Mathematics for Secondary School Teachers: Basics, Connections, and Extensions

Elizabeth G. Bremigan
Ralph J. Bremigan
John D. Lorch

E-mail address: egbremigan@bsu.edu
E-mail address: bremigan@bsu.edu
E-mail address: jlorch@bsu.edu

Department of Mathematical Sciences, Ball State University, Muncie, Indiana 47306–0490
## Contents

Chapter 1. Functions
1. The Notion of a Function 2
2. Graphs, Relations, and the Definition of a Function 9
3. An Application/Interlude: Graphing Transformations 14
4. Images and Inverse Images 22
5. Injective, Surjective, and Bijective Functions 28
6. Composite Functions, Identity Functions, and Inverse Functions 36
7. Subsets and Equations 46

Chapter 2. Lines in the Plane
1. Linear Equations 59
2. Intercepts, Slope, and Equations of Lines 61
3. Solving Systems of Linear Equations 68
4. Parameterized Lines 71
5. Parallel Lines, Perpendicular Lines, and Distance 81
6. An Application of Distance: Using Lines to Fit Data 91
7. The Collection of All Lines in the Plane 99
8. Exercises Involving Student Work 105

Chapter 3. Quadratic Polynomials
1. The Squaring Function and Parabolas 108
2. Completing the Square and Solving Quadratic Equations 115
3. Quadratic Polynomials: Graphs, Roots, and Factors 120
4. Application: Quadratic Splines 131
5. Application: Problems Involving Tangency 137
6. Exercises Involving Student Work 143

Chapter 4. Trigonometry
1. Angles 146
2. Triangles, Circles, and Trigonometric Functions 155
3. Laws of Sines and Cosines 167
4. Values of the Trigonometric Functions: Ptolemy’s Theorem, Identities, and Tables 174
5. Trigonometry, Coordinate Geometry, and Linear Algebra 182
6. Application: Why We Use Radians in Calculus 190
7. Application: Periodic Data and Trigonometric Polynomials 192
8. Exercises Involving Student Work 202

Chapter 5. Hyperbolic Trigonometry
1. Hyperbolas and Ellipses 206
2. Hyperbolic Angles and Their Measures 214
3. The Hyperbolic Trigonometric Functions 221
4. Hyperbolic Rotations 225
5. Differential Calculus and the Hyperbolic Trigonometric Functions 231
6. Hyperbolic Trigonometric Functions and the Exponential Function 234
7. Linking the Two Trigonometries 238

Chapter 6. Numbers 241
1. The Natural Numbers and Whole Numbers: An Intuitive View 243
2. The Integers: An Intuitive View 246
3. The Rational Numbers: An Intuitive View 248
4. The Real Numbers: An Intuitive View 257
5. The Natural Numbers and Whole Numbers: A More Rigorous View 262
6. The Integers: A More Rigorous View 270
7. The Rational Numbers: A More Rigorous View 272
8. The Real Numbers: A More Rigorous View 275
9. Exercises Involving Student Work: Fractions and the Number Line 283

Chapter 7. Operations in Number Systems 285
1. Addition and Subtraction in Number Systems 286
2. Multiplication in Number Systems 298
3. Division in Number Systems 307
4. Algebraic Properties in Number Systems: An Intuitive View 315
5. Algebraic Properties in Number Systems: A Survey of Some Rigorous Arguments 321
6. Complex Numbers 326

Chapter 8. Topics in Number Systems 343
1. Arithmetic in the Integers 344
2. Systems of Numeration for Whole Numbers 364
3. Divisibility Tests 375
4. Decimals 377
5. Algebraic and Transcendental Numbers 396
6. Exercises Involving Student Work: Fractions and Decimals 402

Chapter 9. Exponentiation 403
1. Whole Number Exponents 406
2. Integer Exponents 410
3. Rational Exponents 413
4. Real Exponents 420
5. The Real Logarithmic Functions 426
6. A Rigorous Approach to Real Exponential Functions 433
7. Exercises Involving Student Work 438

Chapter 10. Exponential and Logarithmic Functions: History, Computation, and Application 451
1. Logarithms and History: Logarithm Tables 452
2. Logarithms and History: The Natural Logarithm 459
3. Compelling Properties of the Natural Logarithm and Natural Exponential Function 467
4. Applications of Exponential Functions 472

Chapter 11. Transcendental Functions and Complex Numbers 477
1. The Functions \( f(z) = z^n \) 478
2. Roots of Complex Numbers 480
3. Rational Exponents: Roots and Powers of the Base 490
4. The Complex Exponential Function \( e^z \) 492
5. Complex Bases and Complex Exponents 500

Chapter 12. Beyond Quadratics: Higher Degree Polynomials 503
CONTENTS

1. Graphical Features of Cubic Polynomials 504
2. Connection: Roots and Coefficients 516
3. Solving Cubic Equations 526
4. Factoring and the Fundamental Theorem of Algebra 534
5. Application: Newton’s Method and Polynomials 545

Chapter 13. Measurement 553
2. Measuring Geometric Objects 564
3. Exploring Areas of Triangles 578
4. Measuring Figures with Non-integer Dimension 588
5. The dimension of a self-similar set 593
6. Exercises Involving Student Work: Measuring Two-dimensional Figures 596
CHAPTER 1

Functions

Functions are crucially important in all of mathematics, since they arise whenever some entity (the “output”) is determined by another (the “input”).

In school, we learn a tremendous amount about functions. We study the properties and graphs of specific functions like \( f(x) = x^2 \) and \( f(x) = \sin x \). We learn general techniques for graphing more complicated functions like \( f(x) = \frac{(x-1)(x-2)^2}{x-3} \). These techniques may start with simply plotting points, but later we learn about intercepts and asymptotes, and eventually, we discover how derivatives give crucial information about the shape of a graph. Meanwhile, we may become aware that functions occur in other settings in mathematics and elsewhere. For example, we might realize that the “area” can be thought of as a function, for which the inputs are geometric objects like triangles and the outputs are numbers.

This chapter, though, is not intended to recount information about specific functions, or the interplay between calculus and graphing. Rather, our attention will be focused on fundamental issues that arise as soon as functions are defined. We will be interested in questions such as the following: Can a function take on a value more than once? When does a function have an inverse? What does the graph of an inverse function look like? What is the relationship between functions and equations? We also discuss one of the most useful (and for many students, unmastered) topics in the high school curriculum: graphing transformations.

Let’s Go 1. Based on your understanding of the meaning of function, answer and discuss the following questions. (It may be helpful to draw a graph.)

(a) A sprinter races a 100 yard dash in eleven seconds. Can one define a function \( x(t) \), representing the distance she has traveled as a function of elapsed time? Can one define a function \( t(x) \), representing elapsed time as a function of the distance she has traveled? Explain.
(b) A rocket ship flies to the moon and back. Is the ship’s distance to earth a function of elapsed time? Is elapsed time a function of the ship’s distance to earth? Explain.

Let’s Go 2. In school, functions may be described using a variety of representations, such as a formula, a table, or a graph. Each representation has advantages and disadvantages. As an example, suppose the high temperature in degrees Fahrenheit in Pleasantdorf, IN on various days in 2008 is given in Figure 1. (The days selected are the first day of each month.)

<table>
<thead>
<tr>
<th>Day</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.1</td>
</tr>
<tr>
<td>32</td>
<td>24.0</td>
</tr>
<tr>
<td>60</td>
<td>32.7</td>
</tr>
<tr>
<td>91</td>
<td>40.2</td>
</tr>
<tr>
<td>121</td>
<td>71.3</td>
</tr>
<tr>
<td>152</td>
<td>69.3</td>
</tr>
<tr>
<td>182</td>
<td>82.5</td>
</tr>
<tr>
<td>213</td>
<td>85.5</td>
</tr>
<tr>
<td>244</td>
<td>76.3</td>
</tr>
<tr>
<td>274</td>
<td>64.3</td>
</tr>
<tr>
<td>305</td>
<td>45.0</td>
</tr>
<tr>
<td>335</td>
<td>40.2</td>
</tr>
</tbody>
</table>

Figure 1. High temperature as a function of day

(a) Explain how the scenario of this problem suggests a function $f : \{1, 2, 3, \ldots, 365\} \rightarrow \mathbb{R}$.

(b) Do you believe that there is a simple algebraic formula that gives $f$? Explain.

(c) What method might you use to find an algebraic-symbolic approximation for $f$?

(d) Absent a more complete table of values than Figure 1, what method might you use to give a likely value of the high temperature on April 15? On April 11?

(e) Draw a graph of $f$. What difficulties arise? How is the graph of $f$ helpful?

1. The Notion of a Function

1.1. Seeking a definition. Let’s imagine how a high school student might respond to the question, “What is a function?” She might reply by giving examples: “Well, a function is like $y = 10 + 3x$ or...
1. THE NOTION OF A FUNCTION

$f(x) = 4 \sin x$ or $1000e^t$.” This is not a bad response. Indeed, much of our time and effort—in both high school and college mathematics courses—is devoted to mastering very specific functions, such as polynomial, trigonometric, and exponential functions. These functions are important since they are useful for modeling phenomena in the physical world, in business, and in daily life.

Another student might focus on a version of the famous *vertical line test*. He might say, “If a vertical line crosses the graph more than once, then it’s not a function.” This gets close to the key idea in the definition of a function. Suppose we are trying to interpret some graph in the $xy$ plane as the graph of a function, but suppose also that some vertical line crosses the graph more than once. That would mean that two points $(x, y_1)$ and $(x, y_2)$ are both on the graph. This would leave us in a quandary, since if we wanted to define $y$ as a function of $x$, we wouldn’t know whether to choose $y_1$ or $y_2$ as the “correct” value of $f(x)$.

Yet another student might recall hearing about functions as “machines” that transform inputs into outputs. This comes closer to a definition; we still must clarify the meanings of “input,” “output,” and “transform.” First, we must identify a set of inputs, and the usual name for this set is the *domain of the function*. Second, we must identify a set that contains all of the expected outputs of the function. We’ll follow the usual practice in college-level mathematics courses and call this set the *codomain of the function*. The third issue—clarifying the meaning of “transforming an input into an output”—is the thorniest, and we postpone our discussion to Section 2.2.1. For the moment, we will content ourselves with an intuitive understanding; we may use the words “assignment,” “rule,” or “process” to describe what the “function machine” does in transforming a given input $x$ into the output $f(x)$. We are familiar with this concept through examples. For instance, we might consider the process that inputs a number $x$ and outputs the number $10 + 3x$; we are accustomed to writing this as $f(x) = 10 + 3x,$ or perhaps as $x \mapsto 10 + 3x,$ or perhaps more casually as simply “$10 + 3x$” (with the tacit assumption that the variable $x$ represents the input).

Summing up this discussion, we have the following provisional definition:

---

1. Perhaps you expected us to use the word *range* instead. We will have more to say about this in Section 4.1.
2. Were you expecting “$y = 10 + 3x$”? Functional notation will be discussed below in Section 2.3.
1. FUNCTIONS

**Provisional Definition 1.** Let $A$ and $B$ be nonempty sets. We define a *function from $A$ to $B$*, written $f : A \rightarrow B$, to be a rule that assigns, to each element $a \in A$, an element $f(a) \in B$. The set $A$ is called the *domain of the function*, and the set $B$ is called the *codomain of the function*.

The notation $f(a)$ is meant to emphasize something important: *the element $b = f(a)$ that is assigned to $a$ may depend on $a$*. In other words, different choices of the input $a$ are allowed to produce different outputs $f(a)$.

Also, we note that in spite of the description of $f$ as a “rule,” all that is required of $f$ is that $f(a) \in B$ for all $a \in A$. The “rule” need not be given by a simple formula.

**Your Turn 1.** In high school mathematics, the emphasis is usually on the rule and not on the domain and codomain. For example, we might write down the rule $f(x) = 2x$ or $g(x) = \frac{1}{x}$ without calling attention to the intended domain and codomain.

For $f$, and also for $g$, what would you consider reasonable domains? What are reasonable codomains? Is there more than one possible answer? Is there a largest or smallest reasonable choice for the domain or for the codomain? Explain.

**Your Turn 2.** Consider the curve illustrated in Figure 2. Imagine that we are moving along the curve (say, in a counterclockwise direction). As we move, our $x$ and $y$ coordinates change. However, these coordinates are not free to change independently; for example, we can’t move 0.1 units to the right and remain free to choose “any old” change in $y$. The fact that we must stay on the curve constrains us, due to the fact that $x$ and $y$ “satisfy a relation” with respect to each other. In fact, the way that $x$ and $y$ are related to each other along the curve is very explicit: the curve consists of the solutions of the equation $x^2 + y^4 = 1$.

(a) In what sense does the curve, or a portion of the curve, suggest a function $f(x)$ (that is, $y$ as a function of $x$)? Be certain to identify the domain, codomain, and rule for any function you identify.

(b) Likewise, in what sense does the curve, or a portion of the curve, suggest a function $g(y)$ (that is, $x$ as a function of $y$)?

(c) Draw a curve in the plane that defines $y$ as a function of $x$, but does not define $x$ as a function of $y$.
(c) Draw a curve in the plane that defines $y$ as a function of $x$, and also $x$ as a function of $y$. What are the domains of these two functions?

1.2. Why sets? Our provisional definition of a function (Provisional Definition 1) might surprise you because it talks about sets $A$ and $B$. Our experience from high school mathematics courses might be that the inputs and outputs for functions are always numbers. So why are we talking about general sets?

The reason is that there are many important examples of functions where the domain or codomain doesn’t consist of numbers. For example, consider the function that inputs people and assigns to each person his or her weight in pounds. Here the codomain would have to consist of numbers, but the domain consists of people. Likewise, consider the function that assigns to each person his/her blood type. Here, a reasonable choice for the domain is the set of all people (or any subset of this), and the reasonable choice for the codomain is $\{A, B, O, AB\}$. Neither the domain nor the codomain consists of numbers.

1.3. What the definition says, and what it doesn’t say. Let $f : A \to B$ be a function. We’ve said that to each input $a \in A$, the function $f$ assigns an output $f(a) \in B$. In the language mathematicians use, for every $a$ in $A$, there exists exactly one $b$ in $B$ such that $f(a) = b$. Mathematicians sometimes also express this as, for every $a$ in $A$, there is exactly one $b$ in $B$ such that $f(a) = b$. In less technical language—and
veering into imprecision—some might say, *Each input produces exactly one output*, or even, *For every a there’s exactly one b.*

**Your Turn 3.** A student considers \( f(x) = x^2 \), and sees that \( f(1) = 1 \), \( f(2) = 4 \), and \( f(3) = 9 \). He observes that 1, 4, and 9 are all outputs. However, he reads that \( f \) is a function if “each input produces exactly one output.” He observes that \( f \) produces more than one output, and in his mind, it doesn’t seem the case that “each input produces exactly one output.” He concludes that \( f \) is not a function. Discuss the source of his confusion.

To make certain that we are not reading any unwarranted assumptions into our understanding of function, we emphasize two points:

- For a function, it’s entirely possible that some of the elements of the codomain are not the output of any input. For example, for the function \( f : \mathbb{R} \to \mathbb{R} \), \( f(x) = 2^x \), if \( y = -1 \), then there is no \( x \) such that \( f(x) = y \). Thus, \( y \) is an element of the codomain, but \( -1 \) is not an output of the function.

- For a function, it’s entirely possible that two different inputs might produce the same output. For example, for the function \( f : \mathbb{R} \to \mathbb{R} \), \( f(x) = x^2 \), the inputs 3 and \(-3\) both produce the output 9.

**1.4. Exercises.**

1. 
   (a) For each mapping diagram (diagram with dots and arrows) in Figure 3, determine whether the diagram gives a function from \( A \) to \( B \).
   (b) What easy-to-spot property of the mapping diagram should one look for to determine whether the mapping diagram represents a function?

2. Let \( f(x) \) be given by the expressions below. Find all real values of \( x \) such that \( f(x) \) is a real number. (In precalculus and calculus books, this set is often called the natural domain of the function.)
   (a) \( \sqrt{x} \)
   (b) \( \sqrt[3]{x} \)
3. Explain why each of the following is not a function.

