multiple because of the pattern

$$\begin{aligned} u_2 &= 3u_1 = 3 \cdot 2, \\ u_3 &= 3u_2 = 3 \cdot 6, \\ u_4 &= 3u_3 = 3 \cdot 18, \\ &\vdots \end{aligned}$$

It is call the ratio if we rewrite the equations as a ratio of a value to its previous value

$$\begin{aligned} \frac{u_2}{u_1} &= 3, \\ \frac{u_3}{u_2} &= 3, \\ \frac{u_4}{u_3} &= 3, \\ &\dots \end{aligned}$$

For recursively defined sequences, the equation that describes the relationship between consecutive terms of the sequence is called the **recurrence relation**. When the recurrence relation for a sequence x is solved for the next value as a dependent variable in terms of an expression involving of the previous term, we call this map or function the **projection function** because it allows us to *project* future values based on current values.

13.2.2 Functions as Maps

Before we discuss more about projection functions, we take a short diversion to review some core concepts about functions in general. Functions are at the heart of everything we do in calculus. Unfortunately, many students have subtle misconceptions about how to think about functions. We want to use our emphasis on sequences to come to terms with these ideas. Be prepared to think about functions from many different points of view. We begin by thinking of a function as a map between variables.

Definition 13.2.1 Function. Given two related variables, say x and y, such that there is a rule or relation that defines a map $x \mapsto y$, we say that the map is a **function**. (We will make this more precise later.[1]) ◊

Similar to other mathematical objects like variables and sequences, functions are usually represented by symbols for their names. Letter symbols like f or g are particularly common. We would write $f : x \mapsto y$ to say that the name of the function representing the map from a variable x to y is f. The independent variable x is the **input** for the function; the dependent variable y is the **output** of the function.

We often use **function notation**, writing $y = f(x)$. The parentheses in function notation indicate that whatever is inside is the value for the input, *not* multiplication. So this equation says that the dependent variable y is equal to the output of the function f when the input has the value represented by x. We usually read the equation, "y equals the value of f of x." When we have an equation expressing y as an explicit expression involving x, that expression can be used to define the function.

[1] The precise definition needs to address the domain of the function and clarify what is a map when there is no defining expression.

Example 13.2.2 Consider the equation $2x + 5y = 10$, which relates the variables x and y. Because we can solve for the variable y to be a dependent variable,

$$y = -\frac{2}{5}x + 2,$$

the equation defines a function $x \mapsto y$. We can choose any name for this function (other than the symbols x or y, obviously). We might choose to use the name P and write this as a map

$$P : x \mapsto y = -\frac{2}{5}x + 2.$$

Using the usual function notation, we would instead write

$$y = P(x) = -\frac{2}{5}x + 2.$$

□

When we wrote explicit formulas for the value of a sequence with the index as the independent variable, we noted that we had a map from the index to the value of the sequence. That was an example of a function.

Example 13.2.3 For the explicitly defined sequence $x_n = 3n - 2$, $n = 1, 2, 3, \ldots$, the equation defines a map $n \mapsto x_n = 3n - 2$. We could name the function S, for example, and write $S(n) = 3n - 2$ so that $x_n = S(n)$. □

13.2.3 Projection Functions

We now return to the concept of a recursively defined sequence and the projection function. A sequence is recursive if the *same* rule is used to go from one value of the sequence to the next. For our earlier example of an arithmetic sequence, we had a pattern of equations

$$\begin{aligned} x_2 &= x_1 + 3, \\ x_3 &= x_2 + 3, \\ x_4 &= x_3 + 3, \\ &\vdots \end{aligned}$$

The rule was always the same, but the symbols were different because they involved different index values.

We need some notation to capture the idea of consecutive values of a sequence. If n represents the value of the index for a sequence, then $n + 1$ represents the value of the index for the subsequent term of the sequence and $n - 1$ represents the value of the index for the preceding term of the sequence. For example, if $n = 3$, then x_n would be x_3, x_{n+1} would be x_4, and x_{n-1} would be x_2. All of our equations follow the pattern

$$x_n = x_{n-1} + 3,$$

where the first equation corresponds to $n = 2$, the second to $n = 3$, and so forth. Equivalently, we could think of all of the equations as following the pattern

$$x_{n+1} = x_n + 3,$$

but now the first equation comes from $n = 1$ and the second from $n = 2$. This equation which relates x_n and x_{n-1} is called a **recurrence relation** or

recursive equation for the sequence. Because the equation has been written with x_n as a dependent variable in terms of the value of x_{n-1}, we actually have a projection function.

Definition 13.2.4 Projection Function. Suppose a sequence x is defined by a recurrence relation of the form

$$x_n = f(x_{n-1}).$$

The function f defining the relation $x_{n-1} \mapsto x_n$ is called the **projection function**. Equivalently, we could write the recurrence relation in terms of a previous value as

$$x_{n+1} = f(x_n).$$

$\Diamond$

For a sequence with a recurrence relation $x_n = x_{n-1} + 3$, the projection function is $f : x_{n-1} \mapsto x_n = x_{n-1} + 3$. The function f takes a value as an input and maps it to an output that is that value plus 3. I find it useful to imagine the independent variable in a function as if it were a box, as in

$$f(x_{n-1}) = x_{n-1} + 3 \quad \Leftrightarrow \quad f(\square) = \square + 3.$$

Whatever is between the parentheses for the input of a function will go in the box of the formula. With this understanding, any variable can be used as a placeholder for the input: $f(x) = x + 3$ and $f(a) = a + 3$ describe the same mapping rule.

When we have a projection function for a sequence defined recursively, we can apply the function repeatedly to calculate values for the sequence. Many sequences can use the same projection function. We need an **initial value** to begin the process.

Example 13.2.5 A sequence $u = (u_n)_{n=0}^{\infty}$ is defined recursively by the projection function $f(x) = 2x - 5$ and an initial value $u_0 = 3$. Find the next four terms of the sequence.

Solution. The projection function defines the map $u_{n-1} \mapsto u_n$ according to the rule $f(x) = 2x - 5$ or $f(\square) = 2 \cdot \square - 5$. It tells us that if we use an input coming from a sequence value, the output of the function will be the next sequence value. We were given an initial value $u_0 = 3$. Using the recursive equation for $n = 1$ and the preceding value $u_0 = 3$ as the input to f, the output will be the value u_1:

$$u_1 = f(u_0) = f(3) = 2(3) - 5 = 1.$$

Now that we know u_1, we can use that value as an input to get u_2, and so on:

$$\begin{aligned} u_2 &= f(u_1) = 2(1) - 5 = -3, \\ u_3 &= f(u_2) = 2(-3) - 5 = -11, \\ u_4 &= f(u_3) = 2(-11) - 5 = -27. \end{aligned}$$

$\square$

Sometimes, a recurrence relation is not written with the new sequence value isolated. To identify the projection function, we need to solve for the new value of the sequence.

Example 13.2.6 A sequence is defined by the recurrence relation

$$w_{n+1} - w_n = 1.4w_n - \frac{3}{w_n}, \quad n \geq -1,$$

and an initial value $w_{-1} = 1$. Find the recursive equation corresponding to the projection function, $w_n \mapsto w_{n+1}$, and find $f(x)$. Use this to find w_0 and w_1.

Solution. To find the recursive equation, we need to solve for w_{n+1}.

$$w_{n+1} - w_n = 1.4w_n - \frac{3}{w_n}$$
$$w_{n+1} = 2.4w_n - \frac{3}{w_n}$$

This recursive equation gives us a map from a current value w_n to the next value w_{n+1} in the sequence, $w_n \mapsto w_{n+1}$, which is the projection function

$$w_{n+1} = f(w_n) = 2.4w_n - \frac{3}{w_n}.$$

This means that using an input x gives

$$f(x) = 2.4x - \frac{3}{x}.$$

Once we have the map, we can repeatedly use the projection function to find subsequent values of the sequence.

$$\begin{aligned} w_{-1} &= 1 \\ w_0 &= f(w_{-1}) = f(1) \\ &= 2.4(1) - \frac{3}{1} = -0.6 \\ w_1 &= f(w_0) = f(-0.6) \\ &= 2.4(-0.6) - \frac{3}{-0.6} = 3.56 \end{aligned}$$

□

13.2.4 Arithmetic and Geometric Sequences Revisited

The arithmetic and geometric sequences have simple explicit formulas. We use these formulas to illustrate the idea of a sequence as a map from the index to the sequence value.

The explicit formula for an arithmetic sequence is a special case of a linear function. The increment of the sequence represents the slope. The initial value gives us a known point. Knowing how many steps away from the given point along with the increment allows us to compute other sequence values.

Example 13.2.7 Find the explicit formula for the arithmetic sequence $x = (7, 10, 13, 16, \ldots)$. Use the formula to find x_{100}.

Solution. The initial value $x_1 = 7$ means our function will be a map $n \mapsto x_n$ that takes an input $n = 1$ to an output $x_1 = 7$. The increment of 3 that appears in the recursive equation $x_n = x_{n-1} + 3$ means that the function increases by 3 for every increment of the index by 1. That is, the slope is $m = +3$. The value of x_n will equal 7 plus 3 times the number of increments in the index,

$$x_n = 7 + 3(n - 1).$$

We could find an equivalent expression after using the distributive property,

$$x_n = 4 + 3n.$$

Because we now have the function $n \mapsto x_n$, we can find the value of the

sequence for any index using this expression. To find x_{100}, we use $n = 100$ and the map $n \mapsto x_n$,

$$x_{100} = 4 + 3 \cdot 100 = 304.$$

□

The following theorem provides the formula for the explicit formula of any arithmetic sequence.

Theorem 13.2.8 Explicit Formula of Arithmetic Sequences. *An arithmetic sequence x with an increment β so that $x_n = x_{n-1} + \beta$ and with a given initial value x_k has an explicit representation $n \mapsto x_n$ given by*

$$x_n = x_k + \beta \cdot (n - k).$$

The explicit formula for a geometric sequence is a special case of an exponential function. The multiple for the sequence corresponds to the base of the exponential. An initial value gives us a known point. The formula will count how many increments the index has changed and multiply by the base to that power.

Example 13.2.9 Find the explicit formula for the geometric sequence $u = (48, 24, 12, 6, 3, \ldots)$. Use the formula to find u_{20}.

Solution. The sequence u is geometric because the ratio of the sequence value to its predecessor is always the same.

$$\begin{aligned}
\frac{u_2}{u_1} &= \frac{24}{48} = \frac{1}{2} \\
\frac{u_3}{u_2} &= \frac{12}{24} = \frac{1}{2} \\
\frac{u_4}{u_3} &= \frac{6}{12} = \frac{1}{2} \\
\frac{u_5}{u_4} &= \frac{3}{6} = \frac{1}{2}
\end{aligned}$$

The recursive formula for the sequence multiplies the previous sequence value by $\rho = \frac{1}{2}$,

$$u_n = \frac{1}{2} \cdot u_{n-1}.$$

The initial value $u_1 = 48$ means our function will be a map $n \mapsto x_n$ that takes an input $n = 1$ to an output $x_1 = 48$.

Each time the index is incremented by 1, the value of the sequence is multiplied by $\rho = \frac{1}{2}$. We can count the number of increments for the index n by the expression $n - 1$. Because repeated multiplication is a power, we obtain an explicit formula

$$u_n = u_1 \cdot \rho^{n-1} = 48 \cdot \left(\frac{1}{2}\right)^{n-1}.$$

Using the properties of powers, this is equivalent to

$$u_n = \frac{48}{2^{n-1}}.$$

With the function $n \mapsto u_n$, we can find the value of the sequence for any index using this expression. To find u_{20}, we use $n = 20$ and the map $n \mapsto u_n$,

$$u_{20} = \frac{48}{2^{19}}.$$

If we rewrite $48 = 16 \cdot 3 = 2^4 \cdot 3$ and then simplify the fraction, this is equivalent to

$$u_{20} = \frac{3}{2^{15}} = \frac{3}{32768}.$$

□

The general formula for a geometric sequence is provided in the following theorem.

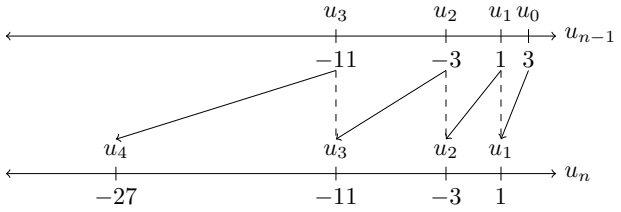
Theorem 13.2.10 Explicit Formula of Geometric Sequences. *A geometric sequence x with a multiple ρ so that $x_n = \rho \cdot x_{n-1}$ and with a given initial value x_k has an explicit representation $n \mapsto x_n$ given by*

$$x_n = x_k \cdot \rho^{(n-k)}.$$

13.2.5 Graphical Representations of Projections

If we think about a function as a map between two number lines, then the process of using a projection function to find values in a sequence can be visualized using such a mapping. Consider two number lines. The top number line will represent the current value of the sequence or the input of the function. The bottom number line will represent the next value of the sequence or the output of the function. The projection function defines the rule for how we go from the input to the output. Because the process repeats, we also go from the output number line to the same value on the input number line.

Example 13.2.11 For the sequence defined by the projection function $f(x) = 2x - 5$ and initial value $u_0 = 3$, the mapping used to generate the first few terms of the sequence can be represented graphically as shown below.

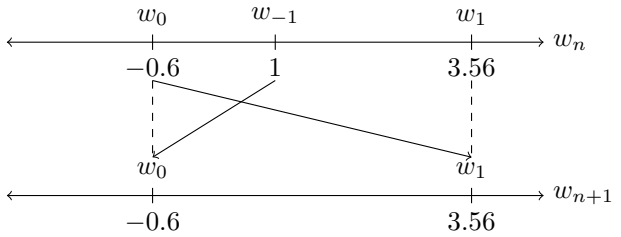


□

Example 13.2.12 For the sequence defined by the projection function

$$w_{n+1} = 2.4w_n - \frac{3}{w_n}$$

and initial value $w_{-1} = 1$, the mapping used to generate the first few terms of the sequence can be represented graphically as shown below.



□

The graph of a function f shows all points (x, y) in the plane where $y = f(x)$. Where a mapping visualizes a function as going from the input number line to the output number line, a graph visualizes a function by thinking of the x-axis as the input number line and the vertical position of the graph in the

y-direction as the output. It is as if there were a separate vertical number line parallel to the y-axis (and perpendicular to the x-axis) through every value on the x-axis. The point on the graph corresponds precisely to the location of the output for the function.

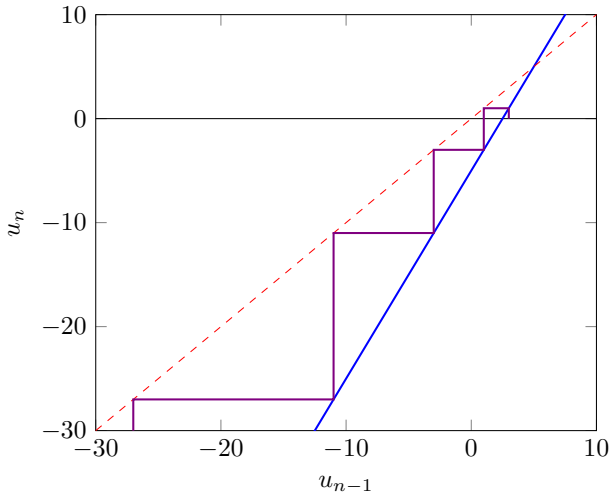
We can use the graph of a projection function to visualize how to generate a recursive sequence. The algorithm we use follows the same pattern as we used in the mapping and generates what is called a **cobweb diagram** in the plane.

Algorithm 13.2.13 Generating a Cobweb Diagram. *A cobweb diagram for a sequence x with projection function f and initial value x_0 is generated by the following steps.*

1. *Graph $y = f(x)$ and $y = x$ in the same plane.*
2. *Label the x-axis x_{n-1} (for the previous value) and the y-axis x_n (for the next value).*
3. *Find the initial value x_0 on the x-axis. The initial point will be $(x_0, 0)$ representing the* current sequence value.
4. *Draw a* vertical line *from the point representing the current sequence value to the graph $y = f(x)$. This corresponds to using the map to find the* next sequence value.
5. *Draw a* horizontal line *from the point for the next sequence value to the graph $y = x$. This corresponds to resetting the current sequence value based on the most recent next sequence value.*
6. *Repeat the last two steps as many times as desired.*

Example 13.2.14 Draw the first four iterations of the cobweb diagram for the sequence u with projection function $f(x) = 2x - 5$ and initial value $u_0 = 3$.

Solution. We start by drawing the graphs $y = f(x) = 2x-5$ and $y = x$ on the same graph. We then start with a point on the x-axis at $x = 3$ corresponding to the value of $u_0 = 3$. We want to use this value to find the next value in the sequence u_1. This use the projection function, so we go up to the value of the function $f(3) = 1$, drawing a vertical line segment to the point $(3, 1)$. We now know $u_1 = 1$ and we need to use this as a new input for the projection function. So the next step is to draw a horizontal segment to $y = x$ at the point $(1, 1)$. Now that our x-value is 1, we can repeat the process and use the function to find $u_2 = f(1) = -3$, drawing a vertical segment down to the point $(1, -3)$ and then a horizontal segment to $(-3, -3)$.



□

13.2.6 Summary

- A sequence x is recursive when the relation between consecutive values of the sequence is the same for every index. An equation describing this relation is called a **recurrence relation**. If we can solve for x_n as a dependent variable with x_{n-1} as the independent variable, the corresponding equation is called the **recursive equation**.

- An arithmetic sequence with common difference c has a projection function $f(x) = x + c$, a recursive relation

$$x_n = x_{n-1} + c,$$

  and an explicit formula given a known value for x_k,

$$x_n = x_k + \beta(n - k).$$

- A geometric sequence with common ratio ρ has a projection function $f(x) = \rho x$, a recursive relation

$$x_n = \rho \cdot x_{n-1},$$

  and an explicit formula given a known value for x_k,

$$x_n = x_k \cdot \rho^{n-k}.$$

- We think of functions as maps from the value of one variable to the value of another variable. For a sequence x, the map from the index n to the sequence value x_n is called the explicit function of the sequence. For a recursive sequence, the map from one sequence value x_{n-1} to the next sequence value x_n is called the projection function.

- The graph of a function uses values on the x-axis as input values and the vertical position of the graph as output values. A cobweb diagram uses the graph of a projection function with repeatedly updated inputs to generate a visual representation of a recursive sequence.

13.2.7 Exercises

Determine if each sequence is arithmetic, geometric, or neither. For each sequence that is arithmetic or geometric, (i) state the recursive equation for the sequence, (ii) find the projection function $f(x)$, (iii) state an explicit formula for the sequence, and (iv) use the explicit formula to find the value with index 20.

1. $u = (u_n)_{n=0}^{\infty} = (-8, -2, 4, 10, \ldots)$

2. $t = (t_k)_{k=2}^{\infty} = (27, 23, 19, 15, \ldots)$

3. $v = (v_k)_{k=-2}^{\infty} = (12, 16, 21, 27, \ldots)$

4. $w = (w_k)_{k=1}^{\infty} = (4, 20, 100, 500, \ldots)$

5. $z = (z_i)_{i=0}^{\infty} = (27, 18, 12, 8, \ldots)$

Each problem gives a projection function and an initial value that together determine a sequence recursively. Find the four terms following the initial value. Illustrate the sequence as a map between two number lines.

6. $P = (P_t)_{t=0}^{\infty}$ with $P_0 = 400$ and projection function $f(x) = x + 25$.

7. $u = (u_n)_{n=0}^{\infty}$ with $u_0 = 3$ and projection function $f(x) = 1.5x + 1$.

8. $u = (u_n)_{n=0}^{\infty}$ with $u_0 = -3$ and projection function $f(x) = 1.5x + 1$.

9. $w = (w_i)_{i=1}^{\infty}$ with $w_1 = 4$ and projection function $f(x) = 2.5x - 6$.