(a) \( A = \) all parents, \( B = \) all people,
\[ f : A \to B, \quad f(x) = x's \text{ children.} \]
1. FUNCTIONS

(b) \( A = \) all people, \( B = \) all people,
\[ f : A \rightarrow B, \ f(x) = x's \text{ firstborn child} \]

(c) \( A = \) all parents of at least one son, \( B = \) all males,
\[ f : A \rightarrow B, \ f(x) = x's \text{ firstborn child} \]

4. What would be reasonable choices for the domain and codomain for the following functions?

(a) \( f(x) = x's \text{ biological father} \)
(b) \( f(x) = x's \text{ spouse} \)
(c) \( f(x) = x's \text{ firstborn child} \)

5. Explain why each of the following is not a function.

(a) \( f : \mathbb{N} \rightarrow \mathbb{N}, \ f(x) = \sqrt{x} \)
(b) \( f : \mathbb{R} \rightarrow \mathbb{R}, \ f(x) = \arcsin(x) \)
(c) \( f : \mathbb{N} \rightarrow \mathbb{R}, \ f(x) = \pm \sqrt{x} \)

6. Explain why a sequence (intuitively, an infinite ordered list \( a_1, a_2, a_3, \ldots \)) should be defined to be a function whose domain is the set of natural numbers.

7. According to our provisional definition, a function involves a “rule” that assigns, to each point in the domain, a point in the codomain. No guarantees are made that the rule is given by a simple or transparent process.

(a) Give an example of a function with domain \( \{100, 101, \ldots, 200\} \) and codomain \( \{0, 1, \ldots, 9\} \) whose “rule” is not immediately discernable.

(b) Consider the function \( f \) with domain \( \{1, 2\} \) and codomain \( \mathbb{Z} \), given by the rule \( f(1) = 5, \ f(2) = 8 \).

Find three different polynomial functions that produce the function \( f \) when restricted to the input set \( \{1, 2\} \).

---

3In this and similar exercises, ignore issues of polygamy / plural marriage.
2. Graphs, Relations, and the Definition of a Function

From our experience in school, we are familiar with the task of graphing functions. The notion of graph makes sense for functions in general, and turns out to be the key to a rigorous definition of function.

2.1. The graph of a function. Let’s begin with graphing. Why do we graph functions in the first place?

One reason is that the graph of a function is the closest thing we have to a “table of values” in the case when the domain is infinite. Consider, for example, the function \( f : \mathbb{R} \to \mathbb{R}, f(x) = e^x \). We can’t “list” all possible values of \( e^x \) for every real number \( x \). However, from the graph of \( f \), we can read off the value of \( e^x \) (with a reasonable degree of accuracy) from the graph of \( f \), for any value of \( x \) that is depicted in the graph.

Your Turn 4. Figure 4 contains a graph of \( x \mapsto e^x \) for \(-4 \leq x \leq 4\). Use the graph to estimate values of \( e^{-2} \), \( e^1 \), and \( e^4 \). Then find accurate values of \( e^{-2} \), \( e^1 \), and \( e^4 \) using your calculator, and compute the percentage errors you made in estimating the values from the graph. What attributes of the graph made it easy or difficult to obtain accurate estimates?

In addition, having the graph of a function \( f : \mathbb{R} \to \mathbb{R} \) at hand allows us to understand many properties and behaviors of the function and its graph at a glance. Such properties/behaviors include where the function is increasing/decreasing, the graph’s concavity, its intercepts and its asymptotes.

Now it’s time to ask what we really mean by “the graph of a function,” first for a function from \( \mathbb{R} \) to \( \mathbb{R} \), and then for a function from any set \( A \) to any set \( B \). Suppose we have a function from \( \mathbb{R} \) to \( \mathbb{R} \), such as \( x \mapsto e^x \). The first thing we realize is that we draw the graph in a plane! Why do we do this? The reason is that we need “ordered pairs” to represent a pair of points: the element (here, a number) \( x \) in the domain and the element \( y \) in the codomain.

Then, which points in the plane do we “color in” when we draw the graph? They are the points \((x, y)\) such that \( y = f(x) \), in other words, the points of the form \((x, f(x))\).

\(^4\)Of course, we could give a table of values of \( e^x \) for the \( x \)-values \(-3, -2, -1, 0, 1, 2, 3\) and have a quite good idea of the behavior of the function \( x \mapsto e^x \). However, there are plenty of functions whose behavior is not at all transparent just from their values at small integers. Can you think of such a function?
Once we realize how we graph functions from $\mathbb{R}$ to $\mathbb{R}$, it’s easy to make a provisional definition of the graph of a function from a set $A$ to a set $B$, based on our provisional definition of a function:

**Provisional Definition 2.** Let $f : A \to B$ be a function. We define the **graph of $f$** to be $\{(a, f(a)) : a \in A\}$. The graph of $f$ is a subset of the Cartesian product $A \times B$. \(^5\)

**Your Turn 5.** Let $A = \{1, 2, 3\}$ and $B = \{x, y, z\}$. Let $f : A \to B$ be defined by the rule $f(1) = x$, $f(2) = x$, and $f(3) = y$. Draw $A \times B$ and the graph of $f$. (Suggestion: you might start by drawing $1, 2, 3$ on a horizontal axis and $x, y, z$ on a vertical axis.)

**Your Turn 6.** If $f : A \to B$ is a function, explain why for each each $a \in A$, there is exactly one point on the graph of $f$ whose first entry is $a$.

### 2.2. Functions, rigorously defined.

In Section 1.1, we were careful to call our definition of a function **provisional**, since we never clarified the meaning of “a rule that assigns, to each element of $A$, an element of $B$.” In this section, we will give a rigorous definition of function that gets around this difficulty.

\(^5\)Recall that the **Cartesian product** $A \times B$ is defined to be $\{(a, b) : a \in A$ and $b \in B\}$, a collection of ordered pairs.
2.2.1. *A rigorous definition of function.*

**Definition 3.** Let $A$ and $B$ be nonempty sets.

(a) A relation between $A$ and $B$ is any subset of $A \times B$.

(b) Let $S$ be a relation from $A$ to $B$. We say that $S$ is a function from $A$ to $B$ if given any $a \in A$, there is exactly one element of $S$ whose first entry is $a$.

**Your Turn 7.** Let $A = \{1, 2\}$ and let $B = \{p, q\}$. List all the relations between $A$ and $B$, and identify those that are functions from $A$ to $B$. (You should find sixteen relations and four functions.)

2.2.2. *Putting the ‘f’ in ‘function’.** Let $S \subseteq A \times B$ be a function (in the sense of Definition 3). Given any $a \in A$, there is a unique element $(a, b) \in S$. We define the symbol $f(a)$ by $f(a) = b$. Note that we have defined “$f(a)$” without yet attaching any meaning to the symbol “$f$” by itself.

**Your Turn 8.** For each of the four functions in Your Turn 7, identify $f(1)$ and $f(2)$.

2.2.3. *From Definition 3 back to Provisional Definition 1.* Let $A$ and $B$ be nonempty sets. Let’s start with a function from $A$ to $B$ in the sense of Definition 3. For each $a \in A$, we have a meaning for the symbol $f(a)$ (it is an element of $B$). We now consider $\{(a, f(a)) : a \in A\}$. Provisionally, we called this “the graph of $f$,” but in the sense of Definition 3, it is exactly the function! And while we used $S$ as shorthand for the function (a subset of $A \times B$), it makes more sense simply to use the letter $f$. Thus from the rigorous point of view, we can replace Provisional Definition 1 with the following:

**Definition 4.** Let $f \subset A \times B$ be a function. By definition, the graph of $f$ is a synonym for $f$. Note that by our definition of the symbol $f(a)$ in Section 2.2.2, we have

$$f = \text{graph}(f) = \{(a, f(a)) : a \in A\}.$$ 

Where does this leave us? We have rigorous definitions of *function* and *the graph of a function*. We also have an intuitive notion of function, reflected in Provisional Definition 1, as a “rule” or “assignment,” and some of the language and notation involving functions in Section 2.3 below is motivated by this intuitive
understanding. We will freely use such language and notation. This is pedagogically sound, since the notion of function as assignment is powerful and rooted in our school experience. With an eye toward rigor, however, we note that everything we write about functions can be expressed in terms of Definition 3; several Exercises in this chapter will ask the reader to translate notions involving functions from the setting of Provisional Definition 1 to the setting of Definition 3.

2.3. Functional notation. There is not complete uniformity in mathematicians’ preferences about notation for functions. For example, we have indicated that the functional rule that inputs \( x \) and outputs \( 10 + 3x \) can be notated in a variety of ways, such as “\( f(x) = 10 + 3x \)” and “\( x \mapsto 10 + 3x \)” What advantages accrue to different choices of notation?

The notations “\( f(x) = 10 + 3x \)” and “\( x \mapsto 10 + 3x \)” share the advantage of making it absolutely clear that \( x \) is being used as a variable representing elements of the domain. For example, if we were to write “the function \( tx^2 \)” it might be unclear whether we are regarding \( tx^2 \) as a function of the pair of variables \( t \) and \( x \), or whether we are regarding \( tx^2 \) as a function of \( x \) with \( t \) being an (unspecified) constant. These situations would be distinguished by the notation “\( f(t, x) = tx^2 \)” or “\( (t, x) \mapsto ax^2 \)” in the former case, as opposed to “\( f(x) = tx^2 \)” or “\( x \mapsto tx^2 \)” in the latter case.

Why might we prefer “\( f(x) = 10 + 3x \)” to “\( x \mapsto 10 + 3x \)”? The introduction of the letter \( f \) makes it convenient to write down the domain and codomain, employing the notation “\( f : \mathbb{R} \to \mathbb{R}, f(x) = 10 + 3x \)” Moreover, if we planned on “doing something” with the function, then it’s handy for it to have a name (for example, having invested the time to write “\( f(x) = 10 + 3x \),” we concisely can record the derivative as “\( f'(x) = 3 \)” ). On the other hand, if naming a function would only be an unnecessary distraction, it may be preferable to denote the rule as \( x \mapsto 10 + 3x \), without bothering to use \( f \) or \( g \) or any other letter as a name for the rule.

Some favor the notation “ \( y = 10 + 3x \)” for the rule \( x \mapsto 10 + 3x \). From a strict mathematical point of view, the introduction of the symbol \( y \) merely makes a choice of variable for points in the codomain, and this is not necessary to state the rule for the function. However, this notation is pedagogically important in terms of students’ development of the concept of a function. Students grow accustomed to “graphing” equations like \( x^2 + y^2 = 1 \) or \( y = 10 + 3x \), in the sense of plotting the points in the \( xy \) plane that satisfy
the equation. Then, the notion of the graph of a function (say the function \( x \mapsto 10 + 3x \)), as defined in Provisional Definition 2 in Section 2.1 above, is already familiar to students through the task of graphing the equation \( y = 10 + 3x \). In trying to interpret the equation \( y = 10 + 3x \) in functional terms, the reader does need to identify \( x \) as the input variable and \( y \) as the output variable. This involves more than mere conventional use of variable names, since the equation \( y = 10 + 3x \) also defines \( x \) as a function of \( y \), by the rule \( y \mapsto \frac{y - 10}{3} \).

On occasion, many of us are willing to sacrifice correctness for the sake of brevity. For example, we are comfortable writing “the function \( 10 + 3x \)” as shorthand for “the function whose rule is \( x \mapsto 10 + 3x \).” Likewise, a high school teacher might instruct his students to “graph the function \( e^x \)” instead of the more longwinded “graph the function \( f \) defined by \( f(x) = e^x \).”

2.4. Exercises.

1. Let \( A \) and \( B \) be finite sets. Suppose that \( A \) has \( a \) elements and \( B \) has \( b \) elements. Let \( f : A \to B \) be a function.

   (a) How many elements are in \( A \times B \)?

   (b) How many elements are in the graph of \( f \)? (Does the answer depend on \( f \)?)

2. In the text, we wrote that the graph of a function helps produce a table of values. Conversely, a table of values helps produce a graph. For “ordinary functions” from \( \mathbb{R} \) to \( \mathbb{R} \) (the sort one studies in a calculus class), what are the advantages of having data in tabular form? What are the advantages of a graph? Try to think of several advantages of each.

3. This exercise points to the special nature of a function as opposed to a general relation. Recall that if \( A \) and \( B \) are sets, then any subset of \( A \times B \) is termed a relation from \( A \) to \( B \).

   (a) Suppose \( f : A \to B \) is a function. Why is the graph of \( f \) an example of a relation from \( A \) to \( B \)?

   (b) Explain why there can be relations from \( A \) to \( B \) that are not graphs of functions.

   (c) Suppose that \( A \) is a set with five elements and \( B \) is a set with three elements. How many relations from \( A \) to \( B \) exist? How many of them are graphs of functions?
More generally, if $A$ has $m$ elements and $B$ has $n$ elements, what fraction of the relations from $A$ to $B$ are graphs of a function from $A$ to $B$?

4. This exercise illustrates limitations to the usefulness of graphs. Consider the function $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = e^{-1/x^2}$. (We define $f(0) = 0$.)

(a) Why is it reasonable to define $f(0) = 0$?

(b) Using technology, construct a table of values, for the inputs $\{0, 0.1, 0.2, 0.3, 0.4\}$.

(c) Using your answer to (b), attempt to construct a rough graph of $f(x)$ over the interval $[0, 0.4]$.

(d) What property of the function makes it so difficult to construct a meaningful graph?

(e) Use technology (graphing calculator or computer algebra system) to graph $f$. Discuss your results.

3. An Application/Interlude: Graphing Transformations

One of the most important ideas in the precalculus curriculum is that certain transformations of the graph of a function (a shift in the vertical or horizontal direction, a stretch in the vertical or horizontal direction, or a reflection over the $x$ or $y$ axis) can be achieved by a simple change in the formula for the function. Conversely, a function like $2\sin(x - \frac{\pi}{4})$ can be graphed by starting with the graph of $\sin x$ and performing a succession of such transformations. Hence, it’s important to know and understand the precise relationship between the graphing transformations and the changes to the expression of the function.

We begin our investigation in Example 1, with one particular graphing transformation (a horizontal shift). The rest of the work on graphing transformations is yours to do, through a succession of exercises. In Exercise 1, you will explore the other basic graphing transformations, and will see why the rules taught in precalculus indeed are valid. In further exercises, you will consider whether the order in which graphing transformations are performed is important, explore how graphing transformations move key features of graphs such as intercepts and asymptotes, and demonstrate your understanding by graphing functions using graphing transformations.

3.1. An example: a horizontal shift of a graph.

---

6In this section, we consider real-valued functions whose domain is also contained in the real numbers.
Example 1. Let \( f(x) = x^3 - 3x^2 = x^2(x-3) \). Its graph is given in Figure 5. Note the points \( A(-1,-4), \) \( B(0,0), \) \( C(1,-2), \) \( D(2,-4), \) \( E(3,0), \) and \( P(c,f(c)) \). You should think of the point \( P \) as a variable point on the graph.

Let \( g \) be the function whose graph is obtained from the graph of \( f \) by a horizontal shift to the right by 1 unit. Our goals are to understand how the graphing transformation is related to a transformation of the formula of \( f(x) \), and to find a symbolic expression for \( g(x) \). We accomplish this goal through the following three activities.

(i) Let \( A', B', C', D', E', P' \) be the points obtained from \( A, B, C, D, E, P \) by applying the graphing transformation. We can easily compute that these points are \( A'(0,-4), B' = (1,0), C' = (2,-2), D' = (3,-4), E' = (4,0), \) and \( P' = (c+1,f(c)) \).

(ii) We can verify our work in (i) by plotting the graphs of \( f \) and \( g \) on the same coordinate axes (see Figure 5).

(iii) We concentrate on the point \( P'(c+1,f(c)) \). By definition, this is a point on the graph of \( g \), so it can be written as \( (x,g(x)) \) for some \( x \). Equating these two points, we have \( c + 1 = x \) and \( f(c) = g(x) \). We easily obtain \( g(x) = f(c) = f(x-1) \). Thus, the symbolic expression for \( g(x) \) is obtained from the
symbolic expression of $f(x)$ by replacing $x$ with $x - 1$. For the specific function $f$ in this example, we obtain $g(x) = (x - 1)^3 - 3(x - 1)^2$ (which we could expand as a polynomial if we so desired).

3.2. Exercises.

3.2.1. Graphing transformations and expressions for functions.

1. Mimic the exposition in Example 1 to obtain a formula for $g(x)$, relative to each of the following graphing transformations applied to the graph of $f(x)$.

   (a) Shift upward by 3 units
   (b) Vertical stretch by a factor of 2
   (c) Horizontal stretch by a factor of $\frac{2}{3}$ (each point on the graph of $f$ moves farther from the $y$ axis in the graph of $g$)
   (d) Reflection across the horizontal axis
   (e) Reflection across the vertical axis

For each of these graphing transformations, you should execute the following steps:

   (i) Compute the coordinates of the points $A', B', C', D', E'$, and $P'$.
   (ii) Copy the graph of $f$, and use your answer to (i) to graph $g$ on the same coordinate axes.
   (ii) Take your answer for the coordinates of $P'$, and set this ordered pair equal to $(x, g(x))$. Equate the first coordinates, and obtain an expression for $c$ in terms of $x$. Finally, use the second coordinates to obtain a formula for $g(x)$ (in terms of $x$).