10. $w = (w_i)_{i=1}^{\infty}$ with $w_1 = 5$ and projection function $f(x) = 2.5x - 6$.

11. $z = (z_j)_{j=0}^{\infty}$ with $z_0 = 16$ and projection function $f(x) = \sqrt{x}$.

Each problem defines a sequence recursively. Give the formula of the projection function $f(x)$. Create the cobweb diagram for the sequence corresponding to the first five values of the sequence.

12. $Q = (Q_t)_{t=0}^{\infty}$ with $Q_0 = 3$ and $Q_t = Q_{t-1} + 4$.

13. $c = (c_k)_{k=0}^{\infty}$ with $c_0 = 10$ and $c_{k+1} = 0.75c_k$.

14. $S = (S_n)_{n=0}^{\infty}$ with $S_0 = 10$ and $S_n = 0.8S_{n-1} + 4$.

15. $P = (P_t)_{t=0}^{\infty}$ with $P_0 = 1$ and $P_{t+1} = \dfrac{20P_t}{P_t + 10}$.

13.3 Computing Sequence Values

13.3.1 Overview

When working with sequences, we often need to generate many sequence values. It is quite cumbersome to do this by hand. There would be a lot of repetition. Computers should be used to compute.

This section focuses on developing some basic skills in using computers to generate and plot sequences. Spreadsheets are one tool that can be used to compute and plot sequences. A spreadsheet is essentially a blank table; you create computational rules for individual entries within the table. For sequences, those rules correspond to the formulas for generating the sequence.

An alternative to spreadsheets is writing a computational script in a programming language. This might seem intimidating, especially if you have never done any programming. However, many modern scripting languages use commands very similar to the mathematical statements we already use. Scripts have the advantage of reducing the amount of work required to generate data quickly.

13.3.2 Sequences in Spreadsheets

One way to generate a sequence is with a spreadsheet. Common spreadsheet applications include Microsoft Excel, Apple Numbers, and Google Drive Sheets. A spreadsheet essentially starts as a giant blank table with rows and columns. The rows are numbered starting at 1 and the columns are labeled by letters. (After the first 26 columns, columns are labeled by pairs of letters.) Every cell in the table is identified by its column and row, called its **address**. So cell `B4` would be the cell in the fourth row of the second column.

Spreadsheets are designed to perform calculations based on the values of other cells. Suppose that `A1` contained the number 3 and `A2` contained the number 5. If you were to type in cell `A3` the formula `=A1+A2` (including the equal sign), then `A3` would show the value 8. However, it internally remembers the formula. If you were to change the values in either `A1` or `A2`, the value in `A3` would automatically be updated. We can take advantage of these calculations to compute the values of sequences.

For our first example, we look at how to use a spreadsheet to generate a sequence defined explicitly.

Example 13.3.1 Use a spreadsheet to generate values for the sequence defined by

$$x_n = \frac{3n}{4n+5}, \quad n = 0, 1, 2, \ldots.$$

Solution. Our sequence is defined explicitly by the map $n \mapsto x_n$. In our spreadsheet, we will use one column to define a subsequence of the values for the index. Then we will define a second column for the values of the sequence.

We will make the first column, `A`, contain the values for the index. The first entry in the column will be a label. So type `n` in the cell `A1`. Our first value for the index i $n = 0$, so type `0` in cell `A2`. The next value will be $n = 1$, so we type `1` in cell `A3`.

Now, it would be very tedious to type all of the values for the index once entry at a time. We take advantage of technology to have the spreadsheet apply a pattern to complete the rest of the column. Select the two cells we have already created, `A2` and `A3`. You should see a grab box, usually in the bottom right corner. If you drag that box down the column and then let go,

the spreadsheet will follow your pattern for all of the cells you select before releasing. You can make this column of index values as long as you desire, within the limits of the program you use.

We are now ready to create the column containing the values of the sequence. We start with a label for the column in `B1`, for example typing `x_n`. The rest of the column needs to use the map

$$n \mapsto x_n = \frac{3n}{4n+5}.$$

The value of input for this map, the index, is in column `A`. We want the value of the output for the map placed in column `B`. In cell `B2`, we want the output based on an input from `A2`, so we type `=(3*A2)/(4*A2+5)`.

We want to have this process repeated for the rest of the column, but we do not want to type a formula for each cell. We want the software to fill the column automatically. Notice that if you copy the formula from `B2` and paste it into `B3`, the formula is automatically adjusted to refer to `A3` instead of `A2`. This type of automatic modification of a formula is called **relative addressing**, where a pasted formula uses the relative position of a calculation. In this case, the relative position is to use the cell immediately to the left of the output.

Pasting the formula into every cell in the column is faster than typing a formula in every cell, but it is still too much work. We let the spreadsheet do all of the work by *filling* the remaining cells at once. We can do this by repeating the process earlier. Select the cell with a valid formula in `B2`, and you will again see the grab box in the corner. Click on the box and drag down the column. When you release the selection, the formula will be filled into all of the selected cells, adjusted to use relative addresses.

You now have two columns: the index in column `A` and the values in column `B`. You can create a graph by selecting the two columns of data and inserting a scatter plot. □

The second example illustrates how to use spreadsheet to compute the values of a recursively defined sequence.

Example 13.3.2 Use a spreadsheet to generate values for the sequence defined by

$$x_n = 2.3x_{n-1}(1 - 0.1x_{n-1})$$

with an initial value $x_0 = 1$.

Solution. We again want a column for the index values, which for convenience we will put in column `A`. The same steps as in the previous example apply. Put a label `n` in `A1`. Start with values `0` in `A2` and `1` in `A3`. Use the filling tool of the spreadsheet to extend the pattern for as far down the column as you desire.

In the next column `B`, we will put the values of the sequence. Put the label `x_n` in `B1`. Enter the initial value `1` in `B2` since $x_0 = 1$. For the rest of the values in the column, we will use the recursive map,

$$x_{n-1} \mapsto x_n = 2.3x_{n-1}(1 - 0.1x_{n-1}).$$

In the table, the previous value x_{n-1} corresponds to that value in the cell directly above the cell in question. Consequently, to compute x_1, we select `B3` and type `=2.3*B2*(1-0.1*B2)`. The value of x_1 should appear in the cell.

The rest of the column can be calculated by selecting `B3`, where we just typed the formula, and using the applications fill down feature. Notice that if you change the value for the initial condition in `B2`, every subsequent entry in the column is automatically updated. If you create a graph using columns `A` and `B`, the graph also is updated whenever the initial value is updated. □

The final example illustrates using parameters in our calculations.

Example 13.3.3 Use a spreadsheet to generate values for an arbitrary arithmetic sequence defined by

$$x_n = x_{n-1} + \beta$$

with an initial value $x_0 = a$, where a and β are parameters. Include the explicit and recursive calculations side-by-side in the table.

Solution. When we have parameters, we need to use part of our table to enter those values. You could put them anywhere in the table that is convenient. We will put them to the left of the generated table. There are two parameters: β and a. In cell `A1`, type the label `beta`. Then enter a value in the neighboring cell `B1` such as `3` for $\beta = 3$. In cell `A2`, type the label `a`. Then enter a value in the neighboring cell `B2` such as `8` for $a = 8$. The labels are primarily for our convenience to remember the meaning of the parameter values.

The parameters occupy part of the first two columns. You could start the remainder of your table below the parameters, but I find it more convenient to keep values in their own columns. We will skip column `C` to create a gap between our parameters and our sequence table. Column `D` will have the index values, column `E` will have the explicitly computed sequence values, and column `F` will have the recursively computed sequence values. Start by putting the labels in the first row. You might use `explicit` in `E1` and `recursive` in `E2`.

Create the values for the index in column `D` in the same way as described above. The explicit formula for our arithmetic sequence is given by

$$n \mapsto x_n = a + \beta n.$$

In the table, the index n is always found using relative table location. That is, the index used in `E2` will be found to the left in `D2`. The values for the parameters, however, will always be found in the same table positions.

Our parameters need to use **absolute addresses**, which spreadsheet indicate by putting a dollar sign in front of the column and row. To create the formula in `E2`, we will type `$B$1` to represent β and `$B$2` to represent a. This means our spreadsheet entry in `E2` is `$B$2+$B$1*D2`. If you copy and paste this into `E3`, you should see that only the cell representing the index is updated to `$B$2+$B$1*D3`. Selecting one of these cells and filling the rest of the column will finish generating the explicitly calculated values.

To create the column with recursively calculated values, we start by putting the initial value in `F2`. Because that is our parameter a, stored in `B2`, we enter the simple formula `=$B$2` to use that initial value. To apply the recursive relation

$$x_{n-1} \mapsto x_n = x_{n-1} + \beta,$$

we enter `=F2+$B$1` in `F3`. The rest of the column can be automatically filled with this formula.

If you did this correctly, you should see that the explicit and recursive columns contain the same values even though they were calculated in different ways. □

13.3.3 Computer Programming

Spreadsheets are useful, but typing a script can be even more efficient. In a spreadsheet, the use of cell references is a little awkward. There are some advanced techniques where you can reference cells by names instead of reference. However, if you want to adjust the size of your table and change the number of rows, you essentially need to repeat the dragging and filling steps.

Writing computer scripts in a programming language is one of the most efficient approaches. A free online tool called SageMath uses a scripting language based on the Python programming language. The online version of this text has interactive cells where you can try the scripts directly. Otherwise, you can use the following website and type the scripts: https://sagecell.sagemath.org/.

The idea of a scripting language is that the computer will store values in memory associated with names of your choosing. You can includes commands in your script to display the values or even to create graphs. To create a sequence of values, there are two fundamental ideas to understand: memory assignment and looping.

First is the idea of memory assignment. In a script, we tell the computer to perform a computation and then to store the result somewhere in memory. That memory location is assigned a variable name. The pattern for this step is the form `name = calculation`, where `name` is replaced by whatever name you want associated with the memory and `calculation` is replaced by the expression used in the calculation. The calculation can use variable names for any memory previously saved.

Below is a very short script. It will store a value of 3 in memory associated with the name `x`. It will then calculate $x^2 - 5x$ and store the result in memory associated with the name `y`. Finally, it will show the values of x and y as results. You will notice extra lines that begin with the # symbol. These are called comments and the script ignores them. We use comments in scripts to remind ourselves or to explain to others what is happening.

```
# Store the value for x
x = 3
# Calculate x^2-5x and store the value as y
y = x^2 - 5*x
# Show the results
show(x,y)
```

```
3
-6
```

For a sequence, we repeat the same process of calculation many times. Repeating a computation is called a **loop**. In a script, a loop is usually associated with a variable (a named memory location) that is associated with a given list of values. The basic scripting pattern in Python and Sagemath to repeat a computation for each value in a list is as follows.

```
for name in list:
    repeated block
```

The `name` is replaced by whatever variable name you want for the associated memory location. The `list` is replaced by a list of values that will be used for `name`. The `repeated block` is a collection of scripting commands, indented exactly four spaces to the right of the `for` statement. The script will take the values from the `list`, one at a time and in order, and will then go through the `repeated block` immediately after the value has been placed in the memory associated with `name`.

In Python or SageMath, we create a list of consecutive integers using the `range` function. The command `range(integer)`, where `integer` is replaced by any expression representing an integer, creates a list of consecutive integers starting at 0 and ending at the integer before the `integer` used. For example, `range(5)` creates the list $(0, 1, 2, 3, 4)$ and `range(3)` creates the list $(0, 1, 2)$. The following two scripts are equivalent, but the looped version is much more

efficient.

```
# Unlooped
n=0
x=3*n+1
show(n)
show(x)
n=1
x=3*n+1
show(n)
show(x)
n=2
x=3*n+1
show(n)
show(x)
```

```
# Looped
for n in range(3):
    x=3*n+1
    show(n)
    show(x)
```

We are almost ready to create a script that generates a table of values. The last step is creating a table with two values on the same line. We can do this with the `print` command and value formatting. In place of the `show` commands, we will use `print(format % (values))`, with `format` being a format string and `values` being a comma-separated collection of expressions to be formatted. An integer uses format string `%d` while a decimal value (a floating-point) uses format string `%f`. We can include a tab character using `\t`. Our improved script to generate a table with twenty values is now given.

```
# Loop to calculate a table of values.
for n in range(20):
    x=3*n+1
    print("%d\t%f" % (n,x))
```

The script can be quickly modified to generate a table as large or small as desired. If we want to create a graph of these sequence values, we need to create a table in memory. SageMath expects a table to be graphed as a list of points. We can modify our script to create an empty table before the loop, and then add (append) the individual points one at a time in the loop. In SageMath, a graphic is itself an object stored in memory, so we give it a name and then show it. A scatterplot is created using the `list_plot` command, which has an option to label the axes with a given list of names.

```
# Create an empty list named "dataPoints"
dataPoints = []
# Loop to calculate a table of values.
for n in range(20):
    x=3*n+1
    print("%d\t%f" % (n,x))
    # Append the current point to our list.
    # Each point is itself a list with two entries.
    dataPoints.append( [n,x] )
# Create the scatter plot with labels on the axes
myGraph = list_plot(dataPoints, axes_labels=["$n$","$x_n$"])
# Show the resulting figure with given width/height
```

```
show(myGraph, figsize=[4,3])
```

Once you have a script, such as the one above, you can just modify it for a new problem. A table for any explicit sequence can be calculated using the script above simply by modifying the first line in the repeated block to match the explicit formula. More values can be generated by modifying the `range` command. If the table is not wanted, just remove the `print` command.

To create a table for a recursive sequence, we need to make another modification. A recursive sequence uses the previous value of a sequence to compute the next value. In a script, we can use a memory assignment command using the variable name in the expression and then storing the result back into the original memory location. The old value is replaced by the new value.

The following script illustrates how to generate a table and a graph for a recursively defined sequence with recurrence

$$x_n = 1.05x_{n-1} - 10$$

and initial value $x_0 = 400$.

```
# Create an empty list named "dataPoints"
dataPoints = []
# Set the initial value.
# We use the name "x" for the currently-stored sequence value
x = 400
# Loop to calculate a table of values.
for n in range(20):
    # Because the loop starts at n=0, we print and append
        data first.
    print("%d\t%f" % (n,x))
    dataPoints.append( [n,x] )
    # Before we end the repeat block, we update for the next
        value.
    # The formula on the right uses the old value.
    # The answer replaces what was in memory for the next
        loop block
    x=1.05*x-10
# Create the scatter plot with labels on the axes
myGraph = list_plot(dataPoints, axes_labels=["$n$","$x_n$"])
# Show the resulting figure with given width/height
show(myGraph, figsize=[4,3])
```

When we have parameters in a model, we just need to add a few memory assignment commands at the beginning of the script. The following script is a generalization of the previous script for a recursive model

$$x_n = (1 + r)x_{n-1} - w$$

where r and w are parameters. When using variables in scripts, remember that the symbols must exactly match. Uppercase and lowercase letters are not the same—`w` and `W` are different.

```
# Assign parameters
r = 0.05
w = 10
# Create an empty list named "dataPoints"
dataPoints = []
# Set the initial value.
# We use the name "x" for the currently-stored sequence value
```

```
x = 400
# Loop to calculate a table of values.
for n in range(20):
    # Because the loop starts at n=0, we print and append
        data first.
    print("%d\t%f" % (n,x))
    dataPoints.append( [n,x] )
    # Before we end the repeat block, we update for the next
        value.
    # The formula on the right uses the old value.
    # The answer replaces what was in memory for the next
        loop block
    x=(1+r)*x-w
# Create the scatter plot with labels on the axes
myGraph = list_plot(dataPoints, axes_labels=["$n$","$x_n$"])
# Show the resulting figure with given width/height
show(myGraph, figsize=[4,3])
```

13.4 Dynamic Models Using Sequences

Overview. A dynamic model considers how quantities change in time. Sequences are often useful for such models. Many populations, including some plants and animals, reproduce on an annual cycle. It thus makes sense to census these populations on an annual basis so that the population is measured as a sequence. Financial models, such as paying off a loan or receiving amortized payments on a contract, involve interest accrual and periodic payments. In these cases, the balance of the loan or fund is a sequence relative to the number of periods. Even for quantities that do not change at such regular periods, we might take measurements at equal spacings for our own convenience. This also will result in a naturally observed sequence.

This section focuses on the formulation and interpretation of models using sequences. We often develop models by considering gain terms and loss terms. For example, in population growth, gains include births and immigration; losses include deaths and emigration. The development of a model involves creating formulas that compute or approximate the size of these terms based on the state of the system.

We will consider simple models that represent various rates of change. For a population model, we might consider the rate of births or the rate of deaths. For a financial model, we might consider the rate of interest or the rate of payment. Finding simple but meaningful models for different rates allows us to predict overall changes of the system. We will analyze the overall rate of change to understand the behavior of the model.

13.4.1 Population Models

Populations are frequently modeled using sequences. Many population are adapted to reproduce on an annual cycle, so it makes sense that such populations might be censused on an annual basis. Even for populations that reproduce throughout the year, it might still make sense to measure the population at the same time to measure year-over-year growth or decline. Fast growing populations like bacteria or some species of insects might be measured on even shorter time scales, such as hourly (bacteria) or weekly (insects). Sequences are appropriate in these circumstances because we are interested in the population size at specific times rather than at all possible times.

There are many variables that determine how a population changes. Some of these are unpredictable. Unpredictability or randomness is called **stochasticity**. Populations are subject to environmental stochasticity and demographic stochasticity. Environmental effects might include temperature fluctuations or variation in rainfall. Demographic stochasticity includes the randomness in number of offspring (e.g., seeds or eggs) or randomness in mortality or the timing of development.

In spite of these random effects, it is often the case that the size of the population can be approximately predicted knowing the population of the previous year. Recursive equations using projection functions provide the mathematical framework for modeling these sequences. We will use P as our population sequence and will develop the projection function f that relates consecutive values of the population as

$$P_n = f(P_{n-1}).$$

Population sizes change because individuals are entering and leaving the population. Growth in the population includes births as well as immigration. Decline in the population includes deaths as well as emigration. The quantities

measuring the number of births, deaths, and migration events per year are rates of change. For the state of our system, we include a variable representing the size of the population as well as a variable for each of the rates of change. In a more complex model, we might have variables for the number of individuals at different ages or stages of development. In principle, each state variable corresponds to its own sequence.

For example, consider a population that only changes from births and deaths. Let P be the size of the population, let B be the annual birth rate, and let D be the annual death rate. These variables are each measured annually and can be considered as sequences. Our index variable t will measure the time in years. The population sequence will satisfy a recurrence relation

$$P_{t+1} - P_t = B_t - D_t.$$

This equation simply states that the net change in the population, called the **forward difference** $\Delta P_t = P_{t+1} - P_t$, is equal to the number of births (a gain) minus the number of deaths (a loss). We usually consider a reference time at $t = 0$ so that the first value in the sequence would be P_0.

We will explore a variety of models based on different assumptions for how the rate of births and deaths relate to the size of the population.

Constant Rates. The simplest model would be that the numbers of births and deaths are constant values every year. For such a model, the forward difference is also constant, $\Delta P_t = \Delta P = B - D$. The resulting recursive equation becomes

$$P_{t+1} = P_t + \Delta P,$$

which we recognize as an arithmetic sequence with an increment ΔP. Using Theorem 13.2.8, we know the explicit formula for this sequence is given by

$$P_t = P_0 + \Delta P \cdot t.$$

Such a population either increases linearly (if $B > D$), decreases linearly (if $B < D$), or is constant (if $B = D$).

Example 13.4.1 This example considers a dynamic graph for constant birth and death rates. There are sliders for the birth rate B and the death rate D and the initial population is also adjustable. The resulting population sequence automatically updates to visualize the result. Such a model gives an arithmetic (linear) sequence.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 13.4.2

□

Constant Per Capita Rates. Of course, it is not realistic to think that a population has the same number of births, regardless of how large the population is. Rather, we would expect that the population will see more births when the size of the population itself is larger. The simplest model for this would be that the number of births is proportional to the size of the population. That is, we expect that there is a parameter b so that $B = b \cdot P$. This parameter is the proportionality constant and is called the **per capita birth rate**. The phrase "per capita" literally means per head. If we rewrote the equation relating B and P as

$$b = \frac{B}{P},$$

we see that the model is really saying that the total number of births in a year divided by the population that year is always the same constant. In a similar way, we might expect the number of deaths to be proportional to the population size,

$$D = d \cdot P,$$

where d is the **per capita death rate**.