2.

   (a) On the basis of Exercise 1: to translate the graph of $f(x)$ to the right by $d$ units, we replace $x$ with $\underline{\hphantom{1}}$ in the expression for $f(x)$.
   (b) To stretch the graph of $f(x)$ horizontally by a factor of $c$, we replace $x$ with $\underline{\hphantom{1}}$ in the expression for $f(x)$.
   (c) Suppose we had the graph of $\sin(x^2 + 7x + 5)$ conveniently available to us. If we shifted this graph to the right by 1 unit, what function would we be graphing? (Find a symbolic expression for the function, obtained from the expression $\sin(x^2 + 7x + 5)$.)
3. AN APPLICATION/INTERLUDE: GRAPHING TRANSFORMATIONS

(d) A common incorrect answer to (c) is $\sin(x^2 + 7x + 4)$. Why is this incorrect? What misunderstanding lies behind the mistake?

(e) Is it easy to obtain the graph of $\sin(x^2 + 7x + 4)$ from the graph of $\sin(x^2 + 7x + 5)$? Give either a conceptual explanation or a response using graphing technology.

3.

(a) Draw the graph of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ (your choice).

(b) Using this graph, draw the new graph that results when the graph is stretched vertically (each point doubling its distance to the $x$ axis), and the new graph that results when the graph is compressed vertically (each point halving its distance to the $x$ axis). Call the first of these the graph of $g$, and the second, the graph of $h$.

(c) Find symbolic expressions for $g(x)$ and $h(x)$ (in terms of $f(x)$).

(d) It is sometimes difficult to choose and to interpret language relating to stretching transformations.

For each of the following instructions, would you produce the graph of $g$ or the graph of $h$?

(i) Stretch the graph of $f$ vertically by a factor of 2.

(ii) Stretch the graph of $f$ vertically by a factor of $\frac{1}{2}$.

(iii) Compress the graph of $f$ vertically by a factor of 2.

(iv) Compress the graph of $f$ vertically by a factor of $\frac{1}{2}$.

4. Repeat Exercise 3, using horizontal transformations instead of vertical ones.

5. Let $f(x) = e^{2x+5} \sin(5x^2 + 4x + 20) + \cos(1 + \ln(x^2 + 3))$. Suppose the graph of $f$ is shifted right 3 units and shifted upwards by 10 units. Find a symbolic expression for the new function being graphed. Simplify your answer.

3.2.2. Combining a horizontal transformation with a vertical transformation.

6.

(a) Let $f$ be a function. Find a formula for the function whose graph is obtained from the graph of $f$ by shifting two units to the left and then by stretching vertically by a factor of three.
the function whose graph is obtained from the graph of $f$ by stretching vertically by a factor of three, and then by shifting two units to the left.

(b) George and Laura are trying to graph $2\sin(3x)$. Laura wants to do the horizontal compression followed by the vertical stretch, but George insists on the reverse order. Which method is correct? Why?

(c) John wants to move up, and Teresa is willing to move to the left. Can they graph $\sin(x + \frac{\pi}{4}) + 1$? What two graphing transformations would they use? Does the order of the transformations matter?

(d) State a general principle about transformations, on the basis of the three examples above.

3.2.3. Combining a horizontal stretch and a horizontal shift.

7. (a) Given a function $f$, suppose the graph of $f$ is compressed horizontally by a factor of 3 (each point moves to a point whose distance to the $y$ axis is one third of its original distance). Suppose that after this transformation, the new graph is shifted to the left by two units. Find a symbolic expression for the function with this graph (in terms of $f$).

(b) Suppose that the graph of $f$ is transformed with the same two transformations as above, but performed in the opposite order. Find a symbolic expression for the function with this graph.

(c) Suppose we wished to graph $f(\frac{x}{3} + 5)$, using the graph of $f$. If we wished to obtain the graph of $f(\frac{x}{3} + 5)$ by a horizontal stretch followed by a horizontal shift, what would be the amounts of the stretch and the shift? If we wished to obtain the graph by a horizontal shift followed by a horizontal stretch, what would be the amounts of the shift and the stretch?

8. Calvin is doing his trigonometry homework, and is supposed to graph $\sin(2x - \frac{\pi}{2})$, using his knowledge of the graph of $\sin x$. He sees the answer on his graphing calculator, but is trying to understand how it comes from graphing transformations. He tries horizontally compressing the graph of $\sin x$ by 2 and shifting right by $\frac{\pi}{2}$, but it gives the wrong graph.

(a) If we actually did the sequence of graphing transformations that Calvin contemplated, what function would we be graphing?
(b) If we tried to graph \(\sin(2x - \frac{\pi}{2})\) using a compression followed by a shift, by what amount should we shift? Explain clearly.

(c) Hobbes takes pity on Calvin, and points out that Calvin could graph \(\sin(2x - \frac{\pi}{2})\), first with a horizontal shift, and then a horizontal compression. By what amount should Calvin shift? Explain clearly.

9. A function \(f(x)\) is graphed in Figure 6, and three points are included on the graph. Let \(g(x) = f(2x + 1)\).

(a) The graph of \(g\) can be obtained from the graph of \(f\) by a succession of two graphing transformations. State what the two graphing transformations are (including which one is performed first and which one is performed second). For each of these two graphing transformations, explain both the algebraic substitution in the formula for the function, and the effect on the graph.

(b) The three marked points on the graph of \(f\) produce, in a natural way, three points on the graph of \(g\). What are the coordinates of these three points?

(c) With help from your answer to (b), graph \(g\).
3.2.4. **Domain, range, and intercepts.**

10. Let $f$ be a function whose domain is $(3, \infty)$ and whose range is $[-2,1]$. What are the domain and the range of $7f(4x + 50) + 13$?

11. Suppose the graph of $f$ has a vertical asymptote of $x = 9$ and a horizontal asymptote of $y = 17$. What are the equations of the asymptotes of the graph of $-4f(-7x + 2) + 9$?

12. Suppose that the $x$-intercepts of the graph of $f$ are $\{4,7\}$. Find the $x$-intercepts of the graphs of the following functions. (If it is impossible to tell from the given information, explain why.)

   (a) $f(x - 3)$
   (b) $2f(7x + 1)$
   (c) $7f(x) + 1$

13. For which of the basic graphing transformations (vertical stretch, vertical shift, horizontal stretch, horizontal shift) applied to the graph of a function $f$, is it easy to find the $x$-intercepts of the new graph? Explain.

14. Suppose the $y$-intercept of the graph of $f$ is known. Is it easy to find the $y$-intercept of the graph that results from

   (a) a vertical shift of the graph of $f$?
   (b) a vertical stretch of the graph of $f$?
   (c) a horizontal shift of the graph of $f$?
   (d) a horizontal stretch of the graph of $f$?

15. Let $f(x) = x^3 - 7x^2$ and let $g(x) = 3f(x + 10)$. What are the roots of $g$? Solve the problem in two different ways: first by explicitly computing a factorization of $g(x)$, and second, by finding the roots of $f$ and then using graphing transformations.

16. Find the $x$-intercepts of the graph of the function $\sin(2x)$. Find the values of $x$ such that $\sin(2x) = 1$. How is this information useful in graphing $\sin(2x)$?
17. Repeat Exercise 16 with the function $\sin(x - \pi/3)$.

18. Use graphing technology to give an effective illustration of Exercises 16 and 17.

3.2.5. **Graphing specific functions.**

19. Suppose you wanted to graph the following functions. Identify the function with whose graph you would begin, and then the succession of graphing transformations you would perform. Finally, graph the function. Be certain to mark units on the $x$ and $y$ axes, or to explicitly mark the $x$ and $y$ coordinates of several points on the graph. (You may wish to check your answer using graphing technology, but you should attempt to construct the graphs only using your understanding of graphing transformations.)

   (a) $\tan(x - \pi/6)$
   (b) $(3 \sin x) + 1$
   (c) $\sin \frac{x}{4}$
   (d) $\cos(4x - \pi)$

3.2.6. **Additional topics.**

20. Let $f(x) = 2^x$. For this function, any shift to the left or right actually is equivalent to a vertical stretch! Check this:

   (a) A shift left by 1 unit corresponds to a vertical stretch by what factor?
   (b) A vertical stretch by a factor of $\frac{1}{8}$ corresponds to what horizontal shift?
   (c) A shift right by $c$ units corresponds to a vertical stretch by what factor?

21. Which graphing transformations (vertical stretches and shifts; horizontal stretches and shifts; reflections) preserve distances between points?

22. One can consider transformations of subsets of the plane that are not graphs of functions. For example, suppose $S \subset \mathbb{R}^2$ is given by $F(x, y) = 0$. Let $T$ be the subset of $\mathbb{R}^2$ obtained by applying one of the graphing transformations listed below to $S$. Find a function $G(x, y)$ such that $T$ is the subset of $\mathbb{R}^2$ where $G(x, y) = 0$. 
(a) Shift upward by 3 units
(b) Shift to the right by 1 unit
(c) Vertical stretch by a factor of 2
(d) Horizontal stretch by a factor of $\frac{1}{2}$ (each point on the graph of \( f \) moves farther from the \( y \) axis in the graph of \( g \))
(e) Reflection across the horizontal axis
(f) Reflection across the vertical axis

Check your answers for \( F(x, y) = (x - y)(x - 2) \).

4. Images and Inverse Images

A function \( f : A \rightarrow B \) sets up relationships between subsets of \( A \) and \( B \). The concepts and notation from this section are crucial for all developments in the rest of this chapter.

4.1. Images of points and sets. We are all accustomed to evaluating functions at elements of the domain. For example, evaluating \( f(x) = x^2 - 10x + 23 \) at 5, we obtain \( f(5) = -2 \). This makes sense for any function \( f : A \rightarrow B \): if \( a \in A \), then \( f(a) \) is an element of \( B \), and is called the image of \( a \) under \( f \). Often, mathematicians think of \( f \) as “pushing elements of \( A \) forward into \( B \).” It is also useful to push forward subsets of \( A \):

**Definition 5.** If \( S \) is a subset of \( A \), we define \( f[S] = \{ f(s) : s \in S \} \). The set \( f[S] \) is a subset of the codomain, and is called the image of \( S \) under \( f \).

Here are two ways to think about \( f[S] \):

- According to the definition, we obtain \( f[S] \) by evaluating \( f \) at each point \( s \) in \( S \), and then by collecting all the images \( f(s) \) into a set.

- We can also think about \( f[S] \) in terms of a test condition on the elements of the codomain \( B \), meaning a statement about \( b \in B \) that is true exactly when \( b \) is in \( f[S] \). Namely, given \( b \in B \), then \( b \) is an element of \( f[S] \) exactly when there exists an element \( s \in S \) such that \( f(s) = b \).
Note the difference in these two views of $f[S]$. In the first, we are running through the elements of $S$, and evaluating $f$ on each of them. In the second view, we are running through the elements of $B$, and “testing” each to see whether it is in $f[S]$.

**Your Turn 9.** Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2 - 10x + 23$.

(a) For the set $S = \{0, 1, 2\}$, compute the set $f[S]$. Which of the two interpretations of $f[S]$ (collection or test condition) are you using?

(b) Let $S$ be the closed interval $[0, 10]$. Is $b = 2$ an element of $f[S]$? Which interpretation of $f[S]$ (evaluation or test condition) are you using to solve this problem?

The set $f[A]$ is called the range of the function. It is a subset of the codomain. The range is just a special case of the image of a subset, so we can reiterate our two interpretations:

- We obtain the range by computing $f(a)$ for each $a \in A$, and collecting all these values into a set.
- Alternatively, we can obtain the range by the following test condition on elements of the codomain: an element $b \in B$ is in the range of $f$ exactly when there exists at least one element $a \in A$ such that $f(a) = b$.

**Your Turn 10.** Suppose $f : \mathbb{R} \to \mathbb{R}$ is a function. Carefully explain how we can describe the range of $f$ using the graph of $f$. Try to give two explanations, one for each of the two interpretations of the range.

### 4.2. Inverse images of sets.

We just showed that a function $f : A \to B$ “pushes forward subsets of $A$ to subsets of $B$.” We can also “pull back subsets of $B$” using the function:

**Definition 6.** If $T$ is a subset of $B$, we define $f^{-1}[T] = \{a \in A : f(a) \in T\}$. The set $f^{-1}[T]$ is a subset of $A$ and is called the inverse image of $T$ under $f$ or the pre-image of $T$ under $f$.

Just as with images, there are two ways to think about inverse images:

- One interpretation involves a test condition on elements of the domain $A$. To find $f^{-1}[T]$, we check each element $a$ of $A$. We compute $f(a)$ and check whether it is an element of $T$. If it is, we include it in $f^{-1}[T]$. If it’s not, we don’t.

- Here is a second interpretation. To compute $f^{-1}[T]$, we run through each element $t \in T$. We compute $\{a \in A : f(a) = t\}$ (this is exactly $f^{-1}([t])$, the inverse image of a set containing a single point).
We collect the inverse images; the collection of all elements obtained by this process (running through all elements \( t \in T \)) is exactly \( f^{-1}[T] \). Symbolically, we have that \( f^{-1}[T] = \bigcup_{t \in T} f^{-1}\{t\} \).

**Your Turn 11.** Which of the two interpretations (test condition or collecting inverse images) is closer to the actual definition of \( f^{-1}[T] \)?

**Your Turn 12.** Let \( f : \mathbb{R} \to \mathbb{R} \), \( f(x) = x^2 \).

(a) Let \( T \) be the closed interval \([50,300]\). Is \(-8 \in f^{-1}[T] \)? Which interpretation of \( f^{-1}[T] \) (test condition or collecting inverse images) are you using?

(b) If \( T = \{5,4,0,-1\} \), compute \( f^{-1}[T] \). Which interpretation of \( f^{-1}[T] \) (test condition or collecting inverse images) are you using?

(c) If \( T \) is the closed interval \([4,9]\), compute \( f^{-1}[T] \). (Start with the description \( \{a \in A : f(a) \in T\} \) and insert the appropriate specifics of \( A \), \( f \), and \( T \). You will obtain an inequality that you will have to solve.)

**Your Turn 13.** Let \( f : \mathbb{R} \to \mathbb{R} \) be a function and let \( T \subset \mathbb{R} \). Carefully explain how you can obtain \( f^{-1}[T] \) using the graph of \( f \).

Is it correct to think of the inverse image \( f^{-1}[T] \) as the image of \( T \) under a function \( f^{-1} : B \to A \)? The answer is yes exactly when an inverse function \( f^{-1} \) exists. Conditions on the existence of an inverse function will be given in Section 6.3.2.

### 4.3. Exercises.

1. Carefully explain how you can find \( f^{-1}[[5,4,0,-1]] \) and \( f^{-1}[[4,9]] \) from the graph of \( f(x) = x^2 \), where the domain of \( f \) is taken to be \( \mathbb{R} \). (This exercise should help reinforce what you learned in Your Turns 12 and 13.)

2. Let \( f : \mathbb{R} \to \mathbb{R} \), \( f(x) = -x^2 + 10x - 23 \). Let \( T = \{-7,-5,2,6\} \).

(a) Compute \( f^{-1}[T] \) (by solving four equations).

(b) Graph \( f \). Then, use the graph to illustrate how one obtains \( f^{-1}[T] \) from \( T \).
3. Suppose that $A$ and $B$ are finite sets and $f : A \to B$ is given by a table of values. For some value $b \in B$, how would you find $f^{-1}([b])$?

4. Let $A = \{1, 2, 3\}$ and $B = \{x, y, z\}$. Let $f : A \to B$ be defined by the rule $f(1) = x, f(2) = x$, and $f(3) = y$.
   
   (a) For each of the eight subsets $S$ of $A$, compute $f[S]$.
   
   (b) For each of the eight subsets $T$ of $B$, compute $f^{-1}[T]$.
   
   (c) State the domain, codomain, and range of $f$.

5. Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$.
   
   (a) Sketch the graph of $f$.
   
   (b) What is the range of $f$?
   
   (c) Let $S = \{-5, -4, 0, 1, 5, 9\}$. Compute $f[S]$.
   
   (d) Let $T = \{-9, -7, 0, 1, 16\}$. Compute $f^{-1}[T]$.
   
   (e) Let $S = [-1, 4)$, a half-open interval. Compute $f[S]$.
   

6. Here is a modified version of Exercise 5. Let $f : \mathbb{Z} \to \mathbb{Z}$, $f(x) = x^2$. (Carefully note the domain and codomain!)
   
   (a) Let $S = \{-5, -4, 0, 1, 5, 9\}$. Compute $f[S]$.
   
   (b) Let $T = \{-9, -7, 0, 1, 16\}$. Compute $f^{-1}[T]$.
   
   (c) What is the range of $f$? (Is it easy to describe?)
   
   (d) Sketch a reasonable portion of the graph of $f$.

7. Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$. Let $C = \{-3, 0, 3, 4, \sqrt{21}\}$. Compute each of the following:
   
   (a) $f[C]$
   
   (b) $f^{-1}[C]$
   
   (c) $f[f^{-1}[C]]$
   
   (d) $f^{-1}[f[C]]$
8. Let $A$ be the set of all humans, and let $B$ be the set of all human males. Define $f : A \to B$ by the rule $f(a) = a$’s biological father.

(a) In words, describe the range of $f$.

(b) If $y \in B$, then (in words) what is the set $f^{-1}([y])$?

(c) If $x \in A$, then (in words) what is the set $f^{-1}([x])$?

(d) If $y \in B$, then what is the set $f[f^{-1}([y])]$. Your answer should depend on whether $y$ is in the range of $f$.

9. Define $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \cos(x)$.

(a) Draw the graph of $f$.

(b) If $T$ is the closed interval $[\frac{1}{2}, 1]$, find $f^{-1}[T]$.

(c) What is the range of $f$?

(d) If $T$ is the closed interval $[\frac{1}{2}, 6]$, find $f^{-1}[T]$. How does question (d) differ from question (b)?

10. Let $f : (-\infty, \infty) \to (0, \infty)$, $f(x) = 2^x$.

(a) Compute $f(1)$, $f(3)$, $f([3])$, and $f(\log_2 9)$.

(b) Compute $f^{-1}([16])$ and $f^{-1}([15])$.

(c) Draw the graph of $f$.

(d) What is the range of $f$?

The next two exercises foreshadow our discussion of functions and equations in Section 7.

11. Let $f : \mathbb{R} \to \mathbb{R}^2$, $f(t) = (\cos t, \sin t)$.

(a) Describe the range of $f$.

(b) Draw the graph of $f$. (It will help first to consider (domain of $f$) $\times$ (codomain of $f$) and realize that the graph lives inside this.)
(c) Compute \( f^{-1}[\{(1, 0)\}] \). (Here \((1, 0)\) is a point in the \(xy\) plane, not an open interval.)

(d) Compute the inverse image under \( f \) of the first quadrant in the plane (that is, \( \{(x, y) : x, y \geq 0\} \)).

(e) In general, for a “reasonable” function with domain \( \mathbb{R} \) and codomain \( \mathbb{R}^2 \), what do you expect to be true about the graph? (Will it be a curve? A surface?)

(f) In general, for a “reasonable” function with domain \( \mathbb{R} \) and codomain \( \mathbb{R}^2 \), what do you expect to be true about the range? (Will it be a curve? A surface? Can you think of any exceptions to this principle?)

12. Let \( f : \mathbb{R}^2 \to \mathbb{R} \), \( f((x, y)) = x^2 + y^2 \).

(a) Describe the range of \( f \).

(b) Draw the graph of \( f \). (Again, it will help first to draw (domain of \( f \)) \times (codomain of \( f \)) and recall that the graph lives inside!)

(c) Let \( T \) be the closed interval \([4, 5]\). Compute \( f^{-1}[T] \).

(d) In general, for a “reasonable” function with domain \( \mathbb{R}^2 \) and codomain \( \mathbb{R} \), what do you expect to be true about the graph? (Will it be a curve? A surface?)

(e) In general, for a “reasonable” function with domain \( \mathbb{R}^2 \) and codomain \( \mathbb{R} \), what do you expect to be true about the range?

(f) Describe \( f^{-1}[\{1\}] \). How is this related to part (a) of Exercise 11?

13. Let \( f : \mathbb{R}^2 \to \mathbb{R}^2 \), \( f(x, y) = (2x + y, x - y) \). Let \( C \) be the set containing the single point \((3, 5)\) \(\in\) \(\mathbb{R}^2\).

(a) Compute \( f[C] \).

(b) Compute \( f^{-1}[C] \).

(c) Let \( S \) be the solid square in the plane with corners at \((0, 0)\), \((0, 1)\), \((1, 0)\), and \((1, 1)\). Compute \( f[S] \). (If you have had a course in linear algebra, you might recall a fact about the image of a parallelogram under a linear transformation of the plane.)

14. Let \( f : A \to B \) be a function.

(a) Prove that if \( S \subseteq A \), then \( f^{-1}[f[S]] \supseteq S \).

If you have had a vector calculus course, you should interpret “reasonable” function as differentiable function.
(b) Prove that if $T \subseteq B$, then $f[f^{-1}[T]] \subseteq T$.

15. Let $f : A \to B$ be a function. You may have noticed that when we evaluate $f$ at an element $a \in A$, we use round brackets, but when we consider images or inverse images of sets, we use square brackets. The reason for this is that confusion might result if we used round brackets for everything, particularly if $A$ includes elements that are sets. To illustrate this, let $A$ be the set consisting of the two elements $x$ and $\{x\}$, and let $B$ be the set consisting of the two elements 1 and 2. Let $f : A \to B$ be the function that takes the element $x$ to 1 and the element $\{x\}$ to 2.

(a) Suppose that we used round brackets for the images of both points and subsets. Explain the two interpretations of $f(\{x\})$ that would arise.

(b) In contrast, compute $f(\{x\})$ and $f[\{x\}]$, assuming the notation that has been developed in the text.

5. Injective, Surjective, and Bijective Functions

Guys, please improve this introduction! A function $f : A \to B$ represents a way in which elements of $A$ are related to elements of $B$, but it does not force a “pairing” of elements of $A$ with elements of $B$. The mathematical term to express the notion of “pairing” is bijection. The notions of bijection, injection, and surjection are defined and explored in this section.

5.1. Counting points in inverse images. The key property of a function $f : A \to B$ is: given any $a \in A$, there exists exactly one $b \in B$ such that $f(a) = b$. In this section, we investigate a different question: given any $b \in B$, how many elements $a \in A$ have the property that $f(a) = b$? Said another way, this question is, Given an element $b \in B$, how many elements are in the set $f^{-1}[\{b\}] \subseteq A$?

For most functions, the answer to the question is, “It depends on $b$.” For example, for the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$, we know that $f^{-1}[\{b\}]$ has two elements if $b > 0$, one element if $b = 0$, and no elements if $b < 0$. For most functions, it’s difficult to figure out how many elements are in $f^{-1}[\{b\}]$.

However, to understand key properties of functions, often one doesn’t need to know exactly how many elements are in $f^{-1}[\{b\}]$ for all $b$. Instead, one might only need answers for the following questions:
Question 1: For every \( b \in B \), does \( f^{-1}([b]) \) have at most one element?

Question 2: For every \( b \in B \), does \( f^{-1}([b]) \) have at least one element?

The answers to these two questions are important. Imagine a credit card company setting up accounts for its clients. Each account needs to have an account number, so we can consider a function \( f \) from the set of accounts to the set of positive integers, where

\[
f(\text{account}) = \text{the credit card number of the account.}
\]

In the context of this example, Question 1 asks, Given a number, is it the credit card number of at most one account? Clearly, the whole system depends on the answer to this question being yes, since chaos would ensue if two accounts shared the same credit card number.

On the other hand, Question 2 asks, Given a number, is it necessarily the credit card number of some account? You should suspect that the answer is no. In fact, since credit card numbers usually use 16 digits, there are \( 10^{16} \) different credit card numbers possible. On the other hand, estimating that there are about 6 billion people in the world, this means that there are more than a million times as many credit card numbers available as people in the world. Again, this is significant for the credit card system: there would be a lot more fraud involving credit cards if one could write down a number at random and expect it to be a valid number.

As a mathematical example, consider how logarithms are defined. We define the natural logarithm of a (positive) number \( b \) to be the unique number \( a \) such that \( e^a = b \). For this to make sense, \( a \) must exist and be unique. In order for \( a \) to exist, the answer to Question 2 (for \( f(x) = e^x \)) must be yes. Likewise, for \( a \) to be unique, the answer to Question 1 must be yes. (In fact, yes is the correct answer to both questions.)

Questions 1 and 2 motivate the definitions of injective and surjective functions, which we now discuss.

5.2. Injective functions.

**Definition 7.** Let \( f : A \to B \) be a function. We say that \( f \) is injective if any of the following equivalent properties holds:

(i) Given any \( a_1 \) and \( a_2 \) in \( A \), if \( a_1 \neq a_2 \), then \( f(a_1) \neq f(a_2) \).

(ii) Given any \( a_1 \) and \( a_2 \) in \( A \), if \( f(a_1) = f(a_2) \), then \( a_1 = a_2 \).
(iii) Given any \( b \) in \( B \), \( f^{-1}\{\{b\}\} \) contains at most one element. *Note the connection to Questions 1 and 2 from Section 5.1.*

An injective function is also called an *injection* for short. Synonymously, we say that an injective function is *one-to-one.*

**Your Turn 14.** Explain why the three properties in the definition indeed are equivalent.

**Your Turn 15.** Consider the functions indicated by the mapping diagrams in Figure 7.

(a) Use Property (i) in Definition 7 above to determine whether the function each injective.

(b) Using the same functions, compute the number of elements in \( f^{-1}\{\{b\}\} \) for each point \( b \) in the codomain. Use Property (iii) to check again which of the functions are injective.

(c) For a function given by a mapping diagram, how can you spot easily whether the function is one-to-one?

**Figure 7.** Figure for Your Turn 15

**Your Turn 16.**

(a) Consider the function \( f : \mathbb{R} \rightarrow \mathbb{R}, f(x) = x^2 \) and its graph. Is \( f \) injective? Which property in Definition 7 are you using to answer the question? Explain.
(b) If the graph of a function $f : \mathbb{R} \to \mathbb{R}$ is given, how can you spot easily whether the function is injective? (Try to formulate a precise statement about the intersections of certain lines with the graph.)

5.3. Surjective functions.

**Definition 8.** We say that a function $f : A \to B$ is surjective if any of the following equivalent properties holds:

(i) Given any $b \in B$, there exists at least one $a \in A$ such that $f(a) = b$.

(ii) $f[A] = B$. In words, the range of $f$ equals the codomain of $f$.

(iii) Given any $b \in B$, $f^{-1} \{b\}$ contains at least one element.

“Surjective function” is abbreviated as “surjection.” We also say that a surjective function $f : A \to B$ “maps $A$ onto $B$.” Even more loosely, mathematicians often call a surjective function “an onto function.” (This uses the word *onto* as an adjective instead of a preposition.)

**Your Turn 17.** Explain why the three properties in the definition indeed are equivalent.

**Your Turn 18.** Which of the functions in Your Turn 15 are surjective? For what easy-to-spot property of the mapping diagrams are you checking?

**Your Turn 19.** Is the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$ surjective? More generally, if $f : \mathbb{R} \to \mathbb{R}$ is a function whose graph is given, what property of the graph is equivalent to the function being surjective?

5.4. Bijective functions. A function that is both injective and surjective is a “bijective function,” or “bijection” for short. (We’ll see more about bijective functions later, since bijections are exactly the functions that have inverses.)

**Your Turn 20.** If $f : A \to B$ is given by a mapping diagram, how can you tell from the mapping diagram whether the function is a bijection?
Your Turn 21. Let $f : \mathbb{R} \to \mathbb{R}$ be a bijective function. What can you say about the intersection of the graph of $f$ with a horizontal line?

5.5. Exercises.

1. For each of the following functions, state whether the function is injective, and whether the function is surjective. Give a short but persuasive explanation.

   (a) $A =$ the students in your class who have a Social Security number, $B =$ the set of all nine-digit numbers, $f : A \to B, f(\text{person}) =$ that person’s Social Security number

   (b) $A =$ the letters a through z, $B =$ the words in the English language, $f : A \to B, f(\text{letter}) =$ the first word in the Oxford English Dictionary that begins with that letter

   (c) $A =$ the people who voted in the 2008 presidential election, $B =$ the 2008 presidential candidates, $f : A \to B, f(\text{person}) =$ candidate for whom the person voted

   (d) $A =$ the set of all married people, $f : A \to A, f(\text{person}) =$ the person’s spouse.

   (e) $A =$ the set of all English words of five or fewer letters, $B =$ $\{1, 2, 3, 4, 5\}$, $f : A \to B, f(\text{word}) =$ the number of letters in the word

   (f) $A =$ the people in your class, $B =$ the set of letters a through z, $f : A \to B, f(\text{person}) =$ the first letter in the person’s first name

   (g) $A =$ the set of all people, $B =$ the set of all males, $f : A \to B, f(a) =$ a’s biological father

2. Let $f : A \to B, f(x) = x^2$. For each choice of $A$ and $B$ below, state whether the function is injective and whether it is surjective, giving a brief explanation. (In one case, the “function” is not actually a function. Which one?)

   (a) $A = \mathbb{R}, B = \mathbb{R}$
5. INJECTIVE, SURJECTIVE, AND BIJECTIVE FUNCTIONS

(b) \( A = \mathbb{R}, \ B = [0, \infty) \)
(c) \( A = [0, \infty), \ B = \mathbb{R} \)
(d) \( A = [0, \infty), \ B = [0, \infty) \)
(e) \( A = \mathbb{N}, \ B = \mathbb{N} \)
(f) \( A = \mathbb{Z}, \ B = \mathbb{Z} \)
(g) \( A = \mathbb{R}, \ B = \mathbb{Z} \)

3. In reaction to the example in the text involving credit card accounts and credit card numbers, a student says, “The function \( f \) is injective since each account has only one credit card number.” Is the student correct? Explain.

4. Suppose that the domain of a function is a finite set, and the rule of the function is given by a table of values. How could you spot whether the function is injective?

5. Give definitions of injection and surjection that rely on the definition of a function in Definition 3 from Section 2.2.1.

The next four exercises illustrate the rarity of injective and surjective functions.

6.
(a) Suppose \( n \) is a positive integer and \( A \) is a set with \( n \) elements. How many functions \( f : A \to A \) exist? How many of these functions are injective?
(b) Taking your answers from (a), compute the ratio of the number of injective functions and the total number of functions, for \( 1 \leq n \leq 5 \). What conclusion can you draw?

7. Let \( A \) be a set with \( n \) elements and let \( B \) be a set with \( n + 1 \) elements. How many functions exist with domain \( A \) and codomain \( B \)? How many of these functions are injective? What percent of the total number of functions are injective when \( n = 1 \)? When \( n = 10 \)? When \( n = 100 \)?

8. Let \( A \) be a set with \( n + 1 \) elements and let \( B \) be a set with \( n \) elements. How many functions exist with domain \( A \) and codomain \( B \)? How many of these functions are surjective? What percent of the total number of functions are surjective when \( n = 1 \)? When \( n = 10 \)? When \( n = 100 \)?
9. Suppose $A$ has 5 elements and $B$ has 9 elements. How many functions exist with domain $A$ and codomain $B$? How many of these functions are injective?

10. Let $f : \mathbb{R} \to \mathbb{R}$.

(a) Recall (perhaps from a calculus course) the condition on $f$ that says that $f$ is an *increasing function*.

(b) Prove that if $f$ is increasing, then it is injective.

(c) Conversely, suppose that $f$ is injective. Is it necessary that $f$ be either an increasing function or a decreasing function? Give either a proof or a counterexample.

11. Suppose you are the teacher of this course and you are grading exams. You have asked students to complete the definition: *Let $f : A \to B$ be a function. We say that $f$ is injective if ....* Most of your students exactly quote one of the properties in the definition, and you give them full credit. However, a few students write down something else and you have to figure out what to do.

For each of the following “answers,” try to characterize the student’s response. You might use characterizations like “Absolutely correct,” “The student has ‘injective’ confused with another concept,” “The student has the right idea but what is written is incomplete or flawed,” “The student is writing a fact about injective functions but not the definition,” “The statement is comprehensible but false,” or “The student’s response is not comprehensible.” Briefly explain your answer.

(a) For all $b \in B$, $|f^{-1}([b])| \leq 1$.

(b) For every $y$ there’s at most one $x$.

(c) For every $x$ there’s at most one $y$.

(d) If $a_1 \neq a_2$, then $f(a_1) \neq f(a_2)$.

(e) For all $a_1, a_2 \in A$, if $f(a_1) = f(a_2)$, then $a_1 = a_2$.

(f) Every line hits the graph at most once.

(g) Every vertical line hits the graph at most once.

(h) Every horizontal line hits the graph at least once.

(i) Two things in the domain can’t go to the same point in the codomain.

(j) If $f(A) = B$ then $f(a) = f(b)$, $a \neq b$. 
(k) Two arrows can’t end at the same point.

(l) If \( a_1 = a_2 \) then \( f(a_1) = f(a_2) \).

12. Follow the directions for Exercise 11, with respect to the definition Let \( f : A \rightarrow B \) be a function. We say that \( f \) is surjective if . . . .