Using constant per capita birth and death rates leads to a new model for the population. The recurrence relation is defined by

$$\Delta P_t = b \cdot P_t - d \cdot P_t = (b - d)P_t.$$

The recursive equation becomes

$$P_{t+1} = P_t + (b - d)P_t = (1 + b - d)P_t.$$

We recognize this as the equation of a geometric sequence with the ratio $\rho = 1+b-d$. Using Theorem 13.2.10, we know the explicit formula for this sequence is given by

$$P_t = P_0 \cdot (1 + b - d)^t.$$

This form of growth for the sequence is often called **Malthusian growth**.

Example 13.4.3 This example considers a dynamic graph for constant per capita birth and death rates. There are sliders for the per capita birth rate b and the per capita death rate d. The initial population is also adjustable. The resulting population sequence automatically updates to visualize the result. Such a model gives a geometric (exponential) sequence.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 13.4.4

□

In many circumstances, we may not be as interested in the individual values of the per capita birth and death rates b and d as we are in their difference $b-d$. This quantity is called the **net per capita growth rate** and is frequently denoted by the symbol $r = b - d$. In that case, the explicit formula for the Malthusian growth model can be rewritten in the same form as compounded interest,

$$P_t = P_0 \cdot (1 + r)^t.$$

That is, we can interpret r as the decimal value corresponding to percent change in the population year-over-year.

Example 13.4.5 Suppose a population of 2500 has 400 births and 250 deaths in the year. Compare the model for constant births and death rates with the model for constant per capita birth and death rates over the next five years.

Solution. The model for constant birth and death rates assumes that $B = 400$ and $D = 250$ are constants. The recursive equation for the population is then given by

$$P_{t+1} = P_t + 400 - 250 = P_t + 150.$$

In this model, the population increases by a net number of 150 individuals per year with an explicit formula given by

$$P_t = 2500 + 150t.$$

The model for constant per capita birth and death rates assumes the ratios $b = \frac{B}{P} = \frac{400}{2500} = 0.16$ and $d = \frac{D}{P} = \frac{250}{2500} = 0.1$ are constants. The recursive

equation for this model becomes

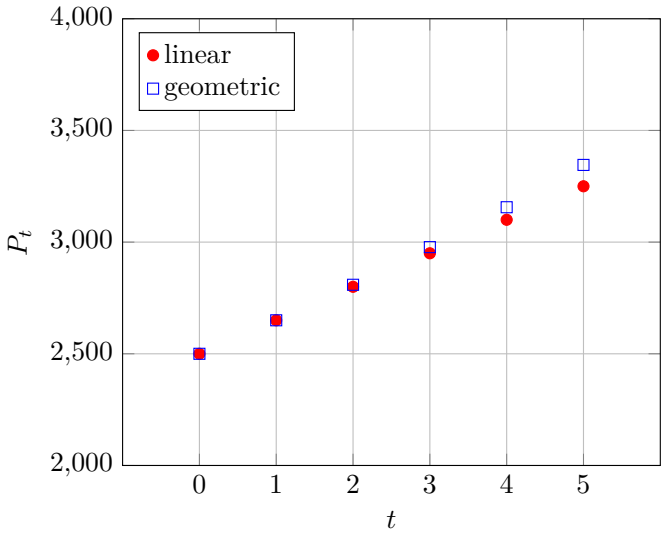
$$P_{t+1} = P_t + 0.16P_t - 0.1P_t = 1.06P_t,$$

with a corresponding explicit formula given by

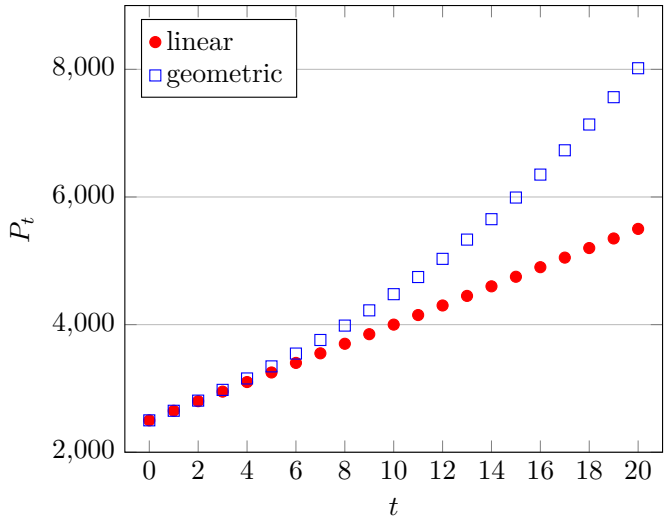
$$P_t = 2500 \cdot 1.06^t.$$

The table and figure below illustrate the growth of these two models. In the table for the Malthusian model (geometric growth), the model predicts non-integer values which I have shown to two decimal places. Of course, a population itself must be integer-valued. When working with mathematical models, we will leave the values exact until we are ready to interpret.

Year t	Linear $P_t = 2500 + 150t$	Geometric $P_t = 2500 \cdot 1.06^t$
0	2500	2500
1	2650	2650
2	2800	2809
3	2950	2977.54
4	3100	3156.19
5	3250	3345.56



The arithmetic and geometric models agree at the initial value and after the first year. But from that point, the geometric model steadily grows faster than the arithmetic model. The geometric model grows each year by the same percentage. Since the population itself is getting larger, the increment of growth is going to be larger each year. The two models diverge from one another even more dramatically as time progresses.



□

13.4.2 Other Models Using Sequences

The mathematical models introduced for sequences of populations can be applied and adapted to other situations. The ideas of per capita growth rates are mathematically the same as those for percentage growth or decay, such as appear in compounded interest investment problems. Any situation where a quantity increases or decreases by a fixed amount or by a fixed proportion or percentage will be modeled using a sequence defined in a similar way.

Example 13.4.6 Car Loan. Suppose you want to buy a car and obtain a loan for \$10000 that includes an annual interest rate of 3%. A bank charges interest in a way that the annual percentage rate is divided equally into the months. The monthly rate of $\frac{3}{12}\%$ applies to the remaining balance of your loan. If you make a monthly payment of \$250, find a model for your remaining loan balance. Use your model to determine when you pay off the loan and the total cost of the loan.

Solution. Start by identifying the relevant variables. Our main concern is the outstanding balance on the loan. Let us use the variable B to represent our sequence. The initial balance on the loan is $B_0 = 10000$.

Next, we identify all sources to changes in the balance. A payment P on the loan reduces the loan balance. Interest I on the loan causes the loan balance to increase. If t represents the number of months since the loan began, then we have a recurrence relation describing how the loan changes,

$$\Delta B_t = B_{t+1} - B_t = -P_t + I_t.$$

Solving for the new balance gives the recursive equation for the loan balance

$$B_{t+1} = B_t - P_t + I_t.$$

For the car loan, the monthly payment is a constant, $P_t = P = 250$. The interest accrued each month is proportional to and depends on the current balance, $I_t = 0.0025B_t$.

The model for our loan balance is given by the recursive equation and the initial value:

$$B_{t+1} = B_t - 250 + 0.0025B_t,$$

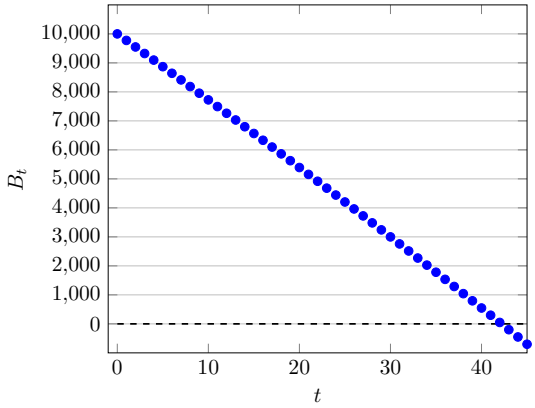
$$B_0 = 10000.$$

Our model is not arithmetic or geometric but a combination of the two. The projection function, $f : B_t \mapsto B_{t+1}$, is the linear function defined by

$$f(x) = x - 250 + 0.0025x = 1.0025x - 250.$$

The values for the loan balance are plotted below, along with a table of values showing when the loan would be paid off.

t	B_t
37	1289.212365
38	1042.435396
39	795.041485
40	547.029088
41	298.396661
42	49.142653
43	-200.734491
44	-451.236327
45	-702.364418



The computer calculations show that the last month in which there is a positive balance is month 42. That month, the remaining balance is \$49.14, which we will pay off completely in month 43. The model continues to use the same rule even after the loan is paid. This explains why the model predicts a negative balance.

We can compute the total amount paid for this loan. For 42 months, we paid \$250.00, followed by a final payment of \$49.14. The total cost of the loan is

$$42(250) + 49.14 = 10549.14.$$

Because the original cost of the car was \$10000, we paid \$549.14 in interest. □

Another example that follows similar dynamics is in mixing solutions.

Example 13.4.7 Mixing Solutions. Suppose that you have 2 liters of salt water that initially has 200 grams salt. You pour out 0.5 liters from your bottle, replace it with a solution of pure water, and then shake well. This is repeated, making your bottle less and less salty. Use a sequence to describe the saltiness of the solution as a function of the number of dilutions.

Solution. Start by identifying the variables. We are interested in the amount of salt in the water. Use S as the variable representing the sequence of total salt (grams) in the water. The concentration C would be $S/2$ (grams per liter). The initial value is $S_0 = 200$. Let n be the variable representing the number of dilutions performed, which we will use as our index for the sequence.

Next, identify what causes the change in the solution. Every dilution, a fraction of the solution is removed, $\frac{0.5}{2}$, along with all salt in that volume. Since the bottle is well-mixed, we have a fourth of the salt remaining taken out of the bottle. The replacement water is pure, so no new salt is added back in.

Based on our discussion, the recursive model for the salt includes only a single loss term:

$$S_n = S_{n-1} - 0.25S_{n-1} = 0.75S_{n-1},$$
$$S_0 = 200.$$

Thus our model is a simple geometric sequence. We have an explicit solution using Theorem 13.2.10:

$$S_n = 200 \cdot 0.75^n.$$

□

13.4.3 Nonlinear Projection Functions

Much more interesting (and surprising) dynamics occur when a sequence is defined by a nonlinear projection function. To motivate one example where this might occur, we return to the ideas of per capita growth for a population.

Recall that our earlier discussion used the idea that the net per capita growth rate was a constant and did not depend on the population size. That is, the number of births and deaths were simply proportional to the total population size. However, this is ultimately not physically possible. When a population gets too large, resources are limited and the population will eventually be unable to sustain such rapid growth. Either the per capita birth rate will decrease or the per capita death rate will increase (or both). Either way, the net per capita growth rate $r = b - d$ will need to decrease as a function of population size. Once we say that one variable decreases with respect to another variable, we can use a mathematical model to capture that idea.

In this case, we want r to be a decreasing function of the population P, $P \mapsto r$. The simplest such model would be a linear function with a negative slope. If we had enough data, we could plot points (P, r) and find a line of best fit. For now, we will use a parametrized model,

$$r(P) = r_0 - \alpha P,$$

where the parameter r_0 is called the **intrinsic net per capita growth rate** (because that is the growth rate for a very small population before resources are limited) and $\alpha > 0$ is the magnitude of the negative slope.

A better parametrization uses the formula for a line given both the intercepts, $(P, r) = (0, r_0)$ and $(P, r) = (K, 0)$, so that

$$r(P) = r_0\big(1 - \frac{P}{K}\big).$$

The value K is called the **carrying capacity** because for $P > K$, the growth rate will be negative (net decrease in population). That is, for any population greater than K, the available resources are inadequate to support such a population.

The final population growth model is based on the model we just found. Recall that a population grows with a recursive model

$$P_{t+1} = P_t + rP_t,$$

where r is the net per capita growth rate. Using the model given above for $r = r_0(1 - \frac{P}{K})$, we construct a nonlinear model for the population sequence,

$$P_{t+1} = P_t + r_0\big(1 - \frac{P_t}{K}\big)P_t = (1 + r_0)P_t - \frac{r_0}{K}P_t^2.$$

This model is called the **discrete logistic model**. The model only makes sense when the parameter r_0 is in the interval $r_0 \in (0, 3)$.

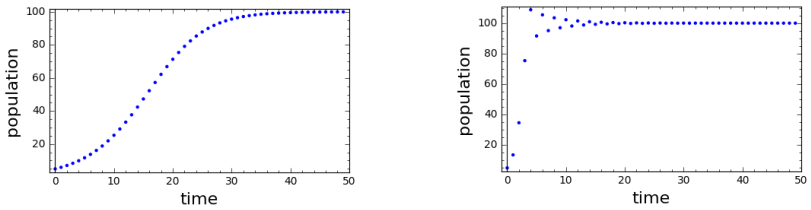
Different behaviors for the population arise, depending on the values of the parameter r_0. A Sage script is provided below that will generate plots of the population sequence for values of the parameters that you specify. The

following graphs were generated by Sage using an initial value $P_0 = 5$ and parameter values $K = 100$ (all plots) and $r_0 = 0.2$, $r_0 = 1.8$, $r_0 = 2.2$ and $r_0 = 2.6$. In addition, a dynamic graph is given where you can adjust the parameters using sliders.

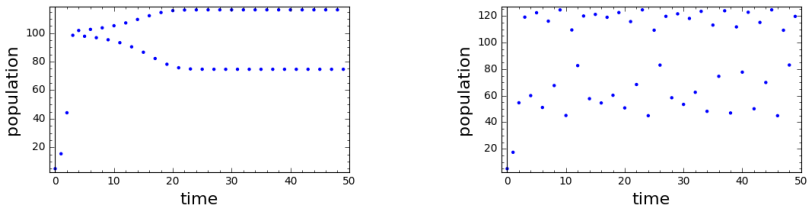
```
# Set the parameter values
r0 = 0.2   # Change this number for different behavior
K = 100    # Change this number for carrying capacity

# Set the initial value
P = 5
# Create what is initially an empty list
data = []
Tmax = 50  # Number of data points to create

# Use a loop repeating the balance update
for t in range(Tmax):
    data.append( (t,P) )
    # Now update the population
    P = P + r0*P - r0/K*P^2
list_plot(data, frame=True,
    axes_labels=['time','population'])
```



(a) $r = 0.2$ **(b)** $r = 1.8$



(c) $r = 2.2$ **(d)** $r = 2.6$

Figure 13.4.8 Logistic growth with $K = 100$ with $P_0 = 5$ and varying values of r.

A deprecated JSXGraph interactive demonstration goes here in interactive output.

Figure 13.4.9 Dynamic graph of the discrete logistic model with variable r and P_0.

13.4.4 Summary

- Sequences can be used to model any quantities that are observed at regular intervals, with populations and financial balances as typical examples.
- A common strategy for building a recurrence model is to add rates of

gain and subtract rates of loss,

$$\Delta x_t = x_{t+1} - x_t = +\text{Gains} - \text{Losses}.$$

- For a population, common gain rates include births and immigration; common loss rates include deaths and emigration.
- A per capita rate is the ratio of the total rate to the population size. It represents the contribution toward the total rate for one individual. The total rate equals the per capita rate times the population size.
- Finding models for individual rates or per capita rates in terms of the population size allows us to formulate a recursive equation for the population sequence. Simple examples are to assume constant rates or constant per capita rates. More complex models might fit models for density-dependent per capita rates.
- We use computers to find values numerically for a sequence based on the recursive equation. This might be through a spreadsheet or through a scripting language like Python. These data allow us to create graphs.

13.4.5 Exercises

1. A population of annual plants has all plants die every year. Before dying, each plant releases 20 seeds which will grow the following year.

(a) Find a recurrence equation for the population.

(b) If $P_0 = 10$, find P_1 and P_2 by hand.

(c) Find an explicit formula for the sequence P_t.

2. A population has constant per capita birth and death rates. When the population is $P = 1000$, there are $B = 200$ births per year and $D = 250$ deaths per year. In addition, this population has a constant immigration rate of $I = 300$ individuals per year.

(a) Find a recurrence equation for the population.

(b) If $P_0 = 1000$, find P_1 and P_2 by hand.

(c) Use a computer to generate a plot of the sequence (t, P_t) for $t = 0, \ldots, 50$.

3. You put \$500 in a bank which pays 1% interest, compounded annually.

(a) Find a recurrence equation for the balance B of your account. What is the initial value, B_0?

(b) Compute B_1 and B_2 by hand.

(c) Find an explicit formula for the sequence B_t.

4. You inherit \$50,000, which you immediately invest. Your investment fund guarantees an annual interest payment of 2%, compounded annually. You withdraw \$2,000 each year to spend.

(a) Find a recurrence equation for the balance of your fund F. What is the initial value, F_0?

(b) Compute F_1 and F_2 by hand.

(c) Use a computer to generate a table and a plot of the sequence (t, F_t) for $t = 0, \ldots, 40$.

(d) How long will the fund last? What was the total value of the inheritance?

5. You purchase a house with a home loan of \$350,000 with an annual interest rate of 4%, which accrues monthly. You choose to make a monthly payment of \$1,500.

(a) Find a recurrence equation for the balance of your loan L. What is the initial value, L_0?

(b) Compute L_1 and L_2 by hand.

(c) Use a computer to generate a table and a plot of the sequence (t, L_t) for a long enough period to determine when the loan is completely paid.

(d) When will you pay off the house loan? What will have been your total cost? How much interest will you have paid?

(e) If you increase the monthly payments to \$1,600, when will you pay off the loan? How much interest will you have paid?

6. A pond with 100,000 gallons of water has a stream flowing in and out at a rate of 5,000 gallons per day. One day, the stream flowing in is polluted with a chemical of 200 grams per gallon. Assuming that the pond mixes the water quickly, develop a model for the amount of chemical in the pond as a daily sequence.

(a) State your variables.

(b) What is your model for how much chemical enters the pond each day?

(c) What is your model for how much chemical leaves the pond each day?

(d) State your recurrence relation and the initial value for your sequence. Determine the resulting recursive equation.

(e) Find an explicit formula for your sequence. How much chemical is in the pond after 30 days?

Appendix A

Mathematics Foundations

A.1 Numbers, Sets and Arithmetic

Numbers started as a conceptual way to quantify count objects. Later, numbers were used to measure quantities that were extensive, such as the geometric ideas of length, area and volume. Arithmetic was developed to provide a numeric representation of physical operations. Combining two quantities or collections corresponds to addition. Repeated addition corresponds to multiplication. Repeated multiplication corresponds to powers or exponents. As these ideas developed, inverse operations were invented to help solve problems, including subtraction, division and roots. However, the introduction of each inverse operation required an extension of the idea of number.

A.1.1 Integers, Rational Numbers and Real Numbers

Numbers conceptually begin with the **natural numbers**, which are the numbers $1, 2, 3, \ldots$. The set of all natural numbers is represented by the symbol $\mathbb{N}$:

$$\mathbb{N} = \{1, 2, 3, \ldots\}.$$

If we include the number 0, we get all possible counting numbers, represented by the symbol $\mathbb{N}_0$:

$$\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}.$$

Both of these sets have an infinite number of elements because there is no upper bound (i.e., for any number you find, there is always a number greater).

The natural and counting numbers are used most basically for ordering and counting elements in sets or collections of objects. For example, consider a set consisting of the basic suits of a standard deck of playing cards, namely hearts, diamonds, spades, and clubs, $\{\heartsuit, \diamondsuit, \spadesuit, \clubsuit\}$. We can count the number of elements by associating each element in the set with one of the natural numbers in order:

$$\begin{aligned} 1 &\mapsto \heartsuit, \\ 2 &\mapsto \diamondsuit, \\ 3 &\mapsto \spadesuit, \\ 4 &\mapsto \clubsuit. \end{aligned}$$

This ordering (which is admittedly arbitrary) allows us to refer to the first, second, third or fourth element in our set. When numbers are used to order elements of a set in this way, we are thinking of numbers as **ordinal numbers**. Because the greatest number used in our ordering was 4, we can say that the number of elements in our set is 4. When numbers are used to count the number of elements in a set, we are thinking of numbers as **cardinal numbers**. Both of these ideas can be extended to sets with infinities of elements. It is not in the scope of this text to deal with these issues, but they are typically addressed in more general discussions of set theory.

However, the elementary ideas of the arithmetic operations of addition and multiplication are often introduced using the ideas of counting. Addition is first defined by joining sets of known size and asking how many elements are in the combined set. For example, 3+5 is interpreted in this context as joining a set of three elements, say $\{a, b, c\}$, to another set of five elements, say $\{z, y, x, w, v\}$, to get a combined set $\{a, b, c, v, w, x, y, z\}$. The size of this new set is 8. The equation

$$3 + 5 = 8$$

is interpreted in this context, namely that "the number of elements in a set formed by joining a set with 3 elements and a set with 5 elements" (3+5) is the same as "the number of elements as a set formed from 8 elements" (8). Multiplication is first defined as adding a certain number of groups of the same size. For example, 3×5 is interpreted as creating a set consisting of 3 groups of 5 elements, which is a new set with 15 elements. This leads to the equation

$$3 \times 5 = 15.$$

Once the arithmetic operations of addition and multiplication are introduced, inverse operations of subtraction and division soon follow. Subtraction corresponds to taking away from a set with $5 - 3$ being interpreted as starting with a set of 5 elements and removing a subset of 3 elements, so that

$$5 - 3 = 2.$$

Division corresponds to determining how many groups of a certain size are in a particular set, with $12 \div 4$ counting the number of groups of size 4 that can be formed from a set of size 12, so that

$$12 \div 4 = 3.$$

Inverse operations have the property that when performed consecutively, the original value remains unchanged. That is, $a + b - b = a$ for any values of a and b because you join and then remove a set of size b to a set of size a, resulting in a set of the original size a. Similarly, $a \times b \div b = a$ because the set $a \times b$ has a groups of sets with b elements.

The problem that arises is that using only natural or counting numbers, there are expressions that have no valid interpretation. For example, the inverse property suggests that $3 - 5 + 5 = 3$, but the intermediate calculation $3 - 5$ does not make sense using counting numbers because there is no interpretation for how to remove 5 elements from a set with only 3 elements. Similarly, although the inverse property suggests $5 \div 3 \times 3 = 5$, the intermediate calculation $5 \div 3$ has no whole number interpretation because grouping 5 into sets of size 3 results in one group of 3 and another group of 2 (the remainder).

In order to resolve this complication, the idea of number itself is extended. Negative numbers resolve the challenge for subtraction. The set of all integers, also called the whole numbers, introduces both positive and negative counting numbers and is represented by the symbol $\mathbb{Z}$:

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}.$$

Although we originally thought of subtraction as the inverse operation of addition, the introduction of negative numbers motivated the idea of a number itself having and inverse with respect to addition, or an **additive inverse**. Every positive and negative number of the same size are inverses because the add to zero,

$$a + -a = 0.$$

With additive inverses, the concept of subtraction is equivalent to addition by an additive inverse,

$$a - b = a + -b.$$

The advantage of this perspective is that it makes clear how to subtract negative numbers — subtracting a negative number is defined as adding the inverse of the negative number, or adding the corresponding positive number.

Just as subtraction led to the development of negative numbers, division motivates the need to extend numbers from just the integers to rational numbers. As soon as we leave the world of whole numbers, our sense of arithmetic actually changes from counting elements in a set to measuring divisible quantities (like length or volume). The standard representation of numbers on a number line illustrates this directly by thinking of numbers as measuring a directed length from an origin (the number 0). There always must be a **unit length** which corresponds to the distance between 0 and 1. Positive numbers are to the right and negative numbers are to the left.

A new interpretation of number requires a new interpretation of arithmetic. Addition of numbers corresponds to combining lengths, with 3+5 meaning we find the number which is found by starting at 0 (the origin), moving three units to the right (to find 3) and then moving five more units to the right (to add 5). Since this the same as the number 8 (starting at 0 and moving 8 units to the right), we know 3+5=8. Multiplication (by integers) will still mean repeated addition, just repeating the displacement interpretation of addition instead of groups.

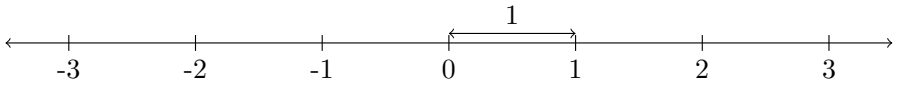


Figure A.1.1 The integers are placed on the real number line with a spacing defined by the unit length.

If we consider the property of consecutive inverse operations, we know that we should get $1 \div 5 \times 5 = 1$. So if we think about the intermediate value $a = 1 \div 5$ (which is not an integer), we can see that it is a value such that $a \times 5 = 1$. In the geometric interpretation, a is a length such that when it is repeated five times, we recover the unit length. This is the unit fraction $\frac{1}{5}$, which is also called the **multiplicative inverse** (or reciprocal) of the integer 5. Other fractions have a similar interpretation, such as $3 \div 4 = \frac{3}{4}$ (dividing three into fourths) being the length such that when it is repeated four times is equivalent to three units.

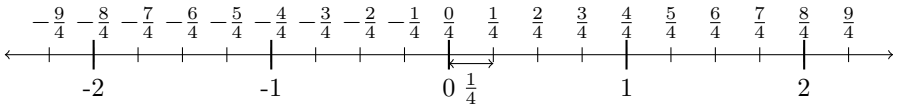


Figure A.1.2 Rational numbers are placed on the real number line using a fractional unit length based on the denominator.

Just as subtraction was found (or defined) to be equivalent to addition by an additive inverse, division is also equivalent to multiplication by a multiplicative inverse. Given any non-zero number $a \neq 0$, the multiplicative inverse $\div a$ is that number so that

$$a \cdot \div a = 1.$$

Then division $a \div b$ is defined by

$$a \div b = a \cdot \div b.$$

This process allows us to define the rational numbers $\mathbb{Q}$. The rational numbers are formed by considering all of the integers, their multiplicative inverses, and all sums and products of those values. It is most commonly defined by

$$\mathbb{Q} = \{p \div q : p \in \mathbb{Z}, q \in \mathbb{N}\}.$$

That is, it consists of all fractions defined by integers.

My goal is not to provide an exhaustive explanation of arithmetic and these representations. That would require, for example, an explanation of what it means to multiply and divide negative numbers. However, let it suffice to say that multiplying two negative numbers together will be a positive number while multiplying a positive number and a negative number will result in a negative number.

Other mathematical operations introduce the need for even more extensions to the idea of number. For example, the mathematical operation of squaring a number has an inverse operation of the square root. The square of a rational number is still a rational number, but the square root of a rational number is not necessarily rational. The most famous historical example is $\sqrt{2}$, which was proved not to be a rational number (according to legend) by the Greek philosopher Pythagoras. The existence of irrational numbers was a closely guarded secret by his followers, the Pythagoreans. The set of **real numbers** is the set of both rational and irrational numbers and is represented by the symbol $\mathbb{R}$. Complex numbers extend the real numbers to include square roots of negative numbers by introducing $i = \sqrt{-1}$ and is defined as

$$\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}.$$

Every real number $a \in \mathbb{R}$ is also complex with $b = 0$.

We will be working almost exclusively with the real numbers. So we very often think in terms of the real number line, which is a continuous and connected curve. Every point on the number line corresponds to a particular real number. Locations correspond to rational numbers if they can be exactly represented using fractional units. An irrational number can never be exactly represented using fractional units.

A.1.2 Sets and Intervals

A **set** is a mathematical collection of objects. The objects that are in the set are called **elements** of the set. Set notation uses curly braces { and } with a description of the elements that belong to the set. When the set has a finite number of elements, we can just list them between the braces. Like other mathematical objects, we can use symbols to represent the set in the same way that variables can be represented by symbols.

Example A.1.3 The set that contains the odd digits could be written

$$O = \{1, 3, 5, 7, 9\}.$$

The set that contains the even digits could be written

$$E = \{0, 2, 4, 6, 8\}.$$

The set that contains the prime digits could be written

$$P = \{2, 3, 5, 7\}.$$

Note: The symbols (names) for these sets, O, E, P, are just used as examples. We could have used any other symbols that might have been convenient. □

The symbol $\in$ is a logical operator used to say that an element is in a set. Using the example sets above, we would say $3 \in O$ (read as "3 is in O") since 3 is an odd digit. But $3 \notin E$ (read as "3 is not in E").

Most useful sets can not be described by listing all of the elements. Instead, we define sets according to a logical rule that describes when an element is in

the set. Such sets start with what is called a **universal set** that classifies what type of elements are being considered. For example, a set containing numbers might have a universal set $\mathbb{Z}$ (only integers) or $\mathbb{R}$ (all real numbers). A typical set would be defined with a statement like the following,

$$A = \{x \in U : \text{logical statement about } x\},$$

where A is the symbol for the set being defined, U is the universal set, x is a symbol being used to represent an arbitrary element of U, and the statement is how you decide if $x \in A$.

Example A.1.4 To define the set of all real numbers between -1 and 1, we would write

$$A = \{x \in \mathbb{R} : -1 < x < 1\}.$$

To define the set of positive real numbers, we would write

$$B = \{x \in \mathbb{R} : x > 0\}.$$

□

We can combine sets to create new sets using unions and intersections. Suppose that A and B represent any two sets. The **union** of the sets, written $A \cup B$, is the combination of sets that includes elements that are in at least one of the sets:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

The **intersection** of the sets, written $A \cap B$, is the combination of sets that includes only elements that are in both of the sets, also thought of as the overlap of the sets:

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

Example A.1.5 Using the sets defined in the examples above, we have the following statements. The union of O (odd digits) and P (prime digits) gives

$$O \cup P = \{1, 2, 3, 5, 7, 9\},$$

which is the same as $O \cup \{2\}$. The intersection of A (real numbers between -1 and 1) and B (positive real numbers) gives

$$A \cap B = \{x \in \mathbb{R} : 0 < x < 1\}.$$

□

Because most sets being studied in calculus come from the real numbers, the universal set is often not explicitly stated. So the following represent the same set:

$$\{x \in \mathbb{R} : 1 < x < 3\} = \{x : 1 < x < 3\}.$$

One of the most common type of sets in calculus is the **interval**. An interval is a set of real numbers representing a connected segment of the real number line. Intervals are defined by their end points. An **open interval** does not include the end points while a **closed interval** does include the end points. An open interval with end points a and b with $a < b$ is represented by the notation using round parentheses

$$(a, b) = \{x : a < x < b\}.$$

A closed interval with the same end points is represented by similar notation using square brackets

$$[a, b] = \{x : a \leq x \leq b\}.$$

If only one end point is included, then the notation uses both parentheses and brackets:

$$[a, b) = \{x : a \leq x < b\},$$
$$(a, b] = \{x : a < x \leq b\}.$$

A set that consists of disjoint intervals can be represented with interval notation using unions. Consider, for example, the interval $[1, 5]$ and remove the two values 2 and 4. This is no longer a single interval but consists of three disjoint intervals, namely $[1, 2)$, $(2, 4)$, and $(4, 5]$. We can write this set as a union of the three intervals.

$$\{x \in [1, 5] : x \neq 2 \text{ and } x \neq 4\} = [1, 2) \cup (2, 4) \cup (4, 5].$$

Notice how the set defined on the left uses curly brackets because the set is defined using a rule, but the interval notation on the right does not include curly brackets because the interval notation defines everything about the sets.

A.1.3 Algebra Properties

One of the guiding principles in interpreting arithmetic in different representations is that we expect the fundamental properties of arithmetic to be satisfied. These include the commutative and associative properties of addition and multiplication and the distributive properties of multiplication over addition. That is, for every system of numbers, we expect the following properties to hold for any numbers a, b, c.

Table A.1.6 Properties of arithmetic of real numbers.

Property	Description
$a + b = b + a$	Addition is Commutative
$(a + b) + c = a + (b + c)$	Addition is Associative
$a + 0 = a$	Zero is Additive Identity
$a + -a = 0$	Additive Inverse Property
$-a = -1 \cdot a$	Finding Additive Inverse
$a \cdot b = b \cdot a$	Multiplication is Commutative
$(a \cdot b) \cdot c = a \cdot (b \cdot c)$	Multiplication is Associative
$a \cdot 0 = 0$	Multiplication by Zero
$a \cdot 1 = a$	One is Multiplicative Identity
$a \cdot \div a = 1$	Multiplicative Inverse Property
$\div a = \frac{1}{a},\ a \neq 0$	Finding Multiplicative Inverse
$a \cdot (b + c) = a \cdot b + a \cdot c$	Left Distributive Property
$(a + b) \cdot c = a \cdot c + b \cdot c$	Right Distributive Property

These basic rules establish the basic properties of arithmetic (and algebra) over the real numbers. (Advanced mathematics considers other structures that have some but not all of the same properties in a subject called abstract algebra.) Other consequences of these properties are often used in algebra. We list some of these below for reference.

Theorem A.1.7 *If* $a \cdot b = 0$, *then* $a = 0$ *or* $b = 0$.

Theorem A.1.8 $(a+b)\cdot(c+d) = ac + ad + bc + bd$ *(FOIL)*

A.2 Algebra Review

A.2.1 Lines and Linear Functions

Lines are perhaps the most important elementary geometric object. A line captures the idea of following a given direction without turning. In ordinary language, we sometimes think of a line as a smooth curve that we could draw. Mathematically, a line would then be a *straight line* or a straight curve that does not bend. Algebraically, we can define a line using an equation involving two variables. This section reviews the basic principles of the algebraic properties of lines.

Definition A.2.1 General Equation of a Line. Every line in the (x, y) plane can be described as the set of points (x, y) that satisfy an equation

$$Ax + By = C \tag{A.2.1}$$

where A, B and C are constants. ◊

There are some special cases that describe horizontal and vertical lines. The (x, y) plane uses x as the horizontal axis (independent variable) and y as the vertical axis (dependent variable). So a horizontal line is a line where the dependent variable is constant while a vertical line is a line where the independent variable constant.

Definition A.2.2 Horizontal Line. A **horizontal line** in the (x, y) plane is the set of points that satisfy an equation

$$y = k \tag{A.2.2}$$

where k is a constant. ◊

Definition A.2.3 Vertical Line. A **vertical line** in the (x, y) plane is the set of points that satisfy an equation

$$x = h \tag{A.2.3}$$

where h is a constant. ◊

All other lines have an equation that involves both variables. We often wish to think of the line as describing the dependent variable as a function of the independent variable. These equations involve the calculation of the slope, which represents a rate (or ratio) of change.

Definition A.2.4 Slope as Rate of Change. Given any two points (x_1, y_1) and (x_2, y_2) on a non-vertical line, the change in the dependent variable $\Delta y = y_2 - y_1$ is proportional to the change in the independent variable $\Delta x = x_2 - x_1$, written $\Delta y = m \cdot \Delta x$. The proportionality constant m is called the **slope** of the line, calculated as the ratio of changes (rate of change)

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}. \tag{A.2.4}$$

◊

Knowing the slope and one point is enough to quickly find an equation of a line.

Definition A.2.5 Point–Slope Equation of Line. Given that a line has slope m and passes through a point $(x, y) = (h, k)$, every point on the line

satisfies the equation

$$y = m\,(x - h) + k. \tag{A.2.5}$$

We interpret k as the starting value for y and the expression $\Delta y = m\,(x - h)$ as the change in y given the change in x, $\Delta x = x - h$. ◊

A special case of the point–slope equation of the line occurs when the point is on the y-axis or, in other words, is a y-intercept.

Definition A.2.6 Slope–Intercept Equation of Line. Given that a line has slope m and passes through a y-intercept $(x, y) = (0, b)$, every point on the line satisfies the equation

$$y = mx + b. \tag{A.2.6}$$

◊

Remark A.2.7 Preparatory mathematics courses often emphasize the slope–intercept equation of a line as if it were the most important. However, the point–slope equation is the preferred equation to use in almost every circumstance.

Another special case of the point–slope equation of a line is when we know the slope and the x-intercept.

Definition A.2.8 Slope and X-Intercept Equation of Line. Given that a line has slope m and passes through an x-intercept $(x, y) = (a, 0)$, every point on the line satisfies the equation

$$y = m(x - a).. \tag{A.2.7}$$

◊

A.2.2 Quadratic Polynomials

Definition A.2.9 A **quadratic polynomial** in a variable x is an algebraic function that is equal to a formula of the form

$$f(x) = ax^2 + bx + c, \tag{A.2.8}$$

where a, b and c are constants called **coefficients**. ◊

The graph of a quadratic function, $y = ax^2 + bx + c$, is a **parabola**. Such a parabola has a mirror symmetry across a vertical line that passes through its **vertex** $x = -\frac{b}{2a}$. Depending on whether the vertex is above, on or below the x-axis and whether the parabola opens up or down, the graph can cross the x-axis twice, once or never. The location of these points are called x-intercepts, **roots** or **zeros** of the function. The values of the roots can always be found using the (((Unresolved xref, reference "thm-quadratic-formula"; check spelling or use "provisional" attribute)))quadratic formula.

Zeros are closely related to factoring. If we know the zeros, then we can immediately rewrite the polynomial in a factored form. On the other hand, if we know the factors, then we can quickly solve for the zeros without using the quadratic formula. This is a consequence of the fundamental properties of numbers in Theorem Theorem A.1.7.

Theorem A.2.10 Factor–Root Theorem for Quadratics. *A quadratic function $f(x) = ax^2 + bx + c$ with real roots $x = r$ and $x = s$ ($r = s$ is possible) is equal to the factored equation*

$$f(x) = a(x - r)(x - s). \tag{A.2.9}$$

A quadratic polynomial that has complex roots is called **irreducible** because it can not be rewritten in a factored form involving only real roots.

There are some tricks to factoring that can be useful to know. Factoring is the reverse process of multiplying by distribution, so we start by noticing what happens when you multiply out two simple factors:

$$(x+a)(x+b) = x^2 + (a+b)x + ab.$$

Notice that the coefficient of x is the sum $a+b$ and the constant term is the product ab. When trying to factor a quadratic, look for numbers that multiply to give the product term and add to give the coefficient of x. This is often a matter of trial and error.

Knowing one root $x = r$ of a quadratic $f(x) = ax^2+bx+c$ so that $f(r) = 0$, we know that $x - r$ is a factor. The other factor can be determined easily.

Theorem A.2.11 Using Roots to Factor Quadratics. *Suppose that $x = r$ is a root of $f(x) = ax^2 + bx + c$ so that $f(r) = 0$. Then*

$$f(x) = (x-r)(ax+d) \tag{A.2.10}$$

where $d = b + ar = -c/r$.

Synthetic division is a procedure that works for quadratics as well as higher order polynomials. This procedure uses a table that starts with the coefficients on the first row. For a more thorough discussion for higher-order polynomials, see Algorithm A.2.19.

Algorithm A.2.12 Synthetic Division (Quadratics). *To divide a quadratic polynomial $f(x) = ax^2 + bx + c$ by the factor $x - r$ (proposed root $x = r$), we will apply the following steps.*

1. *Create a table that will have three rows and three columns. The first row will have the coefficients a, b and c in the three columns. The first entry of the second row will always be 0. So the start of the table will look like the following.*

a	b	c
0	____	____
____	____	____

2. *We finish a column, starting with the first column, by adding the values from the first and second rows, which in this case gives a again.*

a	b	c
0	____	____
a	____	____

3. *We next update the second row of the next column by multiplying the most recent value in the third row by the proposed root value r, which in this case gives ar.*

a	b	c
0	ar	____
a	____	____

4. *Add the values in the second column to update the third row to define the new coefficient $d = ar + b$ (recall Theorem A.2.11) and multiply by r to update the second row.*

a	b	c
0	ar	$ar^2 + br$
a	$d = ar + b$	____

5. *When we finish updating the third row by adding the values in the third column, we discover that the last entry in the third row corresponds to* $f(r) = ar^2 + br + c$.

$$\begin{array}{lll} a & b & c \\ 0 & ar & ar^2 + br \\ a & d = ar + b & f(r) = ar^2 + br + c \end{array}$$

Once the table is complete, we interpret the values in the third row as giving coefficients of the factored polynomial along with the value of the polynomial at $x = r$ *(called the remainder) and can write*

$$f(x) = (x - r)(ax + d) + f(r).$$

If $f(r) = 0$ *(remainder is 0), then this is an actual factorization*

$$f(x) = (x - r)(ax + d).$$

Because synthetic division is quick, this can be a simple way to test for roots and factor simultaneously.