(a) Given any \( a \in A \), there exists \( b \in B \) such that \( f(a) = b \).

(b) Given any \( b \in B \), there exists \( a \in A \) such that \( f(a) = b \).

(c) Given any \( b \in B \), there exists \( a \in A \) such that \( f(x) = y \).

(d) Given any \( b \in B \), there exists \( a \in A \).

(e) For every \( y \) there’s an \( x \).

(f) Every horizontal line hits the graph.

(g) Every vertical line hits the graph at most once.

(h) Every horizontal line hits the graph exactly once.

(i) Every point in the codomain gets hit.

(j) Every point in the codomain has an arrow ending there.

(k) There’s an arrow at every point.

(l) \( b = f(a) \)

13. This exercise shows the close relationship between surjective functions and equivalence relations.

(a) Let \( f : A \rightarrow B \) be a surjective function. We define that \( a_1 \equiv a_2 \) if \( f(a_1) = f(a_2) \). Prove that \( \equiv \) is an equivalence relation.

(b) Conversely, suppose that \( A \) is a set and \( \equiv \) is an equivalence relation on \( A \). Find a set \( B \) and a function \( f : A \rightarrow B \) such that \( f(a_1) = f(a_2) \) exactly when \( a_1 \equiv a_2 \).

14. Let \( f : \mathbb{R} \rightarrow \mathbb{R}, f(x) = x^2(x - 3) \).

(a) Given a real number \( b \), find the number of elements in \( f^{-1}([b]) \). (The answer will depend on \( b \). It will be helpful to draw a rough graph of \( f \), and you probably will need ideas from calculus to complete this exercise.)
(b) Find three intervals whose union is $\mathbb{R}$, such that $f$ is injective as a function on each interval.

(c) Use your answers to (a) and (b) to define the domain and codomain of three bijective functions.

6. Composite Functions, Identity Functions, and Inverse Functions

In precalculus and calculus classes, we learned about specific pairs of inverse functions, such as the exponential and logarithmic function, and the sine function and arcsine function. In this section, we define and explore the notion of inverse function in the general context of functions between sets.

6.1. Composite functions. The notion of composition of functions arises whenever we wish to apply one function to the output of another.

Definition 9. Suppose $g : A \to B$ and $f : B \to C$ are functions. The composite function $f \circ g$ is the function with domain $A$ and codomain $C$, defined by the rule $(f \circ g)(a) = f(g(a))$. (See Figure 8.)

![Figure 8. Composition of two functions](image-url)

**Your Turn 22.** With the same $f$ and $g$ as in Definition 9, does $g \circ f$ make sense? Explain.

**Your Turn 23.** Let $A$ denote the 50 states in the U.S., let $B$ denote the citizens of the U.S., and let $C$ denote the political parties in the U.S. (include “none/independent” as a party). Let $g : A \to B$ assign, to each state, its governor; let $f : B \to C$ assign, to each US citizen, his/her party identification. In words, give the rule for the function $f \circ g$.

**Your Turn 24.** In school, we learn several ways to combine functions, and most of us learned about adding two functions and multiplying two functions before learning about composing two functions. Why have we not discussed addition and multiplication of functions in the general context of “functions between sets”?
6.2. The identity function. If someone asked us, “What is the ‘easiest’ function from $\mathbb{R}$ to $\mathbb{R}$?” how would we respond? We might suggest $f(x) = 0$, or $f(x) = 1$ (both of these are constant functions), or $f(x) = x$ (whose graph is a line).

Suppose instead that someone asked us, “What is the ‘easiest’ function from $A$ to $A$, if $A$ is a set?” We can’t suggest the constant functions $f(x) = 0$ or $f(x) = 1$, since the set $A$ might not contain either zero or one. However, we can still suggest $f(x) = x$, which makes sense for any set $A$:

**Definition 10.** Let $A$ be any set. We define the function $I_A : A \to A$ by the rule $I_A(a) = a$ for all $a \in A$. The function $I_A$ is called the identity function on $A$.

**Your Turn 25.** Let $f : A \to B$ be a function.

(a) What are the domain and codomain of the function $I_B \circ f$?

(b) How can you simplify the rule for the function $I_B \circ f$?

(c) What similar construction can you make involving $I_A$?

(d) In what other mathematical contexts have you used the word “identity”? Are there facts in these contexts that resemble (b) and (c)?

Identity functions are important for defining inverse functions, as we shall see in the next section.

6.3. Inverses of Functions.

6.3.1. Definition of the inverse of a function.

**Definition 11.** If $f : A \to B$ and $g : B \to A$ are functions, then $f$ and $g$ are said to be inverses of each other exactly if $g \circ f = I_A$ and $f \circ g = I_B$, or in other words, if $g(f(a)) = a$ for all $a \in A$ and $f(g(b)) = b$ for all $b \in B$.

**Your Turn 26.** Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = \frac{2}{5}x - 10$, and let $g : \mathbb{R} \to \mathbb{R}$ be defined by $g(x) = \frac{3}{2}x + 15$. Using the definition, show that $f$ and $g$ are inverses of each other.

**Your Turn 27.** Complete Figure 9 by drawing a figure for an inverse function $g$. Make certain that you use the same function $g$ in both portions of the diagram. From the completed diagram, how is it apparent that $f \circ g = I_B$ and $g \circ f = I_A$?
You may be accustomed to writing $f^{-1}$ for “the” inverse function of $f$. This notation presupposes that a function cannot have more than one inverse, which is true:

**Proposition 12.** Let $f : A \to B$ be a function. If $g_1$ and $g_2$ are both inverses of $f$, then $g_1 = g_2$.

**Proof.** The proof comes from simplifying $g_1 \circ f \circ g_2 : B \to A$ in two different ways. On the one hand, $g_1 \circ f \circ g_2 = (g_1 \circ f) \circ g_2 = I_A \circ g_2 = g_2$, but on the other hand, $g_1 \circ f \circ g_2 = g_1 \circ (f \circ g_2) = g_1 \circ I_B = g_1$. \qed

**Your Turn 28.** A friend thinks that the inverse function of $x \mapsto x^3$ should be $x \mapsto \frac{1}{x^2}$. What lies behind your friend’s misunderstanding? How could you convince your friend that he is wrong? Does the function $f : \mathbb{R} \to \mathbb{R}, f(x) = x^3$, have an inverse? Justify your answer.

### 6.3.2. Which functions are invertible?

Must every function have an inverse? The answer is no, but it is easy to characterize the invertible functions:

**Proposition 13.** A function $f : A \to B$ has an inverse if and only if it is bijective.

To verify Proposition 13, we first check that a bijective function does have an inverse. If $f : A \to B$ is a bijection, then by the definitions of *injection* and *surjection*, we conclude that for each $b \in B$, $f^{-1}[(b)]$ contains exactly one element of $A$. We call this element $f^{-1}(b)$. This gives a function $f^{-1} : B \to A$. 

---

**Figure 9.** Figure for Your Turn 27
Your Turn 29. Carefully explain why \( f^{-1} \circ f = I_A \) and \( f \circ f^{-1} = I_B \).

To finish the proof of Proposition 13:

Your Turn 30.

(a) Suppose \( f : A \to B \) is a function that is not injective. Explain why there cannot exist a function \( g : B \to A \) such that \( g \circ f = I_A \).

(b) Suppose that \( f \) is a function that is not surjective. Explain why there cannot exist a function \( g : B \to A \) such that \( f \circ g = I_B \).

6.3.3. Graphs of inverse functions.

Your Turn 31.

(a) Graph the function in Your Turn 27, along with its inverse function.

(b) Graph \( f(x) = x^3 \) and \( g(x) = \sqrt[3]{x} \). Plot at least five points on each graph.

Your Turn 31 may suggest to you a crucial fact about inverse functions. Namely, if \( f : A \to B \) and \( g : B \to A \) are inverses of each other, and if \((a, b)\) is a point on the graph of \( f \), then \((b, a)\) is a point on the graph of \( g \). Indeed, this is not hard to prove:

\[
(a, b) \text{ is on the graph of } f \quad \Rightarrow \quad b = f(a) \quad \Rightarrow \quad g(b) = g(f(a)) \quad \Rightarrow \quad g(b) = a \\
\Rightarrow (b, a) \text{ is on the graph of } g.
\]

Now suppose that \( f \) and \( g \) are functions from \( \mathbb{R} \) to \( \mathbb{R} \). It’s easy to see that in the plane, \((a, b)\) and the “corresponding point” \((b, a)\) are related in a simple way: each is obtained from the other by reflecting over the line \( y = x \). (Try it with a few points.) By our remarks, we can say that the graph of \( g \) is obtained from the graph of \( f \) by reflecting over the line \( y = x \).

Your Turn 32. Suppose that a (bijective) function \( f \) is represented by a table of values. How would you obtain a table of values for \( f^{-1} \)?
6.3.4. Inverting noninvertible functions. Let \( f : A \to B \) be a non-bijective function. We know that \( f \) does not have an inverse function. But how close can we come to producing “something like an inverse function”?

As an example, consider \( f : \mathbb{R} \to \mathbb{R}, f(x) = x^2 \). First off, \( f \) is not surjective, since the range is only \([0, \infty)\). This problem is easily solved: if we just replace the codomain with \([0, \infty)\), we suddenly have a surjective function \( f : \mathbb{R} \to [0, \infty) \).

The more serious problem is that \( f \) still is not injective (some horizontal lines intersect the graph more than once). Here we take a more drastic step: shrinking the domain of the function. In essence, we want to find as large a subset of the domain as possible on which the function is one-to-one. Here, an obvious choice is \([0, \infty)\); this amounts to “ignoring the left half of the graph of \( x^2 \).”

We now have a function \( f : [0, \infty) \to [0, \infty) \), which is bijective, hence it has an inverse function, which clearly is the function \( g : [0, \infty) \to [0, \infty), g(x) = \sqrt{x} \).

![Figure 10. f and g](image)

**Your Turn 33.** Suppose we “ignored the right half of the graph of \( f(x) = x^2 \).”

(a) What are the domain and codomain of \( f \)?

(b) Graph \( f \) and its inverse function \( g \).
(c) What are the domain and codomain of \( g \)?

(d) Guess a formula for \( g(x) \)

(e) Check your answer to (d) by computing \( f(g(x)) \) and \( g(f(x)) \).

This idea—shrinking the domain and codomain of \( f \) to obtain a bijective function—is crucial for defining the inverse trigonometric functions. If we start with \( f : \mathbb{R} \to \mathbb{R}, f(x) = \sin x \), then \( f \) has no inverse, since it’s neither surjective nor injective. We “make \( f \) surjective” by replacing the codomain with \([-1, 1]\). We make \( f \) injective by shrinking the domain so that \( f \) takes on no values more than once: specifically, we take \([-\pi/2, \pi/2]\) for the domain. This makes the sine function a bijective function from \([-\pi/2, \pi/2]\) to \([-1, 1]\), and its inverse function, the arcsine function, has domain \([-1, 1]\) and codomain \([-\pi/2, \pi/2]\).

6.3.5. Procedure for computing inverse functions. Assuming that a function has an inverse, there’s a procedure that often is helpful for computing the inverse. Here is an example. Let \( f : \mathbb{R} \to \mathbb{R}, f(x) = x^3 + 7 \). From the graph of \( f \), it’s easy to see that \( f \) is a bijection. Let \( g \) be the inverse function. We know that \( f(g(x)) = x \) for all \( x \), so \( x = f(g(x)) = (g(x))^3 + 7 \). We now solve \( x = (g(x))^3 + 7 \) for \( g(x) \): we have \((g(x))^3 = x - 7\), so \( g(x) = \sqrt[3]{x - 7} \).

Your Turn 34. Here is slightly different explanation of the procedure for computing inverse functions.

We begin with the equation \( y = x^3 + 7 \).

(a) Interchange \( y \) and \( x \) in the equation, and solve the new equation for \( y \).

(b) Give a convincing mathematical explanation of why this procedure gives the inverse of the function \( f(x) = x^3 + 7 \).

(c) What is the reason for interchanging \( x \) and \( y \)? Is it actually necessary for finding the inverse function?

6.4. Exercises.

1. In Let’s Go 1(a) from the beginning of this chapter, what is the relationship between \( x(t) \) and \( t(x) \)?
2. Let $A$ be the set of all married people, $B$ be the set of all males. Let $g : A \to A$, $g(\text{person})$ = the person’s spouse. Let $f : A \to B$, $f(\text{person})$ = the person’s father. Give the domain, codomain, and rule for $f \circ g$.

3. Let $A$ be any set and let $I_A$ be the identity function. Prove that $I_A$ is a bijection.

4. A beginning calculus student is asked to differentiate the function $F(x) = \sin(x^5)$. He writes $\frac{dF}{dx} = \frac{d\sin}{dx} x^5 + \sin \frac{dx^5}{dx} = \cos x^5 + \sin 5x^4$. What technique of differentiation has the student mistakenly applied? What misconception seems to be responsible for the student’s error?

5. Let $F : \mathbb{R} \to \mathbb{R}$, $F(x) = 3x + 5$. We can write $F$ easily as the composite of two functions. Namely, if we define $f$ and $g$ (both with domain and codomain $\mathbb{R}$) by the rules $f(x) = x + 5$ and $g(x) = 3x$, then $F = f \circ g$.

Find four other ways to write $F$ as the composite of two functions (with domain and codomain $\mathbb{R}$).

6. Suppose that $g : A \to B$ and $f : B \to C$ are injective functions. Prove that $f \circ g : A \to C$ is injective.

7. Suppose that $g : A \to B$ and $f : B \to C$ are surjective functions. Prove that $f \circ g : A \to C$ is surjective.

8. Let $g : A \to B$ and $f : B \to C$ be functions. Answer the following questions, giving either a proof or a counterexample, as appropriate.

(a) If $f \circ g$ is surjective, must $f$ be surjective?

(b) If $f \circ g$ is surjective, must $g$ be surjective?

(c) If $f \circ g$ is injective, must $f$ be injective?

(d) If $f \circ g$ is injective, must $g$ be injective?

9. Using Definition 3 from Section 2.2.1,

(a) Construct a definition of the composite of two functions. (Set up the appropriate notation for the two functions before you give the definition of the composite.)

(b) Construct a definition of the identity function from $A$ to $A$. 

10. Let $A = \mathbb{R} \setminus \{7\}$ (all real numbers except 7) and let $B = \mathbb{R} \setminus \{0\}$. Let $f : A \to B$, $f(x) = \frac{42}{x-7}$.

(a) Explain why $f(x)$ never equals 0. (Given our definition of $f$, why is it important and $f(x)$ be nonzero?)

Now let $g(x) = \frac{42}{x} + 7$.

(b) Explain why $g(x) \neq 7$ for all $x$.

This allows us to define $g$ as a function $g : B \to A$.

(c) Compute $f \circ g$ and $g \circ f$. (You should get $I_B$ and $I_A$, respectively, showing that $f$ and $g$ are inverses of each other.)

(d) Graph $f$ and $g$. Include at least five points on the graph of $f$ and the “corresponding” points on the graph of $g$.

(e) How are the asymptotes of $f$ and $g$ related to each other? How are the intercepts of $f$ and $g$ related to each other?

11. Compute the inverse function of each of the following bijections.

(a) $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 4x + 7$
(b) $f : \mathbb{R} \to \mathbb{R}$, $f(x) = -7(x - 2)^3 + 11$
(c) $f : (0, \infty) \to \mathbb{R}$, $f(x) = \log_{8} x + 5$
(d) $f : \mathbb{R} \setminus \{0\} \to \mathbb{R} \setminus \{0\}$, $f(x) = 1/x$
(e) $f : \mathbb{R} \setminus \{3\} \to \mathbb{R} \setminus \{1\}$, $f(x) = \frac{x-7}{x-3}$

12. Graph each of the functions in Exercise 11, along with their inverse functions. (Use technology if you want. Take time to observe that the graph of $g$ really is the reflection of the graph of $f$ over the line $y = x$, and that $(a, b)$ is on the graph of $f$ exactly when $(b, a)$ is on the graph of $g$.)

13. Let $X$ be the set of all married people, and let $f : X \to X$ be the function $f(x) = x$’s spouse.

(a) Carefully explain why $f$ is indeed a function.

(b) Carefully explain why $f$ is surjective.

(c) Carefully explain why $f$ is injective.
14. Let $C$ denote the set of circles in the plane centered at the origin, and let $\mathbb{R}^+$ denote the positive real numbers. Define $f : C \to \mathbb{R}^+$ by letting the image of a circle be the radius of the circle. Let $g : C \to \mathbb{R}^+$ by letting the image of a circle be the area of the circle.

(a) Explain why $f$ and $g$ are bijections.