Every quadratic can be rewritten in a form $y = a(x - h)^2 + k$ where (h, k) is the vertex of the parabola and a is the leading coefficient and scaling factor. The process of rewriting a quadratic $y = ax^2 + bx + c$ in this vertex form is called **completing the square**. It is based on noticing what happens with expanding the square of a binomial, $(x + a)^2 = x^2 + 2ax + a^2$. The strategy involves adding a term to form a perfect square and subtracting the same term to guarantee the expression does not change.

Algorithm A.2.13 Completing the Square. *A quadratic* $y = ax^2 + bx + c$ *can be rewritten in terms of its vertex by completing the following steps.*

1. *Factor the leading coefficient from the leading two terms*

$$y = a(x^2 + \frac{b}{a}x) + c.$$

2. *Think of the* x *term as being double half its value and add and subtract the square of the half-value:*

$$y = a\Big(x^2 + 2\frac{b}{2a}x + (\frac{b}{2a})^2 - (\frac{b}{2a})^2\Big) + c.$$

3. *Recognize the square of a binomial, group those terms, and regroup the remaining terms:*

$$y = a\Big((x + \frac{b}{2a})^2 - (\frac{b}{2a})^2\Big) + c = a(x + \frac{b}{2a})^2 + \frac{-ab^2}{4a^2} + c.$$

4. *Interpret the results: The vertex is* (h, k) *where* $h = -\frac{b}{2a}$ *(because vertex form uses* $(x - h)^2$*) and* $k = -\frac{b^2}{4a} + c$. *The leading coefficient* a *is a scaling factor that determines the steepness of the parabola and whether the parabola opens up* $(a > 0)$ *or down* $(a < 0)$.

Example A.2.14 Complete the square for $3x^2 - 4x + 1$.

Solution.

1. Group the non-constant terms and factor out the leading coefficient.

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

2. Recognize the coefficient $-\frac{4}{3}$ as double $-\frac{2}{3}$ and use this to complete the square.

$$\begin{aligned} 3x^2 - 4x + 1 &= 3\Big(x^2 + 2 \cdot \frac{-2}{3}x\Big) + 1 \\ &= 3\Big(x^2 + 2 \cdot \frac{-2}{3}x + (\frac{-2}{3})^2 - (\frac{-2}{3})^2\Big) + 1 \\ &= 3\Big(x^2 - \frac{4}{3}x + \frac{4}{9}\Big) - 3(\frac{4}{9}) + 1 \\ &= 3\Big(x - \frac{2}{3}\Big)^2 - \frac{4}{3} + 1 = 3\Big(x - \frac{2}{3}\Big)^2 - \frac{1}{3}. \end{aligned}$$

3. Interpret the results as saying $h = \frac{2}{3}$ (because the completed square is always of the form $(x-h)^2$) and $k = -\frac{1}{3}$. Thus the vertex of the parabola is at $(\frac{2}{3}, -\frac{1}{3})$. The leading coefficient $a = 3$ indicates that the parabola opens up and is three times steeper than the standard parabola $y = x^2$.

□

A.2.3 Polynomials

Linear and quadratic formulas are special cases of polynomials. This section gives an overview of principles about polynomials that are likely to appear in calculus. First, we introduce some basic definitions.

Definition A.2.15 Monomials. A **monomial** is an expression that is a constant multiple of a variable raised to a non-negative integer power, ax^k, where $k = 0, 1, 2, 3, \ldots$ and $a \in \mathbb{R}$.

Examples include $4x^2$ (with $a = 4$ and $k = 2$), $\frac{1}{3}x^7$ (with $a = \frac{1}{3}$ and $k = 7$) and 3 (where $a = 3$ and $k = 0$). The following are not monomials: $3\sqrt{x} = 3x^{1/2}$ (since not an integer power) and $\frac{3}{x^2} = 3x^{-2}$ since the power is a negative integer. ♢

Definition A.2.16 Polynomials. An algebraic expression that can be rewritten as a sum of monomials is called a **polynomial**. The monomials are called the **terms** of the polynomial. The monomial with the highest power is called the **leading term** and its power is called the **degree** of the polynomial. The constant multiples in the monomials are called **coefficients** and the coefficient in the leading term is called the **leading coefficient**.

We usually write a polynomial with terms ordered by decreasing powers, called **standard form**. An abstract representation of a polynomial with degree n is written

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

where the symbols $a_n, a_{n-1}, \ldots, a_0$ represent the coefficients. A missing term is represented by a coefficient zero. ♢

Example A.2.17 $x^4 - 2x^2 + 3x + 1$ is a polynomial with degree $n = 4$. The coefficients are $a_4 = 1$, $a_3 = 0$ (since no x^3 term), $a_2 = -2$, $a_1 = 3$ and $a_0 = 1$.

$2x^2(3x + 1)(x - 2)$ is a polynomial, but must be expanded (multiply out)

to find the coefficients.

$$\begin{aligned} 2x^2(3x+1)(x-2) &= 2x^2(3x^2-6x+x-2) \\ &= 2x^2(3x^2-5x-2) \\ &= 6x^4-10x^3-4x^2 \end{aligned}$$

We see that the polynomial has degree $n=4$ and coefficients $a_4=6$, $a_3=-10$, $a_2=-4$ and $a_1=a_0=0$. □

Every polynomial $p(x)$ is a function whose domain is all real numbers $(-\infty,\infty)$. Values of x for which $p(x)=0$ are called zeros or roots of the polynomial. These roots are related to factors.

Theorem A.2.18 Root–Factor Theorem. *Suppose $p(x)$ is a polynomial of degree n for which $x=c$ is a root, $p(c)=0$. Then $p(x)$ can be written in a factored form*

$$p(x)=(x-c)\cdot q(x)$$

where $q(x)$ is a polynomial of degree $n-1$.

Synthetic division is an algorithm that can both test if a value $x=r$ is a root and determine the coefficients of the factored polynomial $q(x)$ at the same time. Synthetic division for quadratic polynomials (degree $n=2$) is a special case of this process, described in Algorithm A.2.12.

Algorithm A.2.19 Synthetic Division. *Given a polynomial $p(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+\cdots+a_1x+a_0$ and a test value $x=c$, synthetic division is an algorithm for finding coefficients $b_{n-1},\ldots,b_0$ of a polynomial $q(x)=b_{n-1}x^{n-1}+\cdots+b_1x+b_0$ and a remainder r such that*

$$p(x)=(x-c)q(x)+r.$$

The remainder is also the value of the original polynomial at $x=r$, $p(c)=r$, so that when $x=c$ is a root, $p(c)=0$ and $p(x)$ factors as

$$p(x)=(x-c)q(x).$$

The coefficients for $q(x)$ and the remainder are found using the steps below.

1. *We will create a table with three rows and $n+1$ columns. The first row consists of the coefficients of $p(x)$, using 0 for any skipped terms, ordered by decreasing power. The first value of the second row is always 0. The table will start as follows:*

a_n	a_{n-1}	$\cdots$	a_1	a_0
0	____	$\cdots$	____	____
____	____	$\cdots$	____	____

2. *A column will always be completed (finding the third row) by adding the values in the first and second rows.*

3. *Once a column is complete, the second row of the next column is found by multiplying the previous value in the third row by the test value c.*

4. *Repeat these two steps until the table is complete. The third row of the table gives the coefficients and remainder as follows:*

a_n	a_{n-1}	$\cdots$	a_1	a_0
0	$c\cdot b_{n-1}$	$\cdots$	$c\cdot b_1$	$c\cdot b_0$
b_{n-1}	b_{n-2}	$\cdots$	b_0	r

The interpretation is that

$$p(x) = (x - c)(b_{n-1}x^{n-1} + \cdots + b_1x + b_0) + r.$$

Example A.2.20 Use synthetic division with the polynomial $p(x) = x^3 - 6x + 2$ with the test value $x = 2$ and interpret the result.

Solution. Start by identifying the coefficients. Any missed terms have a coefficient of zero,

$$p(x) = x^3 + 0x^2 + -6x + 2.$$

We start the synthetic division table using the coefficients in the first row.

1	0	−6	2
0	____	____	____
____	____	____	____

We then finish filling the table. To find values in the second row, we use the previous result in the third row and multiply by the test value 2. To find the values in the third row, we add the values in the column. The first value in the second row is always 0. The completed table is shown below.

1	0	−6	2
0	2	4	−4
1	2	−2	−2

Once the table is complete, we interpret the values in the third row as coefficients and a remainder. The last value is the remainder, $r = -2$, and the other values are the coefficients of a polynomial whose degree is one smaller than the original, in this case $n - 1 = 2$. That is, the quotient polynomial is $q(x) = x^2 + 2x - 2$. The original polynomial can be written

$$p(x) = (x - 2)q(x) + r$$
$$x^3 - 6x + 2 = (x - 2)(x^2 + 2x - 2) + -2$$

The non-zero remainder means that $x - 2$ is not a factor and also tells us that $p(2) = -2$. □

How do we know which numbers to try? If you have access to a graph of the polynomial, you should use the values for roots that you see. If you do not have access to a graph, then you might be able to use the results of the Integer Root Theorem or Rational Root Theorem so long as all of the coefficients of your polynomial are integers.

Theorem A.2.21 Integer Root Theorem. *If the coefficients of a polynomial*

$$p(x) = a_nx^n + \cdots + a_1x + a_0$$

has all integer coefficients, then the only possible integer roots are factors of the constant coefficient a_0.

Example A.2.22 The polynomial $p(x) = x^3 - 6x + 2$ has only integer coefficients and a constant coefficient $a_0 = 2$. The only factors of a_0 are ± 1 and ± 2. So the Integer Root Theorem guarantees that these four integers are the only four numbers that we need to check if they are roots. A quick test of each of those values (below) shows that $p(x)$ has no integer roots. (Without the theorem, we wouldn't know how many points we had to check.)

x	$p(x)$
1	-3
-1	-5
2	-2
-2	6

□

Theorem A.2.23 Rational Root Theorem. *If the coefficients of a polynomial*

$$p(x) = a_n x^n + \cdots + a_1 x + a_0$$

has all integer coefficients, then a rational number $x = r/s$ (where r and s are integers) might be a root only if r is a factor of a_0 and s is a factor of a_n.

The Integer Root Theorem is a special case of the Rational Root Theorem where $s = 1$ (which is always a factor of a_n.

A.2.4 Absolute Value

The absolute value operation takes a number and finds its magnitude (or distance from zero). Because magnitude is a non-negative value and positive and negative pairs are the same distance from zero, we often imagine that the role of absolute value is to remove a negative sign, $|-3| = 3$. However, when a variable is involved, a negative sign means finding the inverse of a value for which we may not know if it is positive or negative. So it is incorrect to say that $|-x| = x$ (FALSE). The proper definition of absolute value uses a piecewise formula.

Definition A.2.24 Absolute Value.

$$|x| = \begin{cases} x, & x \geq 0 \\ -x, & x < 0 \end{cases}$$

◊

As a function, the graph of absolute value $y = |x|$ gives two lines: $y = x$ when $x \geq 0$ and $y = -x$ when $x < 0$.

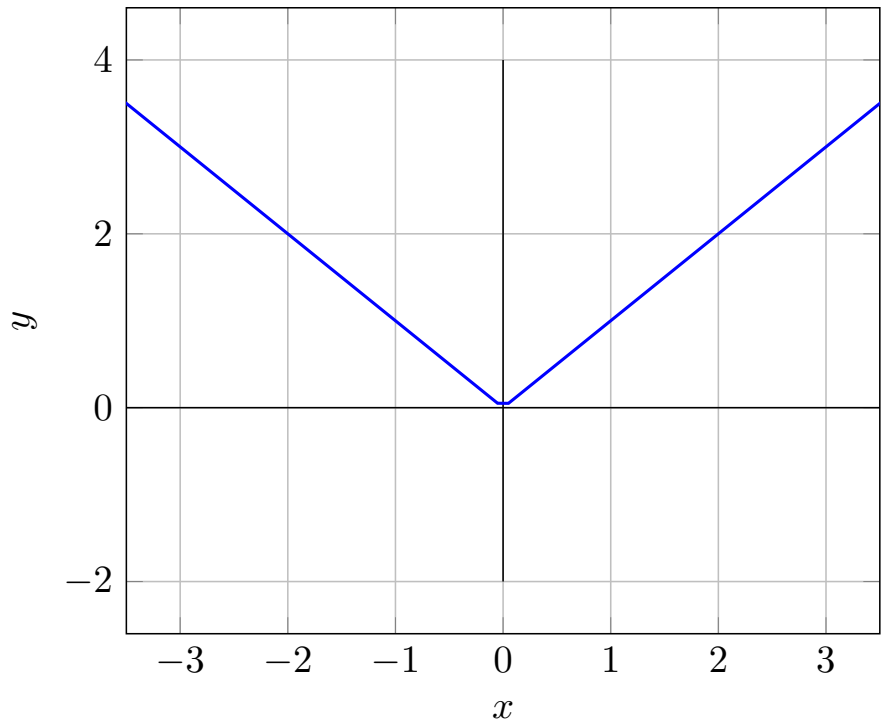


Figure A.2.25 Graph of the absolute value, $y = |x|$

It is sometimes useful to take advantage of an identity between the square root of a square and the absolute value. This is the source of the plus/minus when solving an equation with a square.

Theorem A.2.26

$$\sqrt{x^2} = |x|$$

Example A.2.27 Solve the equation $x^2 = 16$.

Solution. Applying a square root to both sides of the equation, we then get to use the absolute value identity.

$$\sqrt{x^2} = \sqrt{16} \quad \Leftrightarrow \quad |x| = \sqrt{16} = 4$$

The source of plus/minus is that there are two numbers with magnitude 4,

$$x = \pm 4.$$

□

The absolute value splits nicely with multiplication (and division). However, addition of two values with opposite signs shows that absolute values do not add: $|3 + -4| = |-1| \neq |3| + |-4| = 7$. Instead, we have an inequality called the triangle inequality.

Theorem A.2.28 Properties of Absolute Values.

$$|a \cdot b| = |a| \cdot |b|$$
$$\left|\frac{a}{b}\right| = \frac{|a|}{|b|}$$

The triangle inequality is used to show that the absolute value of a sum (or difference) is bounded by the sum of the magnitudes of the individual terms.

Theorem A.2.29 Triangle Inequality.

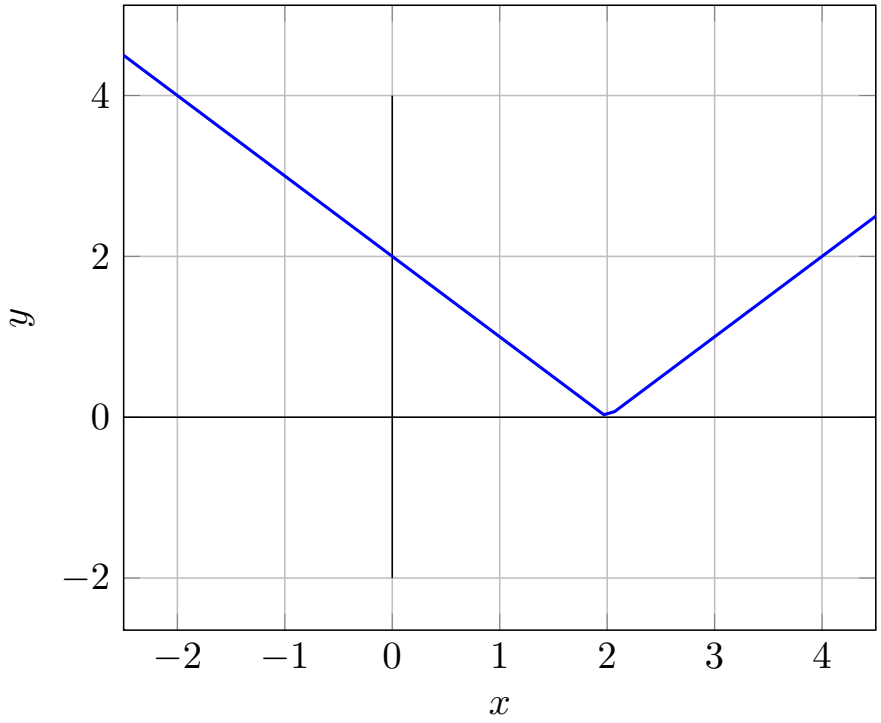
$$|a+b| \leq |a| + |b| \tag{A.2.11}$$
$$|a-b| \leq |a| + |b| \tag{A.2.12}$$

Occasionally we need to apply the triangle in reverse, showing that the absolute value of a sum (or difference) must be bigger than the difference in magnitudes of the parts.

Theorem A.2.30 Reverse Triangle Inequality.

$$|a+b| \geq \big||a| - |b|\big| \tag{A.2.13}$$
$$|a-b| \geq \big||a| - |b|\big| \tag{A.2.14}$$

Absolute value and subtraction is often used to describe the distance between two values. For example, the graph $y = |x-2|$ represents a shift of the graph $y = |x|$ two units to the right, so that instead of measuring the distance of x from 0 it measure the distance of x from 2.



Theorem A.2.31 *The expression $|x-a|$ measures the distance between x and the value a.*

Note that $|x+3| = |x--3|$ so that it represents the distance between x and the value -3.

Appendix B

Trigonometry Basics

B.1 Right Triangles and Trigonometry

B.1.1 Right Triangles

We already reviewed the idea that similar triangles have equal ratios of corresponding sides. This is at the heart of trigonometry using right triangles. An important fact from geometry is that any two triangles that have equal angles are similar. A right triangle, by definition, has one angle that is perpendicular or 90 degrees. Because the sum of angles in a triangle always add to 180 degrees, it really only takes one of the other angles to establish similarity.

We consider an acute angle θ (less than 90 degrees). A right triangle with base angle θ in standard position (with the base horizontal and the angle on the left) is shown in the diagram below. The legs that join at right angles are identified as being adjacent (adj) to the angle or opposite (opp) to the angle, while the side opposite the right angle is the hypotenuse (hyp).

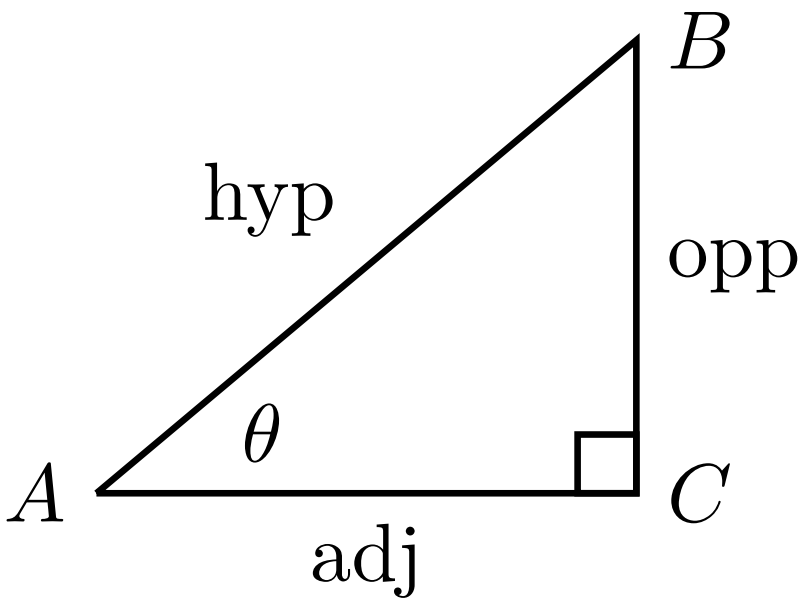


Figure B.1.1 An illustration of similar right triangles with base angle θ in standard position.