(b) In words, explain the meaning of the composite function $g \circ f^{-1} : \mathbb{R}^+ \to \mathbb{R}^+$, and find an explicit formula for $g \circ f^{-1} : \mathbb{R}^+ \to \mathbb{R}^+$.

(d) Do the same for $f \circ g^{-1}$.

15. This exercise gives an example of “inverting a noninvertible function.” Let $f(x) = 3x^2 - 12x + 7$.

(a) Graph $f$ as a function with domain $\mathbb{R}$. Be certain to include the vertex and the intercepts.

(b) Show that as a function from $\mathbb{R}$ to $\mathbb{R}$, $f$ is neither injective nor surjective.

Now consider $f$ as a function from $[2, \infty)$ to $[-5, \infty)$.

(c) Graph $f$, including at least three points on the graph, with their $x$ and $y$ coordinates marked.

(d) Explain why $f$ is a bijection.

(e) Compute the inverse function $f^{-1}$. (You can do this by using the quadratic formula, or by “completing the square.”)

(f) Graph the inverse function, with the three “corresponding” points marked.

(g) What are the domain and range of the inverse function?

16. Find an appropriate domain and codomain for the rule $f(x) = \tan x$ so that the function is bijective.

Graph $f$ along with its inverse function.

17. This exercise points to the importance of considering both $f \circ g$ and $g \circ f$ when determining whether $f$ and $g$ are inverses of each other.

(a) Consider the rule $x \mapsto (\sqrt{x})^2$. Assuming that the domain consists of nonnegative real numbers, graph the function and find the range. Why would we be disinclined to take $\mathbb{R}$ as the domain?

(b) Consider the rule $x \mapsto \sqrt{x^2}$. Assuming that the domain is $\mathbb{R}$, graph the function and find the range.
(c) Let \( f(x) = x^2 \) with domain \( \mathbb{R} \) and codomain \([0, \infty)\), and let \( g(x) = \sqrt{x} \) with domain \([0, \infty)\) and codomain \( \mathbb{R} \).

(i) What is the domain of \( f \circ g \)?
(ii) Is \( f \circ g \) the identity function on its domain?
(iii) What is the domain of \( g \circ f \)?
(iv) Is \( g \circ f \) the identity function on its domain?

18. This Exercise gives an example of “inverting a noninvertible function,” and also points out difficulties that technology may have with inverse functions.

Let \( f(x) = \sin x \), with domain \([-\pi/2, \pi/2]\) and codomain \([-1, 1]\). It is easy to see that \( f \) is a bijection.

Let \( g(x) \) denote its inverse function (commonly called \( \sin^{-1}(x) \) or \( \arcsin(x) \)).

(a) Graph \( f(x) \) and \( g(x) \), carefully observing the domain and codomain of these function.
(b) What are the domain and codomain of \( f \circ g \)? Graph \( f \circ g \).
(c) What are the domain and codomain of \( g \circ f \)? Graph \( g \circ f \).
(d) By hand, carefully graph \( \arcsin(\sin(x)) \), using \([-4\pi, 4\pi]\) as the set of inputs. Explain how this question differs from (b).
(e) Use graphing technology (graphing calculator or computer algebra system) for the problem in (d). Does your choice of technology give the correct answer?
(f) According to our definitions, should the rule \( \sin(\arcsin(x)) \) give a function if one uses the domain \([-2, 2]?)? Why or why not? What does your choice of graphing technology produce if you request the graph of \( \sin(\arcsin(x)) \) with domain \([-2, 2]?\)

19. Having learned about composite functions and inverse functions in Mrs. D.’s prealgebra class, Suzie conducts an exploration. She starts with the functions \( f(x) = x + 5 \) and \( g(x) = 3x \), and computes the composite \( (f \circ g)(x) = 3x + 5 \). She computes the inverse function \( (f \circ g)^{-1}(x) \) to be given by the rule \( \frac{x}{3} - \frac{5}{3} \).

However, she finds that \( f^{-1}(x) = x - 5 \) and \( g^{-1}(x) = \frac{x}{3} \), hence \((f \circ g)^{-1}(x) = (f^{-1} \circ g^{-1})(x) = \frac{x}{3} - 5 \). Find Suzie’s mistake, and explain how to correct it.
20. Let \( f : A \to B \) be any function. Show that there is a bijective function \( g : A \to \text{graph of } f \). Find a formula for the inverse of \( g \).

21. Let \( f : A \to B \) and \( g : B \to A \) be functions. We say that \( f \) is a left inverse of \( g \) and \( g \) is a right inverse of \( f \) if \( f \circ g = I_B \), and we say that \( g \) is a left inverse of \( f \) and \( f \) is a right inverse of \( g \) if \( g \circ f = I_A \).

   (a) Let \( A = \{a, b, c, d\} \) and \( B = \{1, 2, 3\} \). Find functions \( f : A \to B \) and \( g : B \to A \) such that \( f \circ g = I_B \) but \( g \circ f \neq I_A \).

   (b) Can you accomplish the task in (a) if \( A = \{a, b, c\} \)? Explain fully.

   (c) Let \( \sin : \mathbb{R} \to [-1, 1] \) and let \( \arcsin : [-1, 1] \to \mathbb{R} \). Is \( \sin \) a left inverse of \( \arcsin \)? Is \( \arcsin \) a left inverse of \( \sin \)? Explain.

   (d) How would you illustrate the phenomenon in (c) using the “squaring” and “square root” functions?

22. Let \( A \) denote the set of all functions \( h : \mathbb{R} \to \mathbb{R} \) that are differentiable with continuous derivatives, and let \( B \) denote the set of all functions \( k : \mathbb{R} \to \mathbb{R} \) that are continuous. Let \( f : A \to B \) be defined by \( f(h) = h' \) (the derivative of \( h \)). Let \( g : B \to A \) be defined by the rule that \( g(k) = \) the antiderivative \( K \) of \( k \) such that \( K(7) = 4 \).

   (a) Explain why \( f \) and \( g \) actually exist. (You may need some results from calculus. Where do you use the precise definitions of \( A \) and \( B \)?)

   (b) In the language of Exercise 20, is \( f \) a left inverse of \( g \)? Is \( g \) a left inverse of \( f \)? Explain.

23. This exercise is a continuation of Exercise 14.

   (a) Use a computer algebra system to solve \( y = x^2(x - 3) \) for \( x \).

   (b) Use part (a) to find the inverse function for each of the three functions you found in part (c) of Exercise 14.

7. Subsets and Equations

The examples and exercises in this section assume that the reader has some experience with linear algebra and with graphing in \( \mathbb{R}^3 \).
We have emphasized that the notion of a function already relies on the notion of a set (since the domain and codomain of any function are sets). In this section, we will study two methods through which, conversely, functions are helpful in describing sets. We will begin with a statement of the two methods, and work through these methods in the context of some examples. Finally, we’ll conclude with clarification of how these two methods for describing subsets are intimately connected with the problem of finding solutions of equations, which occupies so much of our attention in school.

Let’s Go 3. Below is a list of equations or systems of equations. What do we mean by the “solution” of each equation or system? How would you describe the “solution set” (say, in geometric language)? How would you interpret the instruction to “solve” the equation or system—what would the goal be, and what method would you use?

(a) \( x^2 - 3x + 2 = 0 \)
(b) \( 3x + 2y = 8 \)
(c) \( 3x + 2y + 4z = 1 \)
(d) \( x^2 + y^2 = 1 \) and \( 3x = 2y \)

Are there any ambiguities in these tasks? Is it possible to interpret them in a way that changes the answers?

1. Varying the domain and codomain of a function.

(a) Let \( B \subseteq C \). Merely with this information, show that there is a function with domain \( B \) and codomain \( C \). Call this function \( I_{B \rightarrow C} \). (Hint: if \( B = C \), then your function should be the identity function on \( B \).)

(b) Let \( f : A \rightarrow B \) be any function. Find the domain, codomain, and range of the function \( I_{B \rightarrow C} \circ f \).

(c) Let \( D \subseteq A \). Merely with this information, show that given any function \( f : A \rightarrow B \), there exists a function \( R^A_D f : D \rightarrow B \). This function is called the restriction of \( f \) to \( D \). (Hint: if \( D = A \), then \( R^A_D f \) should be \( f \) itself.)

(d) Let \( f : \mathbb{R} \rightarrow [-1, 1] \) be the sine function. If \( g \) is the sine function, regarded as a function from \([0, \pi/2]\) to \( \mathbb{R} \), how would you write \( g \) in terms of \( f \) and the notation developed in (a) and (c)?
(e) Let $f: A \to B$ and suppose that $E$ is a set with the property that $\text{range}(f) \subseteq E \subseteq B$. Explain how $f$ gives rise to a function with domain $A$ and codomain $E$.

(f) Give an example of the phenomenon described in (e). Be certain to give $A$, $B$, $f$, range($f$), and $E$.

7.1. Two methods of describing a subset. Suppose we have a set $B$ and our goal is to give an explicit description of a particular subset $S \subseteq B$. We now indicate two methods for describing $S$.

**Inverse Image Method.** We can seek a set $C$, a subset $T \subseteq C$, and a function $g: B \to C$ such that $S = g^{-1}[T]$. By this method, we are describing $S$ as the inverse image of another set, namely $T$, under the function $g$. We think of membership in $S$ as given by a “test condition”: given $b \in B$, we “test” whether $b \in S$ by checking whether $g(b) \in T$.

**Image Method.** We can seek a set $A$, a subset $U \subseteq A$, and a function $f: A \to B$ such that $f[U] = S$. In this second method, we are describing $S$ as the image of another set, namely $U$, under the the function $f$. This method produces elements of $S$, since whenever we pick $u \in U$ and compute $f(u)$, we obtain an element of $S$.

The following diagram will help us to remember these methods and our notation for them:

$$
\begin{align*}
A & \xrightarrow{f} B & & \xrightarrow{g} C \\
\cup & & \cup & & \cup \\
U & \xrightarrow{} f[U] = S = g^{-1}[T] & & \xrightarrow{} T
\end{align*}
$$

We explore these methods in the following six exercises.

**Example 1.** Let’s say we are interested in the circle in $\mathbb{R}^2$ that is centered at the origin and has radius 4. In our notation, we are letting $B = \mathbb{R}^2$ and are letting $S$ denote the circle.

The Inverse Image Method depends on realizing that we know an equation for the circle, namely $x^2 + y^2 = 16$. So let us define a function $g: \mathbb{R}^2 \to \mathbb{R}$, by $g(x, y) = x^2 + y^2$ (so $C = \mathbb{R}$). The fact that the circle is given by the equation $x^2 + y^2 = 16$ is just another way of saying that $S = g^{-1}[T]$, where $T = \{16\}$. 

The Image Method depends on our ability to parameterize the circle. Namely, we can write any point on the circle \( S \) in the form \((4 \cos t, 4 \sin t)\), with \(0 \leq t < 2\pi\). We define \( f : \mathbb{R} \to \mathbb{R}^2 \) given by \( f(t) = (4 \cos t, 4 \sin t) \) (we take \( A = \mathbb{R} \)). Then \( S = f[U] \), where \( U \) is the interval \([0, 2\pi)\).

Now let’s consider diagrams that help us understand \( f \) and \( g \).

Since \( g : \mathbb{R}^2 \to \mathbb{R} \), its graph is a subset of \( \mathbb{R}^3 \), namely, the collection of points \( \{(x, y, x^2 + y^2) : x \in \mathbb{R}, y \in \mathbb{R}\} \). (See the first part of Figure 11.) Note that when we intersect the graph with the horizontal plane at height 16, we obtain a circle of radius 4.

**Figure 11. Views of the circle \( S \)**

**Your Turn 35.** Carefully explain the relationship between the circle just described, and the circle \( S \).

To understand \( f \), rather than drawing the graph of \( f \), it’s perhaps easier simply to draw an annotated picture of \( S \subset B \) (see the second part of Figure 11). For a collection of representative points of \( S \), we have labeled the points by their pre-image in \( U \) under \( f \) (that is, by a value of \( t \)). This is a quite standard way of depicting a parameterized curve.

**Your Turn 36.** Consult a vector calculus textbook to find the definitions of level set and contour graph. Are these words associated with the descriptions of \( S \) as an inverse image or as an image? 

---

\(^8\)In fact, we also could have taken \( U \) to be all of \( A \). Our choice of \( U \) as \([0, 2\pi)\) has the merit that the parameterization produces each point on the circle exactly once, or otherwise said, the restriction of \( f \) to \( U \) gives a bijection between \( U \) and \( S \).
Example 2. Suppose that instead of the circle, we wanted to describe an annulus—say the region \( S \) in the plane strictly between the circles of radius 3 and 4, respectively, centered at the origin. As before, we take \( B = \mathbb{R}^2 \) since \( S \) occurs naturally as a subset of \( \mathbb{R}^2 \).

We first use the Inverse Image Method. This time, \( S \) is described easily by inequalities, namely, as the region in the \( xy \) plane given by \( 9 < x^2 + y^2 < 16 \). So the Inverse Image Method gives the following description: if we let \( C = \mathbb{R} \) and if we let \( g : B \to C, g(x,y) = x^2 + y^2 \), then the annulus is the inverse image \( g^{-1}[T] \), where \( T \) is the open interval \((9,16)\).

We now turn to the Image Method. Since the region \( S \) is not a curve, we shouldn’t expect to parameterize it by a subset \( U \subseteq \mathbb{R} \). Instead, it makes sense to parameterize it with a subset of \( \mathbb{R}^2 \). Polar coordinates lurk behind the easiest parameterization: we let \( A = \mathbb{R}^2 \) and let \( f : A \to B, f(r,\theta) = (r \cos \theta, r \sin \theta) \). Then, \( S = f[U] \), where \( U \) is the rectangle \((3,4) \times [0,2\pi) \subset A \).

Your Turn 37. Draw the graph of \( g \). How would you represent the sets \( T \) and \( S \) in this picture?

Precisely how do the set \( T \) and the graph of \( g \) together yield the set \( S \)?

Your Turn 38. Draw a diagram that convincingly illustrates \( S \) as a surface that is parameterized by \( U \).

(How should you modify the second part of Figure 11? How do you indicate the roles of both \( r \) and \( \theta \)?)

Example 3. Consider the unique line \( S \) in the plane that contains the points \((2,0)\) and \((0,3)\).

On the one hand, since an equation of the line is \( 3x + 2y = 6 \), we can describe \( S \) by the Inverse Image Method as \( g^{-1}([6]) \), where \( g(x,y) = 3x + 2y \). Using the Image Method, since we can parameterize the line as \( \{(2 - 2t,3t) : t \in \mathbb{R} \} \), we have that \( S = f[\mathbb{R}] \), where \( f(t) = (2 - 2t,3t) \).

Your Turn 39.

(a) Carefully explain why the line is given parameterically as \( \{(2 - 2t,3t) : t \in \mathbb{R} \} \).\(^9\)

(b) Using our standard notation, what are \( A, B, C, U, \) and \( T \) in this example?

(c) Draw a diagram like Figure 11 that illustrates our two methods of describing the line \( S \).

(d) We could have solved the equation \( 3x + 2y = 6 \) for either \( x \) or for \( y \). Carefully explain how this would give two additional parameterizations of the line.

\(^9\)This topic is discussed in the chapter Lines.
Your Turn 40. Suppose instead that $S$ were the region in the plane lying between the lines $3x + 2y = 6$ and $3x + 2y = 12$. Use both the Inverse Image Method and the Image Method to describe $S$.

Example 4. Suppose the set $B$ consists of the collection of individuals {Alice, Bob, Carol, Denise, Edward}. There is a rather important subset $S$ of $B$, that we can describe in two simple ways, either as the collection of females in $B$, or (by listing) as the set {Alice, Carol, Denise}. These two ways of describing $S$ fit into our general picture of describing subsets either as inverse images or as images. On the one hand, we can define the set $C$ to be the letters {$M, F$} ($M$ for male, $F$ for female) and define $T =$ {$F$}. If we let $g : B \to C$ be the function that associates, to each person, his or her sex, then $S = g^{-1}[$$T$$]$ (the Inverse Image Method). On the other hand, we can let $A = U =$ {$1, 2, 3$} and define $f : A \to B$ by the rule $f(1) =$ Alice, $f(2) =$ Carol, and $f(3) =$ Denise, since then we have $S =$ $f[U]$, the Image Method. 10 While this may seem contrived, it’s exactly the right contrivance to “parameterize” a finite set.

Example 5. Let $S$ be the set of real roots of the polynomial $x^2 - 4x + 1$. Applying the Inverse Image Method, we can describe $S$ by the function $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^2 - 4x + 1$; then $S = g^{-1}[\{0\}]$. As in the last example, $S$ is a finite set (concretely, the set {$2 \pm \sqrt{3}$}), so, by the Image Method, we can describe $S$ as the image of $f : \{1, 2\} \to \mathbb{R}$, defined by $f(1) = 2 + \sqrt{3}$ and $f(2) = 2 - \sqrt{3}$ (a fancy way of listing the roots).