For any right triangle with the same base angle, the ratios of these three sides are always going to found in the exact proportions. These proportions define the six trigonometric values of the angle:

$$\sin\theta = \frac{\text{opp}}{\text{hyp}} \qquad \text{(sine)}$$

$$\cos\theta = \frac{\text{adj}}{\text{hyp}} \qquad \text{(cosine)}$$

$$\tan\theta = \frac{\text{opp}}{\text{adj}} \qquad \text{(tangent)}$$

$$\sec\theta = \frac{\text{hyp}}{\text{adj}} \qquad \text{(secant)}$$

$$\cot\theta = \frac{\text{adj}}{\text{opp}} \qquad \text{(cotangent)}$$

$$\csc\theta = \frac{\text{hyp}}{\text{opp}} \qquad \text{(cosecant)}.$$

The first three proportions are often introduced using a mnemonic ``SOH-CAH-TOA" where the three letters correspond to the first letter of the proportion name (*S*ine), the numerator side (*O*pposite) and denominator side (*H*ypotenuse).

It is sometimes useful to think of drawing special right triangles where one of the sides has unit length (length=1). When the hypotenuse has unit length, the triangle involves sine (opposite) and cosine (adjacent). When the adjacent leg has unit length, the triangle has sides with lengths given by tangent (opposite) and secant (hypotenuse). When the opposite leg has unit length, the triangle has sides with lengths given by cotangent (adjacent) and cosecant (hypotenuse). When I realized this simple fact, the naming of the proportions made much more sense.

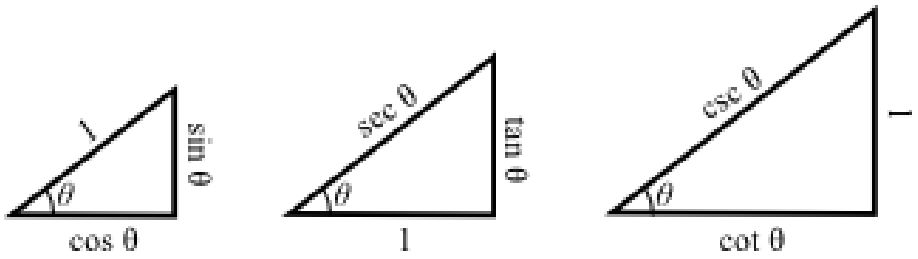


Figure B.1.2 Illustration of the three similar unit right triangles with angle θ.

Theorem B.1.3 Pythagorean Identities. *The Pythagorean theorem states that the sum of the squares of the legs of a right triangle must equal the square of the hypotenuse. Consequently, the trigonometric values of an angle must satisfy:*

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

Proof. Consider any triangle with base angle θ. The Pythagorean theorem guarantees

$$\text{opp}^2 + \text{adj}^2 = \text{hyp}^2.$$

The first identity is found by dividing both sides of this equation by hyp^2. The next two identities are found by dividing by adj^2 and opp^2, respectively. ■

B.1.2 Special Right Triangles

There are two right triangles that often appear in problems because they are geometrically simple. One of these is the isosceles right triangle where the base angle is exactly half of a right angle ($\theta = 45°$). This triangle is often called a 45–45–90 right triangle. The other simple triangle comes from dividing an equilateral triangle in half, so that the base angle is either 30° or 60°, and is called a 30–60–90 right triangle.

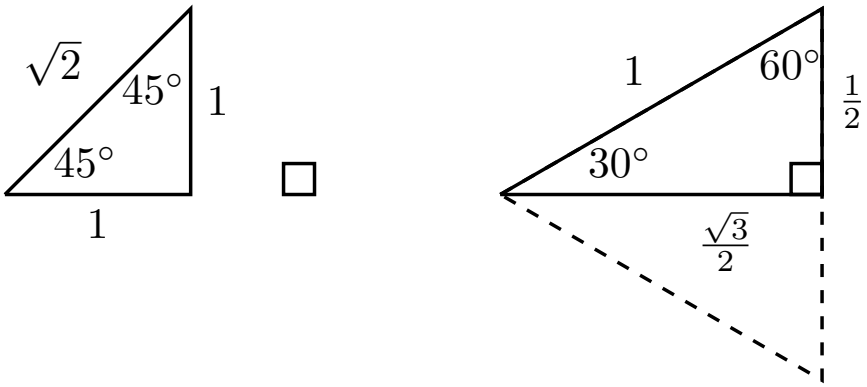


Figure B.1.4 Illustration of 45–45–90 and 30–60–90 right triangles.

It is handy to be able to reproduce these two triangles as quickly as possible. The key is to remember how the triangles were created. For the 45–45–90 triangle, draw an isosceles right triangle and label the lengths of both legs as 1. To find the length of the hypotenuse h, use the Pythagorean theorem:

$$\begin{aligned} 1^2 + 1^2 &= h^2 \\ h^2 &= 2 \\ h &= \sqrt{2} \end{aligned}$$

For the 30–60–90 triangle, recall that this is exactly half of an equilateral triangle. The hypotenuse will have length 1 while the leg opposite the 30 degrees is exactly half that length. Now use the Pythagorean theorem to find the length of the other leg b which bisected the triangle:

$$\begin{aligned} (\frac{1}{2})^2 + b^2 &= 1^2 \\ \frac{1}{4} + b^2 &= 1 \\ b^2 &= \frac{3}{4} \\ b &= \frac{\sqrt{3}}{2} \end{aligned}$$

Once you have the triangles drawn with lengths identified, you can use the triangles to find the proportions that define the trigonometric values. I strongly discourage trying to memorize this table. Learn how the table was created, and that will reinforce the more general principles and ultimately require less mental effort to recall.

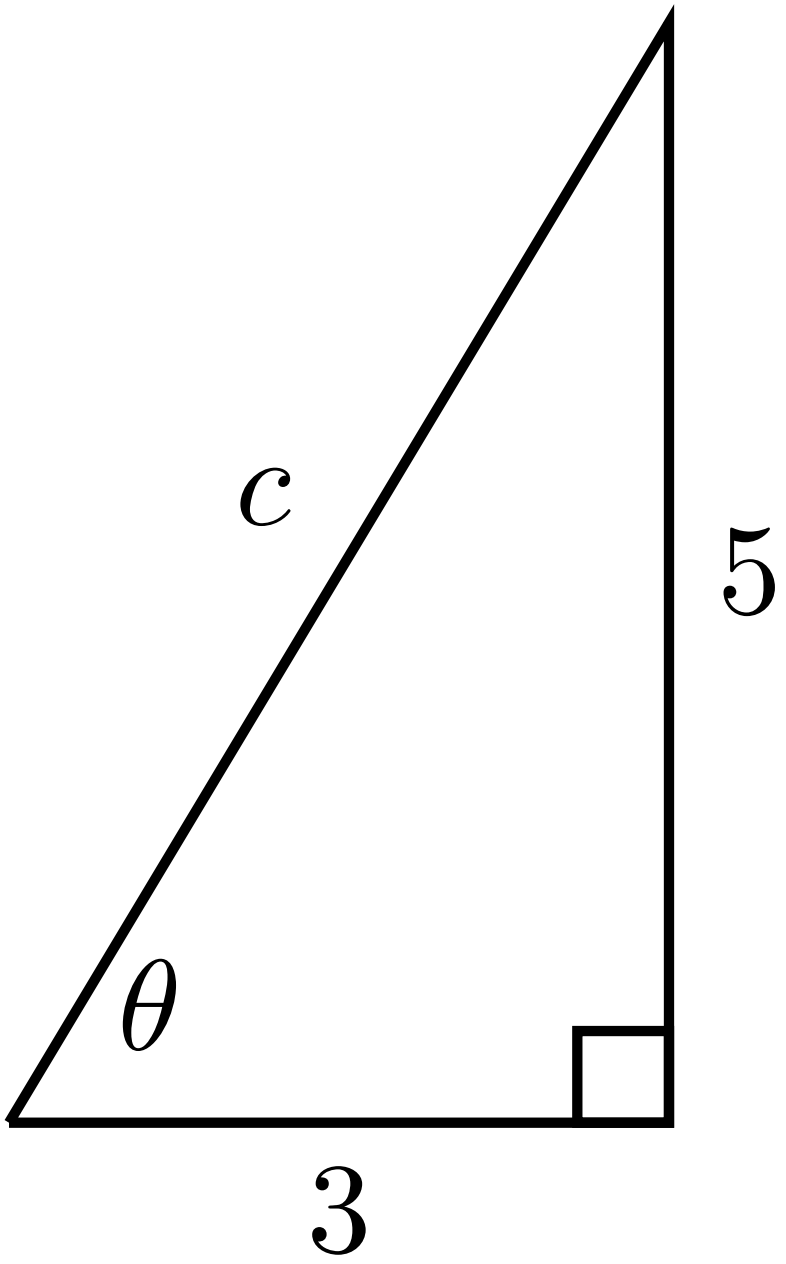
θ	30°	45°	60°
$\sin\theta = \frac{\text{opp}}{\text{hyp}}$	$\sin 30^\circ = \frac{1}{2}$	$\sin 45^\circ = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}$	$\sin 60^\circ = \frac{\sqrt{3}}{2}$
$\cos\theta = \frac{\text{adj}}{\text{hyp}}$	$\cos 30^\circ = \frac{\sqrt{3}}{2}$	$\cos 45^\circ = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}$	$\cos 60^\circ = \frac{1}{2}$
$\tan\theta = \frac{\text{opp}}{\text{adj}}$	$\tan 30^\circ = \frac{\frac{1}{2}}{\frac{\sqrt{3}}{2}} = \frac{1}{\sqrt{3}}$	$\tan 45^\circ = 1$	$\tan 60^\circ = \frac{\frac{\sqrt{3}}{2}}{\frac{1}{2}} = \sqrt{3}$

B.1.3 Examples

The first example illustates how you can use two known lengths for a triangle to find the trigonometric values for the angle.

Example B.1.5 Suppose a right triangle with a base angle θ has a base of length 3 and a height of length 5. What are the trigonometric values associated with θ?

Solution. Start by drawing a diagram, labeling the unknown side (the hypotenuse) with a variable, say c.



Using the Pythagorean theorem, we can find the length of the hypotenuse.

$$3^2 + 5^2 = c^2$$
$$9 + 25 = 34 = c^2$$
$$c = \sqrt{34}$$

Now that we know the lengths of all three sides, we can compute the trigono-

metric values for our angle.

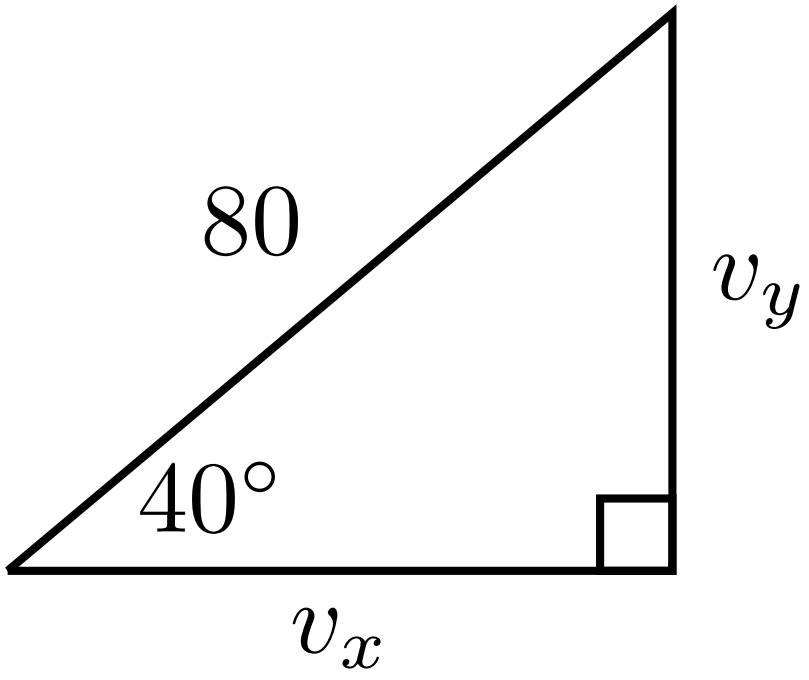
$$\sin\theta = \frac{\text{opp}}{\text{hyp}} = \frac{5}{\sqrt{34}}$$
$$\cos\theta = \frac{\text{adj}}{\text{hyp}} = \frac{3}{\sqrt{34}}$$
$$\tan\theta = \frac{\text{opp}}{\text{adj}} = \frac{5}{3}$$
$$\sec\theta = \frac{\text{hyp}}{\text{adj}} = \frac{\sqrt{34}}{3}$$
$$\cot\theta = \frac{\text{adj}}{\text{opp}} = \frac{3}{5}$$
$$\csc\theta = \frac{\text{hyp}}{\text{opp}} = \frac{\sqrt{34}}{5}$$

□

Our second example illustrates how knowing the trigonometric values associated with an angle allow us to determine the lengths of sides. This occurs a lot in physics in the context of decomposing a force or velocity into two perpendicular components.

Example B.1.6 A golf ball is launched at a speed of 80 $\frac{\text{m}}{\text{s}}$ and at an angle of \(40^\circ\) from the ground. What are the horizontal and vertical components of the ball's velocity?

Solution. My preferred method of solution begins with a figure of a right triangle, where the angle $\theta = 40°$ is the base angle, the hypotenuse is labeled with length 80, and the legs are labeled with variables representing the horizontal velocity v_x and the vertical velocity v_y.



The trigonometric values for $\theta = 40°$ can be easily found using a calculator. (These once were found by looking up the angles in a table.) We can setup equations relating the ratios of lengths and the corresponding trigonometric

values.

$$\sin 40^\circ = \frac{\text{opp}}{\text{hyp}} = \frac{v_y}{80}$$
$$\cos 40^\circ = \frac{\text{adj}}{\text{hyp}} = \frac{v_x}{80}$$

From the first equation, we can solve for v_y:

$$v_y = 80 \cdot \sin 40^\circ \approx 80 \cdot 0.6248 = 51.42.$$

We find the horizontal velocity v_x by solving the second equation

$$v_x = 80 \cdot \cos 40^\circ \approx 80 \cdot 0.7660 = 61.28.$$

So the ball is traveling with a vertical velocity approximately $51.42 \frac{\text{m}}{\text{s}}$ and a horizontal velocity approximately $61.28 \frac{\text{m}}{\text{s}}$. □

B.2 Measuring Arbitrary Angles

B.2.1 Angles and Rotation

In the previous section, we focused on the trigonometry of right triangles, which involve angles smaller than a right angle. The more significant use of measuring angles involve angles of rotation which can be significantly greater than 90 degrees. Before we proceed, we need to discuss how angles are measured.

Have you ever considered why we measure angles in degrees? What is so special about a degree? One thing nice about the number 360 is that it has so many factors, including 2, 3, 4, 5, 6, 8, 9, 10 and 12. (That only misses 7 and 11!) This makes it easy to divide into fractions that simplify cleanly with integers. Having an easily divisible number that is so close to the number of days in the year, so that 1 degree is close to the change in the daily position of stars in the sky, would have been convenient to ancient astronomers.

What other ways are there to measure rotations?

If a right angle is considered the fundamental unit, then we might consider measuring an angle as a percentage of a right angle, so that 90 degrees counts as 100 percent. This idea led to the development of what is called a gradian. So an angle of 1 gradian is exactly 1 percent of a right angle, and there are 400 gradians in a circle.

Alternatively, if we think of a full rotation as the fundamental unit, then we might measure an angle as a fraction of a rotation. I especially like this perspective when dealing with trigonometry on the unit circle. We will let τ (the Greek letter *tau*) be the unit of a complete rotation ($1\tau = 360°$). Since 90 degrees is one quarter rotation and 60 degrees is one sixth rotation, we have $90° = \frac{1}{4}\tau$ and $60° = \frac{1}{6}\tau$. Other conversions can be determined using standard techniques.

All of the preceding methods for measuring angles are based on natural activity of counting. And none of these methods are mathematically superior to one another. Curiously, the mathematically best way to measure an angle is not based on counting integer divisions of a rotation but is based on a unit called the radian which involves measuring arc length in terms of the radius.

B.2.2 Arc Length and Radian Measure of Angles

An arc is a path traced along the circumference of a circle. We can describe an arc by either measuring the length of the path (arc length) or by measuring the angle subtended by the arc. It should be obvious that the arc length is proportional to the angle.

Let s represent the arc length and let θ represent the angle of the arc. To say that $s \propto \theta$ is to say that there is a proportionality constant so that $s = k\theta$. Because s is a measure of length, by geometric similarity, the arc length must also be proportional to the radius of the circle r (or $k \propto r$). This means that there is a constant α so that

$$s = \alpha\theta r.$$

The value of the constant α depends on how we measure angles. To see this, we will use an arc of a complete rotation which has an arc length equal to the circumference of the circle $s = 2\pi r$. If the angle is measured in degrees, $\theta = 360$, then the proportionality constant will be α_{deg} defined by

$$2\pi r = \alpha_{\text{deg}}(360)r \quad \Leftrightarrow \quad \alpha_{\text{deg}} = \frac{2\pi}{360}.$$

If the angle is measured in gradians, $\theta = 400$, then the proportionality constant α_{grad} satisfies

$$2\pi r = \alpha_{\text{grad}}(400)r \quad \Leftrightarrow \quad \alpha_{\text{grad}} = \frac{2\pi}{400}.$$

In a similar way, if the angle is measured in rotations τ, then

$$\alpha_\tau = \frac{2\pi}{1} = 2\pi.$$

The mathematically defined measure of angle called the radian is determined by choosing the proportionality constant α as being convenient instead of choosing the measurement of the angle as being convenient. If we choose $\alpha = 1$, then this requires that we measure a full rotation θ to satisfy

$$2\pi r = 1\theta r \quad \Leftrightarrow \quad \theta = 2\pi.$$

That is, a full rotation is defined as 2π radians.

With radians as the measure of angle, the arc length formula simplifies to

$$s = \theta r.$$

In other words, the angle θ is determined by measuring the length of a subtended arc using the radius as the unit of length. An angle $\theta = 1$ radian has an arc length equal to the radius of the arc. When angles are measured in radians, no units are used; so we would just say $\theta = 1$. We also have, for example,

$$\frac{1}{4}\tau = \frac{\pi}{2}, \tag{B.2.1}$$

$$\frac{1}{2}\tau = \pi, \tag{B.2.2}$$

$$1\tau = 2\pi. \tag{B.2.3}$$

How did you read that last sentence? Did you visualize and interpret what it says, and not just read the symbols? What does each equation mean? Below is a figure illustrating the examples $\theta = 1$, $\theta = \frac{1}{4}\tau$ and $\theta = \frac{1}{2}\tau$. Could you draw a similar figure showing angles corresponding to 30, 45 and 60 degrees? What is the radian measure of those angles?

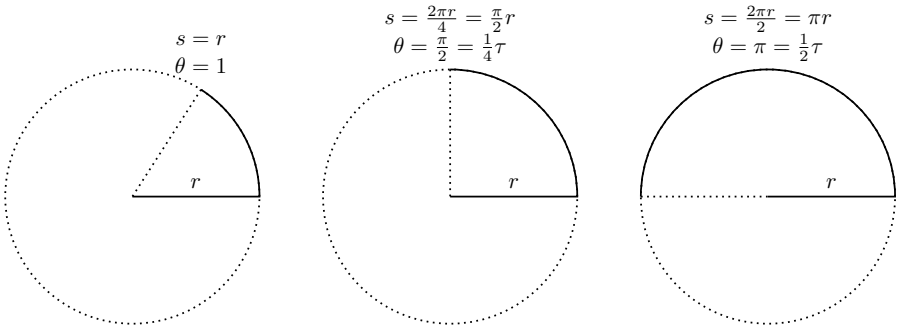


Figure B.2.1 Important arcs measured in radians.