Your Turn 41.

(a) Using our standard notation, identify $A$, $B$, $C$, $U$, and $T$ in this example.

(b) Draw an analogue of Figure 11 for this example. Are the parts equally helpful? Why or why not?

Example 6. Let $S$ be the set of solutions of the equation $x + 2y + 3z = 4$ in $\mathbb{R}^3$. The definition of $S$ is tantamount to an easy Inverse image Method description of $S$, namely, $S$ is the inverse image of the set {$4$} under the function $g : \mathbb{R}^3 \to \mathbb{R}$, $g(x, y, z) = x + 2y + 3z$. How do we describe $S$ as an image? Since $S$ is a plane, it’s no surprise that we should try using $\mathbb{R}^2$ to parameterize $S$. In fact, viewing $S$ as the graph of a function (the function $(x, y) \mapsto 4/3 - x/3 - 2y/3$), we could write $S = f[\mathbb{R}^2]$, where $f : \mathbb{R}^2 \to \mathbb{R}^3$, $f(x, y) = (x, y, 4/3 - x/3 - 2y/3)$. There are many other parameterizations.

10Equally well, we could have taken $A =$ {1, 2, 3, 4, 5}, $U =$ {1, 3, 4}, and defined $f$ by the rule $f(1) =$ Alice, $f(2) =$ Bob, $f(3) =$ Carol, $f(4) =$ Denise, and $f(5) =$ Edward.
7.2. **Solution Sets.** Let’s imagine Example 5 in a classroom setting. Suppose Mrs. D. instructs her algebra class to “find the roots of \(x^2 - 4x + 1\).” If Suzie responds, “The roots are the numbers \(r\) such that \(r^2 - 4r + 1 = 0\),” or (with greater sophistication) “The roots are the elements of the set \(g^{-1}([0])\), where \(g : \mathbb{R} \rightarrow \mathbb{R}, g(x) = x^2 - 4x + 1\)” (the Inverse Image Method), then Suzie has clearly misunderstood her teacher’s intentions. While Suzie has given perfectly correct characterizations of the roots of \(x^2 - 4x + 1\), surely Mrs. D. wants her students to use the quadratic formula, or complete the square, to obtain \(2 \pm \sqrt{3}\).

There are several lessons to be learned from this episode.

- When one is asked to “find the solution set” of a system of equations or inequalities, usually the statement of problem is very close to an Inverse Image Method description of a set.
- Solving the system usually means describing the solution set via the Image Method.
- Implementing the Image Method usually requires techniques that depend on the setting of the problem, and may or may not depend on “solving equations” in the sense of high school algebra.

**Your Turn** 42. We revisit Example 1.

(a) How would you interpret the instruction “Solve the equation \(x^2 + y^2 = 16\)?

(b) How would you interpret the instruction “Find the solution set of the equation \(x^2 + y^2 = 16\)?

(c) When we implemented the Image Method in Example 1, did we complete the task in (a)? In (b)? Explain your answer.

**Your Turn** 43. Imagine that you are back in high school, and your teacher asks you to “solve the inequality” \(x^4 - 3x^3 + 2x^2 > 0\).

(a) Explain how the Inverse Image Method describes (in an uninteresting way) the solution set \(S\) for this problem. Be certain to give \(g\), \(C\), and \(T\).

(b) Solve the inequality \(x^4 - 3x^3 + 2x^2 > 0\), as you would be expected to in a precalculus class. In hindsight, what strategy did you use to complete this task? What other strategies might you have used? Do these strategies differ in significant ways?

(c) How did you choose to present your answer in (b)? What alternate ways could you have used?
(d) Does your answer in (b) represent the set of solutions as the image of a function? If not, could it be made to do so easily?

7.3. Exercises.

2. We revisit Example 6.

(a) How would you interpret the instruction “Solve the equation $x + 2y + 3z = 4$”?

(b) How would you interpret the instruction “Find the solution set of the equation $x + 2y + 3z = 4$”?

(c) When we implemented the Image Method in Example 1, did we complete the task in (a)? In (b)? Explain your answer.

3. Again, we consider Example 6. Let’s imagine that students have been asked to produce a variety of descriptions of $S$, and they have produced the list below. Consider each description. Does it give a “test condition” for membership in $S$? Is it a parameterization of $S$? Is it both? Is it some other sort of description of $S$? (There is some room for disagreement.)

(a) $\{(x, y, z) : x + 2y + 3z = 4\}$

(b) $\{(x, y, 4 - \frac{x}{3} - \frac{2y}{3}) : x, y \in \mathbb{R}\}$

(c) The plane containing $(4, 0, 0)$ that is perpendicular to the vector $(1, 2, 3)$.

(d) The unique plane that contains the points $(4, 0, 0)$, $(0, 2, 0)$, and $(0, 0, 4/3)$.

(e) $\{(4 - 2y - 3z, y, z) : y, z \in \mathbb{R}\}$

4. Mrs. D. has shown how to parameterize the unit circle using trigonometric functions, but Jimmie wonders why he can’t just take the equation $x^2 + y^2 = 1$ and solve for $y$. Show how to use this idea to parameterize the circle. (Find $A$, $U$, and $f$. What must you do to make certain you cover both the top half and the bottom half of the circle?)

5. Describe each of the following subsets of $\mathbb{R}^3$ using the Inverse Image Method:

(a) The unit sphere, centered at the origin, in $\mathbb{R}^3$

(b) The (solid) ball of radius 5, centered at the origin, in $\mathbb{R}^3$
1. FUNCTIONS

(c) The equator of the unit sphere (centered at the origin) in \( \mathbb{R}^3 \)

(d) The hollow vertical cylinder (of infinite height) in \( \mathbb{R}^3 \), centered around the \( z \) axis and having radius 7

(e) The solid vertical cylinder in \( \mathbb{R}^3 \), of height 4, whose base lies on the plane \( z = 0 \), centered around the \( z \) axis and having radius 13

(f) The \( z \)-axis in \( \mathbb{R}^3 \)

6. Describe each of the subsets in Exercise 5 using the Image Method. (There are a variety of ways of doing this, but you probably will want to review the cylindrical and spherical coordinate systems.)

7. Let \( S \) be the line segment with endpoints \((2,0)\) and \((0,3)\). Describe \( S \) by the Image Method and by the Inverse Image Method.

8. Suppose \( S \) is a curve in \( \mathbb{R}^2 \), given by an equation in \( x \) and \( y \). Suppose it is possible to solve the equation for \( y \) in terms of \( x \). Explain how to describe \( S \) by the Image Method (give \( f \), \( A \), and \( U \)). Apply your answer to discuss the curve given by the equation \( xy - 3y = \ln x \).

9. Generalize Exercise 8 for a surface in \( \mathbb{R}^3 \). Give an example. (What condition on the equation will allow you to parameterize the surface easily?)

10. We investigate a good parameterization of a hyperbola.

(a) Graph the hyperbola \( x^2 - y^2 = 1 \).

(b) Show that for all \( t \in \mathbb{R} \), \( f(t) = \left( \frac{e^{t} + e^{-t}}{2}, \frac{e^{t} - e^{-t}}{2} \right) \) lies on the hyperbola.\(^{11}\)

(c) Using a calculator, compute approximations of \( f(-2), f(-1), f(0), f(1), f(2) \) and plot these points along the graph you produced in (a).

(d) From (c), do you believe that the range of \( f \) is the entire hyperbola? From the definition of \( f \), can you determine the range of \( f \)? Discuss.

(e) Find a parameterization of the right-hand branch of the hyperbola that does not use \( f \). (Can you parameterize the branch using \( x \) as a parameter? Using \( y \) as a parameter?)

\(^{11}\)The functions \( \frac{e^t + e^{-t}}{2} \) and \( \frac{e^t - e^{-t}}{2} \) are denoted \( \cosh t \) and \( \sinh t \), respectively, and are called the hyperbolic cosine and sine functions. Completing (b) means verifying the identity \( \cosh^2 t - \sinh^2 t = 1 \).
11. Use a computer algebra system to make a contour plot of the function \( x^2 - y^2 \), and a parametric plot of \( \left( \frac{e^t + e^{-t}}{2}, \frac{e^t - e^{-t}}{2} \right) \), as in Exercise 10. How would you parameterize the level curves of the function \( x^2 - y^2 \) (curves with equation \( x^2 - y^2 = c \) for \( c \in \mathbb{R} \))? 

12. A textbook reports that a particular curve is given by the “parametric equations" \( x = s^2 \) and \( y = s^3 \). 

(a) Has the curve been described by the Image Method or by the Inverse Image Method? Justify your answer. 

(b) How would you describe the curve by the other method? 

In your answers to (a) and (b), you should give \( A, B, C, T, U, f, \) and \( g \). 

(c) By hand, sketch the curve. Briefly explain your method. 

(d) What graphing technology could you use to help sketch the curve? 

13. We investigate the Image and Inverse Image Methods for the intersection of two planes in \( \mathbb{R}^3 \). 

(a) There is a unique plane in \( \mathbb{R}^3 \) that contains the points \((1,0,0), (0,1,0), \) and \((0,0,1). \) What is the equation of this plane? 

(b) Likewise, there is a unique plane in \( \mathbb{R}^3 \) that contains the points \((2,0,0), (0,3,0), \) and \((0,0,4). \) What is the equation of this plane? 

Let \( S \subseteq \mathbb{R}^3 \) be the intersection of the two planes described in (a) and (b). 

(c) Geometrically, what is the nature of \( S \)? (Is it a plane, a line, a point, a circle...?) 

(d) Explicitly describe \( S \) by the Inverse Image Method. (You need to find \( C, T, \) and \( g. \)) 

(e) Explicitly describe \( S \) by the Image Method. (You need to find \( A, U, \) and \( f. \)) 

(f) Did it take more work to accomplish the task in (d) or in (e)? 

(g) Use your answer to (c) to find five elements of \( S \). Would it have been easy to find five points in \( S \) from your answer to (d)? 

(h) Is \((8,-15,8) \in S\)? Is it easier to determine this using your answer to (d) or to (e)? 

14. In \( \mathbb{R}^2 \), solve the system of equations 

\[ \{(y - 2x - 1)(x - 2) = 0, (y - 2x - 1)(y - 3) = 0\} \]
by writing the solution as the union of a few simple subsets of the plane.

15. Let \( S \) denote the ellipse in the plane that includes the point \((0, 1)\) and has foci \(\{\pm(1, 0)\}\). Describe \( S \) by both the Inverse Image and Image Methods. Your answer should include \(A, B, C, T, U, f,\) and \(g\).

16. Parameterize the two surfaces in \(\mathbb{R}^3\) given by the equations below. (Sketches of the surfaces can be found in most vector calculus books.) Try to use the cosine, sine, hyperbolic cosine, and hyperbolic sine functions (see Exercise 10). It will help if you recall some facts about rotation matrices from linear algebra.

\[(a) \quad x^2 + y^2 - z^2 = 1\]
\[(b) \quad x^2 + y^2 - z^2 = -1 \text{ (with } z \geq 1)\]

17. Suppose \( S \) is the graph of a function \( h : M \to N \).

\[(a) \quad \text{Show how to express } S \text{ as } g^{-1}[T] \text{ for some sets } T \subseteq C \text{ and some function } g : B \to C.\]

\[(b) \quad \text{Show how to express } S \text{ as } f[U] \text{ for some sets } U \subseteq A \text{ and some function } f : A \to B.\]

(It might help to think about a specific example, such as \( h : \mathbb{R}^2 \to \mathbb{R}, h(x, y) = x^2 + y^2.\))

18. One of the early tasks that a linear algebra class tackles is solving systems of linear equations. In a paragraph, interpret this task in the context of the methods and notation of this section.

19. In many areas of mathematics, one encounters formulae (for various things) in pairs—one formula for “images” (parameterized objects) and another formula for “inverse images” (level sets).

As an example, suppose you have a surface \( S \) embedded in \(\mathbb{R}^3\). Given a point on the surface, there is a length-one vector that is perpendicular to the surface at that point (a unit normal vector).\(^\text{12}\)

Answer the following questions, perhaps using a calculus book as a reference:

\[(a) \quad \text{Suppose } S \text{ is the level set defined by the equation } F(x, y, z) = c. \text{ Find an expression for a unit normal vector. Justify your answer.}\]

\[(b) \quad \text{Suppose that } S \text{ is given parameterically by functions } x(s, t), y(s, t), z(s, t). \text{ Find an expression for a unit normal vector. Justify your answer.}\]

\(^\text{12}\)Actually there are two such vectors, since if \(v\) is a unit normal vector, then so is \(-v\).
(c) As indicated in the footnote, there are two unit normal vectors at each point, oppositely directed from each other. In your answers to (a) and (b), what characterizes the choice of unit vector that you have obtained?

20. Let \( f : \mathbb{R} \to \mathbb{R}, f(t) = (t, t^2, t^3) \). Let \( S \subset \mathbb{R}^3 \) denote the image of \( f \) in \( \mathbb{R}^3 \).

(a) Attempt a sketch of \( S \). Why is this somewhat difficult? What general statements can you make that describe the curve?

(b) Find a description of \( S \) using the Inverse Image Method.
CHAPTER 2

Lines in the Plane

In this chapter we will explore lines in the plane, which are highly valued for their simplicity and utility, and which command an extensive presence in mathematics. Fundamental mathematical applications involving lines include:

- Simple ‘rules of thumb,’ such as determining temperature by counting cricket chirps, or the rule of 72 for loans. (This is often accomplished by linear regression, or by linear interpolation.)
- Local approximation to curves via tangent lines (that is, the whole of differential calculus).
- Linear programming (used to optimize a linear function on a region with linear boundaries).

As we proceed through the chapter, we will describe lines geometrically and algebraically, forging a connection between the two descriptions. Along the way we will encounter parametric equations for lines as well as the fundamental concepts of parallelism, perpendicularity, and distance. We finish with a discussion of linear regression and an exploration of parameterized families of lines in the plane.

Let’s Go 1. In this exercise we ask you to explore your prior knowledge of lines in the plane.

(a) List as many mathematical topics/concepts as you can that rely upon or are connected with lines in the plane.

(b) What should secondary students learn about lines in the plane?

(c) What aspects of lines in the plane are likely to cause confusion for secondary students?

1. Linear Equations

1.1. Algebraic and geometric interpretations of lines. If asked to describe a line in the plane, most of us would provide an informal pictorial description. For example:

- A line is a curve uniquely determined by two points, or
• A line is a curve with constant ‘steepness’, or
• A line is a curve that provides a path of minimal distance between any two points on the curve.

Given that we all have a powerful intuitive understanding of a line as a geometric object, we might find the following algebraic definition to be unsettling:

**Definition 1.** Let \( A, B, C \) be real numbers.

(a) A linear equation in the variables \( x \) and \( y \) is an equation of the form \( Ax + By = C \).

(b) The linear equation \( Ax + By = C \) is nontrivial if at least one of the numbers \( A, B \) is nonzero.

(c) A line in the \( xy \)-plane is the solution set of a nontrivial linear equation \( Ax + By = C \).

**Your Turn 1.** Refer to Definition 1.

(a) Plot the solutions for each equation. Which equations have solutions that are not lines in the plane?

(i) \( y = 3x + 5 \)  \hspace{1cm} (ii) \( y - \sqrt{2} = 15(x - 4) \) \hspace{1cm} (iii) \( x/5 + y/7 = 1 \) \hspace{1cm} (iv) \( x + 2y + 3z = 6 \)

(v) \( -1483.85x + 0y = 67.9 \) \hspace{1cm} (vi) \( 0x + 0y = 0 \) \hspace{1cm} (vii) \( 0x + 0y = 3 \).

(b) Why is the word ‘nontrivial’ important in part (c) of Definition 1? Explain.

While the definition of a line presented in Definition 1 might seem far removed from our informal picture of a line, it does have a number of significant advantages, such as providing us with the exact location (\( x \)- and \( y \)-coordinates) of each point on a line. Therefore we must reconcile Definition 1 with our intuitive geometric notion of a line. We explore this in Section 2 and in the following Your Turn.

**Your Turn 2.** One intuitive picture of a line is that of a curve determined by two points.

(a) Find a linear equation with solutions containing the points \((1, 3)\) and \((4, 9)\).

(b) Which concepts about lines (not discussed in this section) came to mind as you completed part (a)?