B.2.3 Unit Circle and Standard Position

In order to have a universal reference for an angle, we introduce the standard position of an angle in terms of a unit circle. The equation for a circle with

radius $r = 1$ and a center at the origin $(0, 0)$ is $x^2 + y^2 = 1$. Standard position for an angle always forms an arc on the unit circle that starts on the x-axis at $(1, 0)$ and moves along the circle in the counter clockwise direction for positive angles (clockwise for negative angles).

An angle greater than 2π will wrap completely around the circle and continue. Multiples of 2π $(0, \pm 2\pi, \pm 4\pi, \ldots)$ all end at the same point $(1, 0)$ on the unit circle but correspond to a different number of rotations. In fact, every point on the unit circle has infinitely many different angles (that differ by multiples of 2π) that end at that point when in standard position.

Example B.2.2 Find all angles that terminate at the point $(0, 1)$.

Solution. The point $(0, 1)$ is a quarter turn in the positive direction or three quarters turn in the negative direction. So the angles $\theta = \frac{1}{4}\tau = \frac{1}{4}(2\pi) = \frac{\pi}{2}$ and $\theta = -\frac{3}{4}\tau = -\frac{3}{4}(2\pi) = -\frac{3\pi}{2}$ are two of the angles desired.

Notice that these angles are 2π apart. In fact, we can add any integer multiple of 2π and end at the same spot. One of writing this is to say

$$\theta = \frac{\pi}{2} + 2\pi k, \quad k = 0, \pm 1, \pm 2, \ldots.$$

□

Example B.2.3 Where does the angle $\theta = \frac{28\pi}{3}$ terminate?

Solution. This is the same as dealing with improper fractions or division with remainders. First, recognize that $2\pi = \frac{6\pi}{3}$. Second, determine how many times 6 goes into 28 using division, to find 4 with a remainder of 4 $(28 = 4 \cdot 6 + 4)$. Rewrite the fraction:

$$\theta = \frac{28\pi}{3} = 4 \cdot 2\pi + \frac{4\pi}{3}.$$

To interpret this angle, recognize that the integer multiple of 2π corresponds to 4 complete rotations. Next, determine what fraction of a rotation corresponds to $\frac{4\pi}{3}$:

$$\frac{4\pi}{3} = \frac{2}{3}(2\pi).$$

So we continue another two-thirds of a rotation, which is 60 degrees past a half rotation and 30 degrees short of vertical.

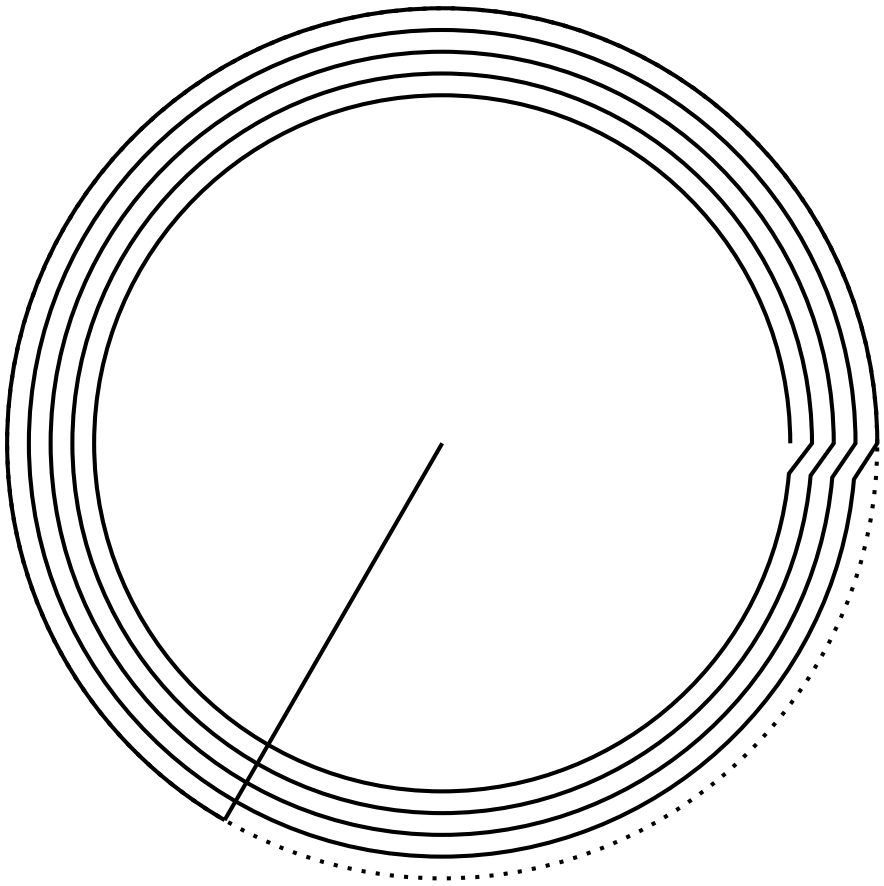


Figure B.2.4 Terminal position of $\theta = \frac{28\pi}{3}$.

□

B.3 Unit Circle Trigonometry

B.3.1 Unit Circle Trigonometry

When we first introduced the trigonometric functions of an angle, we did it for acute angles. What will we do for larger angles, or negative angles? We will use the unit circle standard position of an angle and choose a method that agrees with what we would expect for acute angles.

So consider a positive, acute angle $0 < \theta < \frac{\pi}{2}$ (notice that we continue to use radians) that has been placed in standard position on the unit circle. We have also drawn the corresponding right triangle in standard position. Because the unit circle has radius $r = 1$, the hypotenuse of our triangle has length 1, so that the legs are the cosine (adjacent) and the sine (opposite) of the angle.

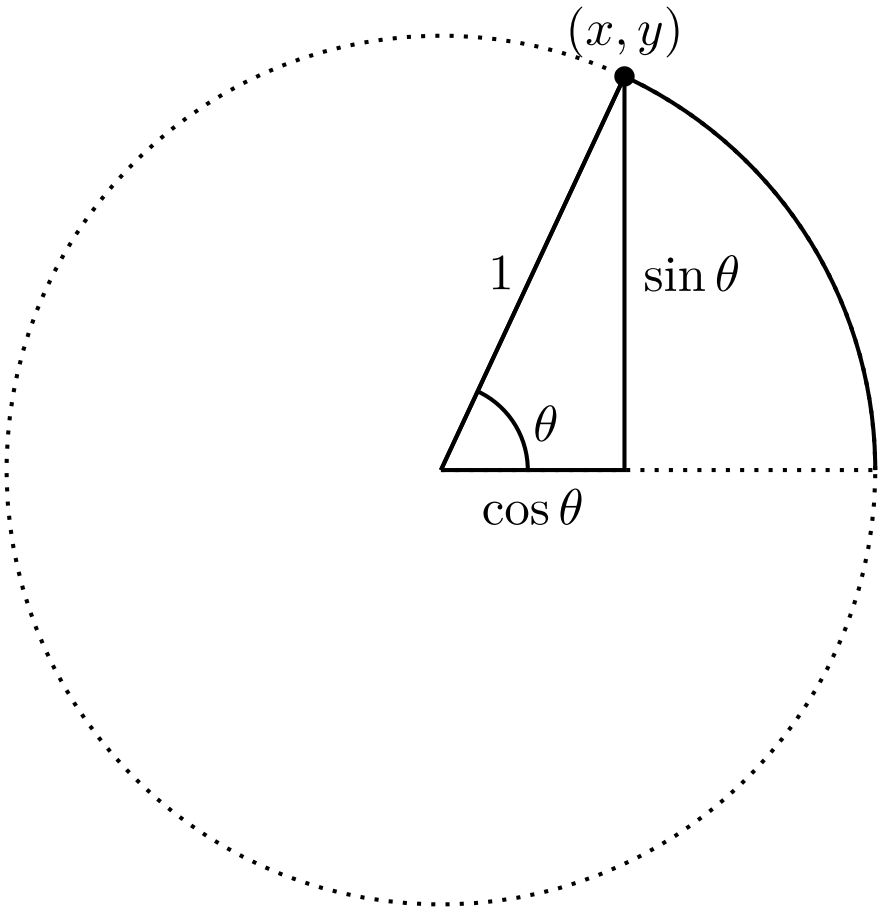


Figure B.3.1 An acute angle on the unit circle and with an associated right triangle in standard position.

By noticing that the length of the legs also corresponds to the x- and y-coordinates, this gives us our desired generalization. For any angle θ, we will define the cosine and sine of the angle as the x- and y-coordinates of the point on the unit circle for the terminal edge of the arc in standard position.

Definition B.3.2 If an angle θ is put in standard position and terminates at a point (x, y) on the unit circle, then the trigonometric functions of θ are defined

by

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

$\Diamond$

I suggest remembering the definitions of cosine and sine in terms of the x- and y-coordinates, and the other functions in terms of the ratios involving sine and cosine. One exception to this general principle is that it can sometimes be helpful to think of the tangent, which is the ratio y/x, as the slope of the terminal edge.

B.3.2 Periodic Behavior of Trigonometric Functions

Because the terminal edge of an arc repeatedly passes through the same points, the unit circle definitions of trigonometric functions create periodic functions. Most of the functions have a period of 2π corresponding to the rotation necessary to return to the same point. However, the tangent and cotangent functions each have a period of π. This is a consequence of the definition involving both x and y in their definitions. The ratio will be the same if both values change sign, which is precisely what happens for a half-rotation in the angle.

How can you remember the graphs of the functions? You can do this while also reinforcing your understanding of the unit-circle definitions by imagining rotating around the unit circle in a counter-clockwise direction. Draw the cosine (x-coordinate), sine (y-coordinate) and tangent (slope) functions as you go.

- Start at angle $\theta = 0$ with a terminal point $(1, 0)$. Using the coordinates and slope of the terminal edge gives:

$$\cos 0 = 1 \qquad \sin 0 = 0 \qquad \tan 0 = 0$$

- Go half a right angle to $\theta = \frac{\pi}{4}$ with a terminal point where $x = y$ at $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Using the coordinates and slope of the terminal edge gives:

$$\cos\frac{\pi}{4} = \frac{1}{\sqrt{2}} \qquad \sin\frac{\pi}{4} = \frac{1}{\sqrt{2}} \qquad \tan\frac{\pi}{4} = 1$$

- Finish the right angle at $\theta = \frac{\pi}{2}$ with a terminal point at $(0, 1)$. Using the coordinates and slope of the terminal edge gives:

$$\cos\frac{\pi}{2} = 0 \qquad \sin\frac{\pi}{2} = 1 \qquad \tan\frac{\pi}{4} = \text{undef.}$$

In general, you should remember that cosine and sine oscillate between peak values of -1 and 1. Paying attention to the unit circle, you will be able to identify the actual points where these are reached. The tangent, which

measures a slope, goes through all possible values with negative values when the angle is in the second or fourth quadrants and positive when the angle is in the first or third quadrants.

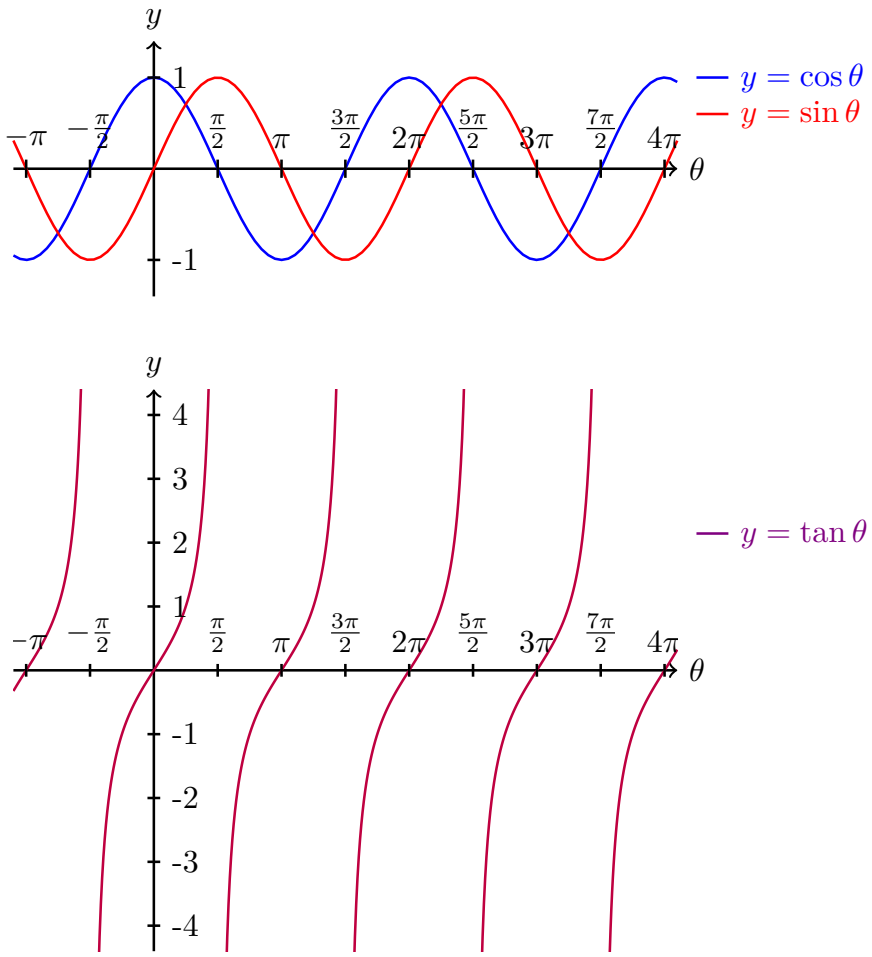


Figure B.3.3 Periodic graphs of the cosine, sine and tangent functions.

When thinking about the tangent function, if you will remember that it involves division by the cosine, then the breaks (vertical asymptotes) occur at every point where the cosine is zero. This exactly corresponds to where the terminal edge of the angle on the unit circle is vertical.

B.4 Inverse Trigonometric Functions

Trigonometric functions are periodic functions defined in terms of the unit circle. A function must be one-to-one in order to have an inverse function. Periodic functions obviously fail the horizontal line test and are not one-to-one. However, we can restrict the function to a domain on which it is one-to-one. For functions defined in terms of the unit circle, we will restrict each domain to include the first quadrant, corresponding to angles from 0 to $\frac{\pi}{2}$, as well as adjacent angles that will guarantee the restricted function is one-to-one and includes the full range of output values.

Recall that the sine function represents the y-coordinate of the point on the unit circle of an angle's terminal edge. The range consists of all numbers in the interval $[-1, 1]$. The first quadrant of angles 0 to $\frac{\pi}{2}$ leads to points on the unit circle with y-values from 0 to 1. Angles just beyond $\frac{\pi}{2}$ repeat the same y-values. We instead use angles in the fourth quadrant from $-\frac{\pi}{2}$ to 0. Altogether, the restricted domain will be $[-\frac{\pi}{2}, \frac{\pi}{2}]$.

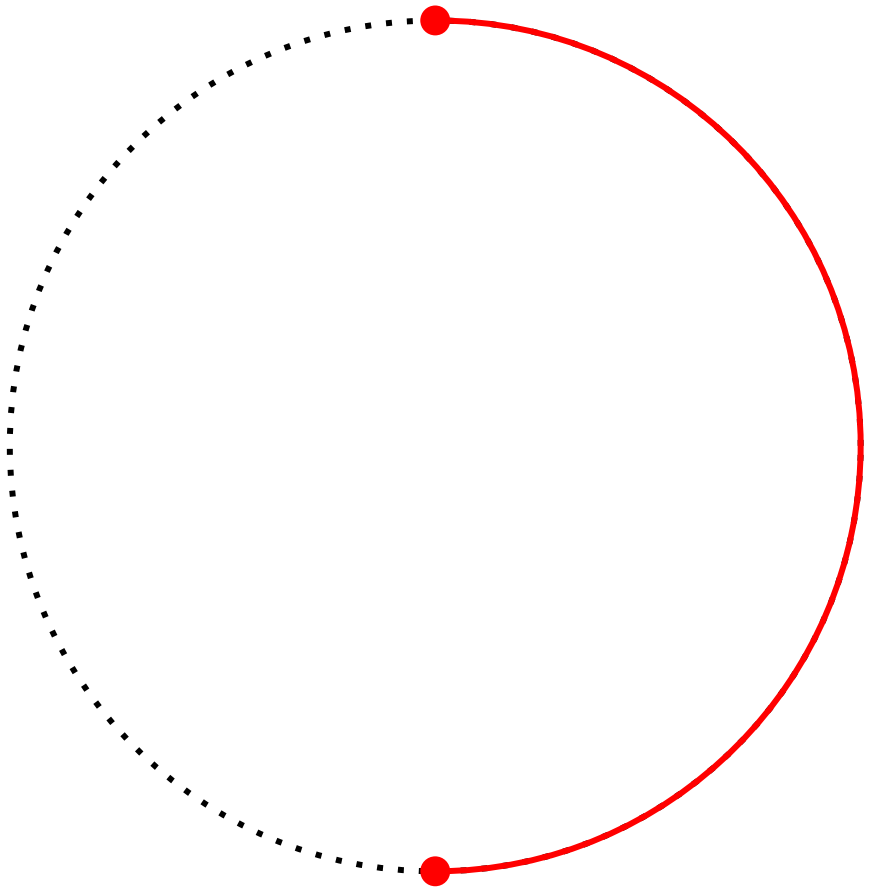


Figure B.4.1 The restricted domain for sine is the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$.

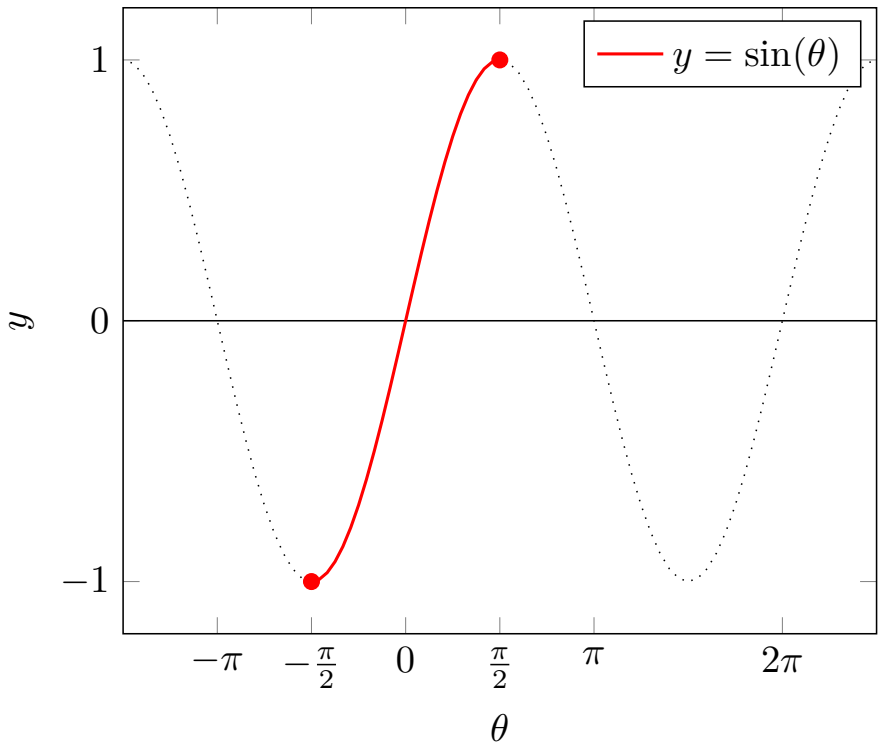


Figure B.4.2 The sine function restricted to the domain $[-\frac{\pi}{2}, \frac{\pi}{2}]$.

The inverse function for the restricted sine function is called the arcsine function or inverse sine function. Because $\sin$ takes an angle θ in radians as the input and gives the y-coordinate on the unit circle as the output, we have $\sin : \theta \mapsto y$. The inverse takes a y-coordinate on the unit circle as the input and gives an angle θ in the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$ as output, we have $\sin^{-1} : y \mapsto \theta$. The graph of the arcsine is shown below.

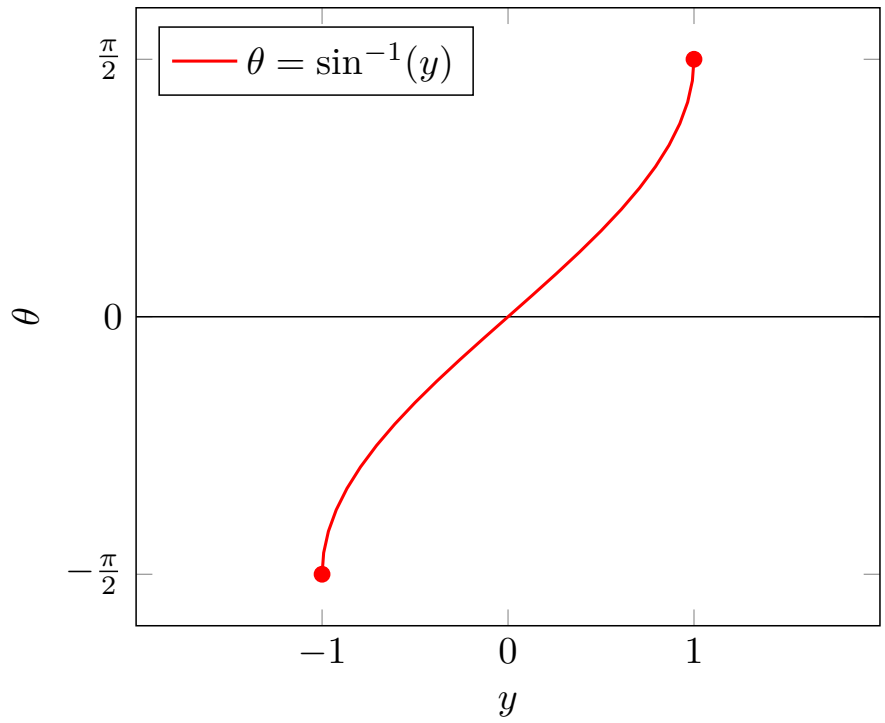


Figure B.4.3 The arcsine function, which is the inverse of the restricted sine.

The cosine function takes an angle θ as the input and returns the x-coordinate of the corresponding point on the unit circle. The first quadrant angles between $\theta = 0$ and $\theta = \frac{\pi}{2}$ have x-coordinates between 0 and 1. To obtain the x-coordinates between -1 and 0 come from angles in the second quadrant. The restricted domain will be the interval $[0, \pi]$.

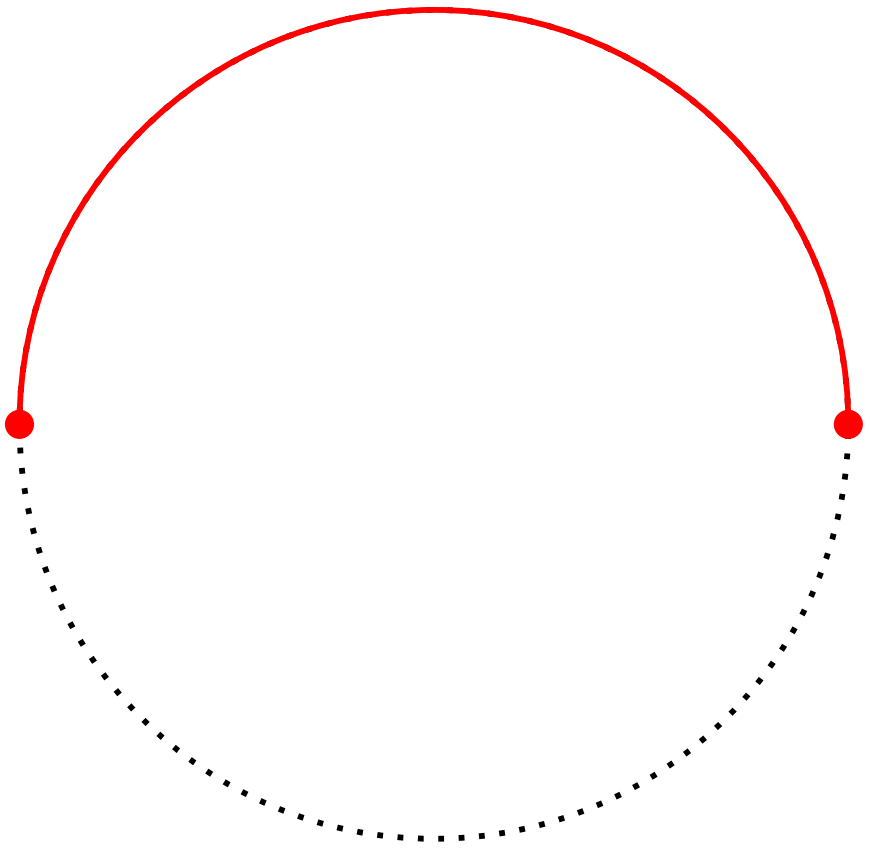


Figure B.4.4 The restricted domain for cosine is the interval $[0, \pi]$.

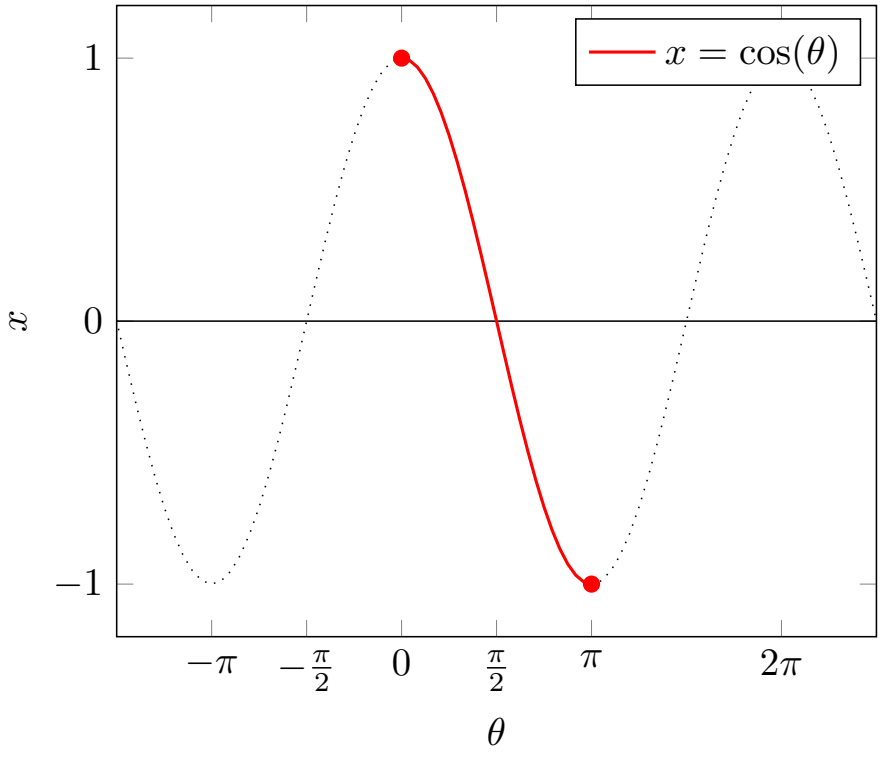


Figure B.4.5 The cosine function restricted to the domain $[0, \pi]$.

The inverse function for the restricted cosine function is called the arccosine function or inverse cosine function. Because $\cos$ takes an angle θ in radians as the input and gives the x-coordinate on the unit circle as the output, we have $\cos : \theta \mapsto x$. The inverse takes an x-coordinate on the unit circle as the input and gives an angle θ in the interval $[0, \pi]$ as output, so we have $\cos^{-1} : x \mapsto \theta$. The graph of the arccosine is shown below.

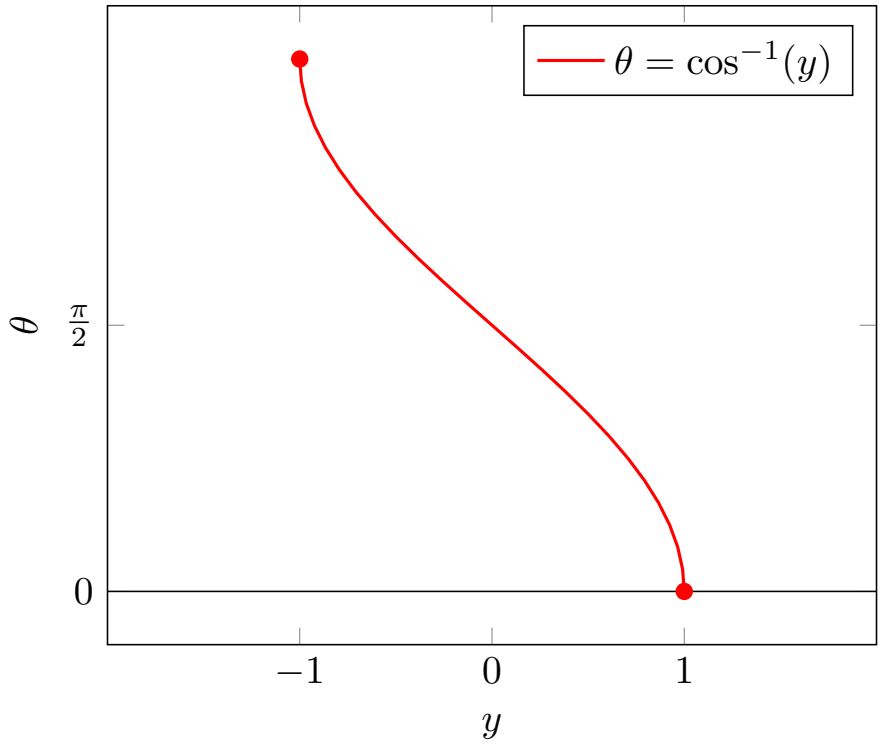


Figure B.4.6 The arccosine function, which is the inverse of the restricted cosine.

The tangent function takes an angle θ as the input and returns the ratio $\frac{y}{x}$ for the coordinates (x, y) of the corresponding point on the unit circle. This ratio is exactly the slope $m = \frac{y}{x}$ of the line joining $(0, 0)$ and (x, y). The first quadrant angles between $\theta = 0$ and $\theta = \frac{\pi}{2}$ correspond to all of the possible positive slopes. To obtain negative slopes, we could use either the second or fourth quadrant. So that the function will be continuous, the restricted domain is chosen as the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$. The end-points are not included because the tangent is not defined at those angles.

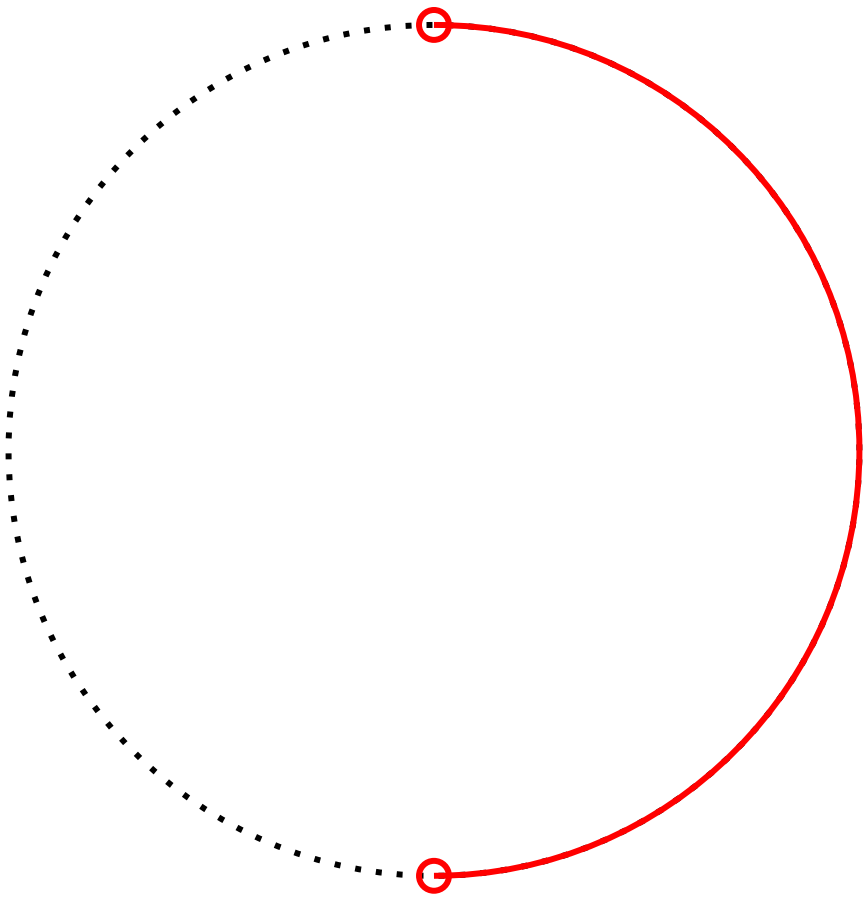


Figure B.4.7 The restricted domain for tangent is the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$.

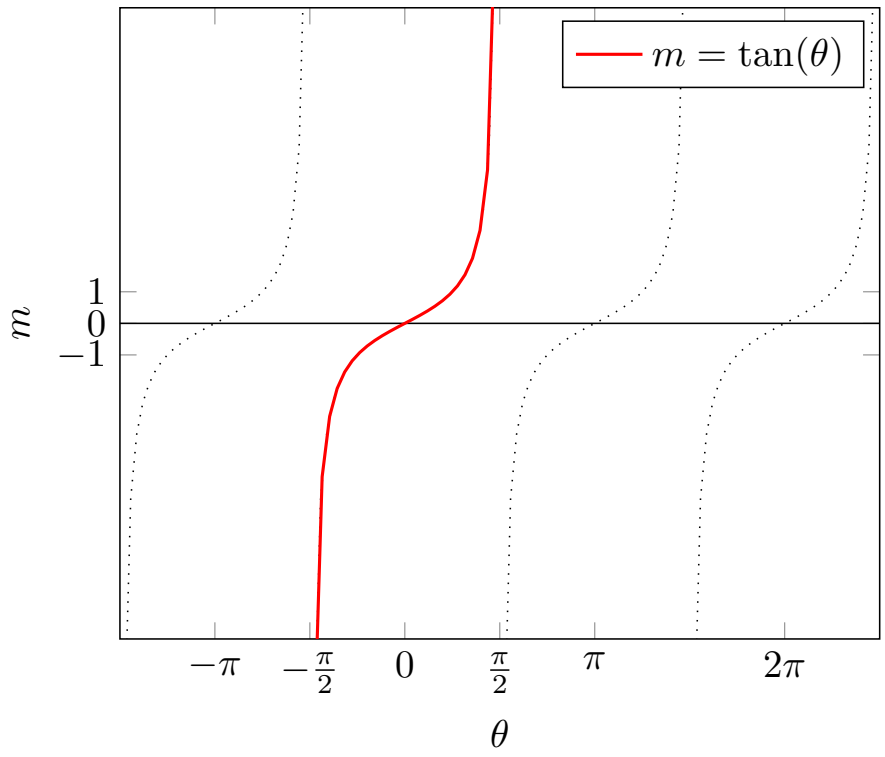


Figure B.4.8 The tangent function restricted to the domain $(-\frac{\pi}{2}, \frac{\pi}{2})$.

The inverse function for the restricted tangent function is called the arctangent function or inverse tangent function. Because $\tan$ takes an angle θ in radians as the input and gives the slope m of the angle, $\tan : \theta \mapsto m$. The inverse takes a slope m as the input and gives an angle θ in the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$ the has this slope, so we have $\tan^{-1} : m \mapsto \theta$. The graph of the arctangent is shown below.

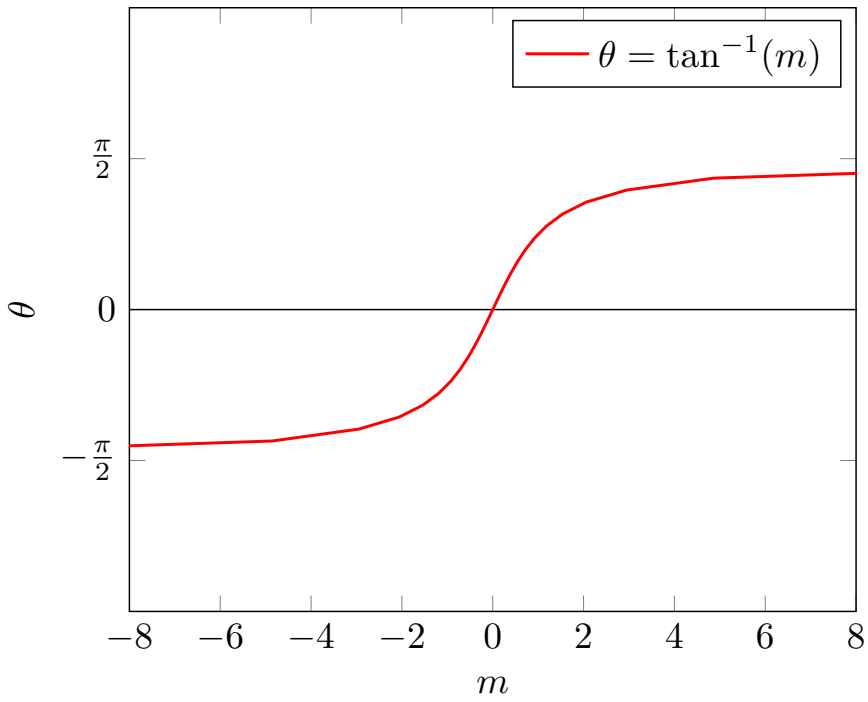


Figure B.4.9 The arctangent function, which is the inverse of the restricted tangent.

The secant, cosecant, and cotangent functions also have restricted domains and corresponding inverse functions. The table below summarizes the restricted domains and ranges for each of the trigonometric functions.

Restricted Function	Domain	Range
$\sin(x)$	$[-\frac{\pi}{2}, \frac{\pi}{2}]$	$[-1, 1]$
$\cos(x)$	$[0, \pi]$	$[-1, 1]$
$\tan(x)$	$(-\frac{\pi}{2}, \frac{\pi}{2})$	$(-\infty, \infty)$
$\cot(x)$	$(0, \pi)$	$(-\infty, \infty)$
$\sec(x)$	$[0, \frac{\pi}{2}) \cup (\frac{\pi}{2}, \pi]$	$(-\infty, -1] \cup [1, \infty)$
$\csc(x)$	$[-\frac{\pi}{2}, 0) \cup (0, \frac{\pi}{2}]$	$(-\infty, -1] \cup [1, \infty)$

The domain and range for the inverse functions are exactly the reverse of the restricted trigonometric functions. The inverse trigonometric functions have multiple representations. For example, the arcsine is sometimes written $\sin^{-1}$ but is also written either arcsin or asin. The table summarizes the information about the inverse trigonometric functions.

Inverse Functions	Domain	Range
$\sin^{-1}(x) = \arcsin(x)$	$[-1, 1]$	$[-\frac{\pi}{2}, \frac{\pi}{2}]$
$\cos^{-1}(x) = \arccos(x)$	$[-1, 1]$	$[0, \pi]$
$\tan^{-1}(x) = \arctan(x)$	$(-\infty, \infty)$	$(-\frac{\pi}{2}, \frac{\pi}{2})$
$\cot^{-1}(x) = \operatorname{arccot}(x)$	$(-\infty, \infty)$	$(0, \pi)$
$\sec^{-1}(x) = \operatorname{arcsec}(x)$	$(-\infty, -1] \cup [1, \infty)$	$[0, \frac{\pi}{2}) \cup (\frac{\pi}{2}, \pi]$
$\csc^{-1}(x) = \operatorname{arccsc}(x)$	$(-\infty, -1] \cup [1, \infty)$	$[-\frac{\pi}{2}, 0) \cup (0, \frac{\pi}{2}]$