1.2. Exercises.

1. Suppose that a given line is horizontal and passes through the point \((x_1, y_1)\). Find a nontrivial linear equation whose solution set is the line. Do the same for a vertical line passing through \((x_1, y_1)\).
2. Suppose that $Ax + By = C$ is a nontrivial linear equation, $(x_1, y_1)$ and $(x_2, y_2)$ are two different points on the line, and $x_1 = x_2$. Prove that $B = 0$, and hence, the solution of the equation is the vertical line containing the two given points.

3. (Some special lines) Let $Ax + By = C$ be a nontrivial linear equation and let $L$ be the corresponding line. Discuss why

   (1) $L$ is horizontal if and only if $A = 0$.
   (2) $L$ is vertical if and only if $B = 0$.
   (3) $L$ contains the origin if and only if $C = 0$.

4. In this exercise, we investigate uniqueness issues in linear equations.

   (a) Three students are asked to produce an equation for the line passing through the points $(1, 3)$ and $(5, \frac{9}{2})$. The students each produce ‘different’ final answers, namely $y - 3 = \frac{3}{8}(x - 1)$, $y - \frac{9}{8}x = \frac{21}{8}$, and $8y - 3x = 21$. Are all of these equations correct? Discuss.

   (b) A problem on an exam states “Find the equation for the line with slope $-6$ and $x$-intercept $\sqrt{3}$.” Considering your answer to part (a), suggest a better way to phrase the problem.

5. Show that if $A_1 x + B_1 y = C_1$ and $A_2 x + B_2 y = C_2$ are two linear equations both possessing $(1, 3)$ and $(4, 9)$ as solutions, then both equations have exactly the same set of solutions. (That is, we’re showing that there is a unique line passing through $(1, 3)$ and $(4, 9)$.)

6. Prove that $Ax + By = C$ and $A'x + B'y = C'$ are equations for the same line if and only if there exists a nonzero real number $\lambda$ such that $A' = \lambda A$, $B' = \lambda B$, and $C' = \lambda C$.

2. Intercepts, Slope, and Equations of Lines

In this section we explore some familiar concepts surrounding lines in the plane, including intercepts, slope, and special equations for lines. Using the slope concept, we forge a connection between the algebraic definition of a line given in Definition 1 and our intuitive geometric picture.
2.1. Slopes of lines. For most curves in the plane, the notion of ‘slope’ (i.e., rate of change of $y$ with respect to $x$, or ‘steepness’) is a local phenomenon. That is, slope depends upon location on the curve. The local nature of slope complicates matters when trying to pin down rates of change for specific curves, and partially drives the existence of entire branches of mathematics (e.g., differential calculus, differential equations) devoted to studying ‘slopes.’ However, for (non-vertical) lines the notion of slope is global: the slope remains constant no matter where we are. This makes the slope behavior of lines relatively easy to understand, and explains why lines are often used to approximate more complicated curves.

We probably all remember that to compute the slope of a line, one computes the ratio \( \frac{y_2 - y_1}{x_2 - x_1} \) for any two distinct points \((x_1, y_1)\) and \((x_2, y_2)\) on the line. Of course, this computation only makes sense if we get the same number for any two distinct points on the line:

**Proposition 2.** Consider a non-vertical line with equation \( Ax + By = C \). Given any two different points \((x_1, y_1)\) and \((x_2, y_2)\) on the line, the ratio \( m = \frac{y_2 - y_1}{x_2 - x_1} \) does not depend on the choice of the two points. It is called the slope of the line. In terms of the original equation of the line,

\[
m = -\frac{A}{B}.
\]

**Proof.** Since \((x_1, y_1)\) and \((x_2, y_2)\) lie on the line, we know \( A(x_1 - x_2) + B(y_1 - y_2) = 0 \) (why?). Further manipulation gives

\[
m = \frac{y_1 - y_2}{x_1 - x_2} = -\frac{A}{B}.
\]

Since \(-\frac{A}{B}\) is a constant, the ratio \( \frac{y_2 - y_1}{x_2 - x_1} \) does not depend on the choice of two points we pick on the line. \( \square \)

What about vertical lines? One might choose to say that the slope is undefined, since \(-\frac{A}{B}\) is not defined if \( B = 0 \). It would be more informative to say that a vertical line “has infinite slope.” However, one cannot distinguish between \( +\infty \) and \( -\infty \) in this case (see Figure 1).
Your Turn 3. How does Proposition 2 help us to reconcile the algebraic definition of a line given in Definition 1 with our intuitive geometric notion of a line? (Several intuitive descriptions of lines are given in a bulleted list at the beginning of Section 1.)

Lines are the unique ‘smooth’ curves in the plane for which the slope computation \( \frac{y_2 - y_1}{x_2 - x_1} \) is constant for all points on the curve\(^1\). This doesn’t happen for other types of curves:

Your Turn 4. Find three points \((x_1, y_1)\), \((x_2, y_2)\), and \((x_3, y_3)\) on the parabola with equation \( y = x^2 \), such that \( \frac{y_2 - y_1}{x_2 - x_1} \neq \frac{y_3 - y_2}{x_3 - x_2} \).

2.2. Intercepts. Suppose \( Ax + By = C \) is a nontrivial linear equation. The \( x \)-intercepts of the corresponding line are defined to be the real number solutions to \( Ax = C \), while the \( y \)-intercepts are the real solutions to \( By = C \).

Your Turn 5. How else might one describe ‘\( x \)-intercept’? ‘\( y \)-intercept’? Argue that your descriptions are consistent with the intercept definitions given immediately above.

\(^1\)There is a good chance you proved this in your calculus class. See exercises below.
2.3. Forms for the equation of a line. Thus far we have focused on the equation \(Ax + By = C\), which is the most general form for the equation of a line in the sense that every line in the plane can be expressed using this equation. Of course there are alternate linear equations, all obtainable from \(Ax + By = C\). While being less general (since they may not apply to certain lines), these alternate forms for the equation of a line are very useful: they have geometric interpretations, and often they are easier to obtain from data. For example, given a point on the line and its slope, the slope-intercept form of the equation of the line is the quickest equation to write down.

These alternate forms are summarized in the following table:

<table>
<thead>
<tr>
<th>Equation</th>
<th>Name</th>
<th>Allowable lines</th>
<th>Info at a glance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y = mx + b)</td>
<td>Slope intercept</td>
<td>nonvertical</td>
<td>slope, y-intercept</td>
</tr>
<tr>
<td>(x = ny + a)</td>
<td>inverse slope intercept</td>
<td>nonhorizontal</td>
<td>reciprocal slope, x-intercept</td>
</tr>
<tr>
<td>(y - y_0 = m(x - x_0))</td>
<td>point-slope</td>
<td>nonvertical</td>
<td>slope, point on line</td>
</tr>
<tr>
<td>(\frac{x}{a} + \frac{y}{b} = 1)</td>
<td>double intercept</td>
<td>nonzero intercepts</td>
<td>x- and y-intercepts</td>
</tr>
</tbody>
</table>

Your Turn 6. Consider the linear equations given in the table above.

(a) Are the \(a\) and \(b\) that appear in the first two equations the same as the \(a\) and \(b\) that appear in the fourth equation?

(b) What is the meaning of \(n\), and why is \(x = ny + a\) called the ‘inverse slope intercept’ form?

(c) Why can’t every equation of the form \(Ax + By = C\) be put into each of the forms listed in the table?

(d) How is the ‘information at a glance’ really there at a glance?

Your Turn 7. Consider the linear equations given in the table above.

(a) Which form is best for identifying \(x\) as a function of \(y\)? For identifying \(y\) as a function of \(x\)?

(b) Given an equation for a line, suppose you wish to graph the line. Which form will allow you to produce a graph most quickly? Why?

(c) You want to find an equation for the tangent line to a given curve at certain point on the curve. Which form will be most useful? Why?

(d) Which form is best for using a graphing device to graph the line?

2.4. Exercises.
1. Find all $x$- and $y$-intercepts of the following lines using the indicated method. Then graph the lines.

   (a) $3x - 7y = 5$ (by first converting to double intercept form)
   (b) $\frac{4}{3}x + \frac{7}{5}y = 2$ (by first converting to double intercept form)
   (c) $0x + 7y = 0$ (by any method)
   (d) $-4x + y = 1$ (by first converting to slope-intercept form)
   (e) $10x + 0y = 0$ (by any method)

2. Find an equation for each line as described, being sure to use each of the four special forms for an equation of a line at least once. Then, convert your equation to the standard form $Ax + By = C$.

   (a) The line with $x$-intercept $-5$ and $y$-intercept 7.
   (b) The line with slope 7 and $y$-intercept 2.
   (c) The line with slope 6 and $x$-intercept 4.
   (d) The line with slope 5 that passes through the point $(4, -7)$.
   (e) The line passing through the points $(4, 7)$ and $(\pi, 7)$.
   (f) The line passing through $(2, 7)$ that makes an angle of $60^\circ$ with the positive $x$-axis.

3. Suppose the line $L$ is given by the equation $Ax + By = C$. For each question below, answer both with a geometric description and with conditions on $A$, $B$, and $C$.

   (a) When does $L$ have a unique $x$-intercept?
   (b) When does $L$ have no $x$-intercept?
   (c) When does $L$ have infinitely many $x$-intercepts?

4. Repeat Exercise 3 for $y$-intercepts.

5. Suppose the line $L$ has equation $Ax + By = C$.

   (a) Give necessary and sufficient conditions on $A$, $B$, and $C$ for $L$ to have positive $x$- and $y$-intercepts.
   (b) Give necessary and sufficient conditions on $A$, $B$, and $C$ for the $x$- and $y$-intercepts of $L$ to be equal.
6. On the same axes, plot the lines $-3x + y = 3$, $-2x + y = 3$, $-x + y = 3$, $0x + y = 3$, $x + y = 3$, $2x + y = 3$, and $3x + y = 3$. What do these lines have in common? What would you expect the line $-40x + y = 3$ to look like?

7. Do the points $(1,7)$, $(12,30)$, and $(22,52)$ all lie on a single line? What is the easiest way to tell?

8. Consider the table of linear equations given in Section 2.3.

(a) Show how each of the equations in the table is obtained from the general equation $Ax + By = C$.

Then express $m, n, a, b$ in terms of $A, B, C$.

(b) Use your answer to part (a) to explain why $m$ represents slope, $n$ represents ‘reciprocal slope’, $a$ represents $x$-intercept, and $b$ represents $y$-intercept. (Hint: Proposition 2 will be helpful.)

9. Let $L$ be a line given by equation $Ax + By = C$. Define $n = -B/A$.

(a) Describe the set of lines for which $n$ is defined.

(b) If $(x_1, y_1)$ and $(x_2, y_2)$ are points on $L$, express $n$ in terms of $x_1, x_2, y_1, \text{and } y_2$.

10. On the same axes, plot the lines $x + 2y = 0$, $x + 2y = 1$, and $x + 2y = 2$. What is the distance between the lines? (Distance here means minimal distance, not ‘vertical distance’ or ‘horizontal distance’.)

11. Show that the point $(1,4)$ is on the line $2x + 3y = 14$. Find the two points on the same line whose distance to $(1,4)$ is 1.

12. Where in the proof of Proposition 2 does one need the fact that the line is nonvertical?

13. Prove that if the line $L$ is known to have slope $m$ and if $(x_0, y_0)$ is known to be a point on $L$, then $y - y_0 = m(x - x_0)$ is an equation for $L$.

14. Consider a line with equation $x/a + y/b = 1$. Find necessary and sufficient conditions on $a$ and $b$ for the line to bound a region with area 20 in the first quadrant.

15. Consider a line with equation $y = mx + b$. 
2. INTERCEPTS, SLOPE, AND EQUATIONS OF LINES

(a) Find necessary and sufficient conditions on $m$ and $b$ for the line to bound a region with area 20 in the first quadrant.

(b) Find all $m$ and $b$ such that the line bounds a region with area 20 in the first quadrant and the point $(2,3)$ lies on the line. Then answer the same question for area 12, and area 10 (still with $(2,3)$ on the line).

16. This exercise makes use of graphing transformations (see Chapter 1, Section 3).

(a) Starting with the graph of $y = x$, suppose that the graph is stretched vertically by a factor of $m$. What is the equation of the new line?

(b) Starting with the graph of $y = x$, suppose that the graph is shifted vertically by the amount $b$. What is the equation of the new line?

(c) Starting with the graph of $y = x$, suppose that the graph is stretched vertically by a factor of $m$ and then shifted vertically by the amount $b$. What is the equation of the new line?

(d) Starting with the graph of $y = x$, suppose that the graph is shifted vertically by the amount $b$ and then stretched vertically by the factor $m$. What is the equation of the new line?

(e) By what amount must the graph of $y = mx$ be shifted horizontally to produce the line with equation $y = mx + b$?

(f) By what angle must the horizontal line with equation $y = 0$ be rotated to obtain the line with equation $y = mx$?

17. If $f(x) = mx + b$ with $m \neq 0$,

(a) Is the inverse function $f^{-1}$ a linear function? If so, find the slope and $y$-intercept for the line corresponding to $f^{-1}$.

(b) Suppose $f(x) = 2x + 3$. Graph $f$ and $f^{-1}$ on the same axes.

18. Let $f(x) = mx + b$ and let $f^{(n)}$ denote the $n$-fold composition of $f$ with itself (e.g., $f^{(1)}(x) = f(x)$, $f^{(2)}(x) = f(f(x))$, etc.).

(a) Give formulas for $f^{(2)}(x)$ and $f^{(3)}(x)$. Are these linear functions?
(b) Deduce a concise formula for \( f^{(n)}(x) \). (Hint: You’ll need to compute the partial sum of a geometric series.)

(c) What happens to \( f^{(n)}(x) \) as \( n \to \infty \) when:
   (i) \( |m| < 1 \)?
   (ii) \( m = \pm 1 \)?
   (iii) \( |m| > 1 \)?

(d) Illustrate your answer to part (i) above by choosing a specific function \( f \) and using mathematical software to graph \( f^{(n)} \) for \( 1 \leq n \leq 50 \).

19. Let \( L \) be a line with equation \( Ax + By = C \). Prove that \( A, B, \) and \( C \) are all nonzero if and only if the \( x \)-intercept exists and is unique, the \( y \)-intercept exists and is unique, and neither is zero.

20. Let \( f \) be a real-valued function defined on the real numbers. Use calculus to prove that \( f(x) = mx + b \) for some constants \( m \) and \( b \) if and only if the quotient \( \frac{y_2 - y_1}{x_2 - x_1} \) is constant for any two points \((x_1, y_1), (x_2, y_2)\) on the graph of \( f \). (Hint: Mean value theorem.)

3. Solving Systems of Linear Equations

Here we discuss methods (elimination and substitution) for solving systems of equations and discover a geometric interpretation of the elimination method.

3.1. Elimination. We begin with the simplest intersection problem—that of finding the intersection of a vertical and a horizontal line:

**Let’s Go 2. Graphs the lines \( x = 2 \) and \( y = 5 \).**

(a) Find the intersection of these two lines. Why is this easy?

(b) Now graph the lines \( 2x + 7y = 9 \) and \( 5x - 10y = 3 \) and find their intersection. Why is this harder than part (a)?

The point of intersection of two intersecting lines is not usually obvious from the given equations, but we know several methods for finding points of intersection, including the method of elimination. For example, in the case of \( 2x + 7y = 9 \) and \( 5x - 10y = 3 \), elimination yields the following results.
Further, the succession of systems in (1) may be viewed geometrically in Figure 2.

**Figure 2.** Geometric view of the elimination procedure

**Your Turn** 8. We learn in linear algebra that elimination is essentially the process of successively trading in a system of equations for a simpler equivalent system, and that the process ends when we can ‘read off’ the solution. Using Figure 2, describe the process of elimination from a geometric standpoint.

**3.2. Substitution.** In addition to the elimination method of solving linear systems of equations, you are also familiar with the method of substitution. For instance, to solve the system

\[
\begin{align*}
2x + 7y &= 9, \\
5x - 10y &= 3,
\end{align*}
\]

one might use the second equation to determine that \( x = 2y + \frac{3}{5} \), and then substitute this information into the first equation to obtain \( 2(2y + \frac{3}{5}) + 7y = 9 \), and hence \( y = \frac{39}{55} \). To finish, we use this value of \( y \) in the equation \( x = 2y + \frac{3}{5} \) to obtain \( x = \frac{111}{55} \).

---

\(^2\)Here’s some terminology you may have forgotten: A *system* of equations is a collection of equations. A *solution* to a system of equations must satisfy every equation in the system. Two systems are *equivalent* if they possess the same solutions.
Your Turn 9. **Elimination and substitution are both valid methods for solving linear systems of equations.**

(a) **Construct two systems of equations where the elimination method seems preferable.** Then, **construct two systems of equations where the substitution method seems preferable.** Give explanations for your choices, including an explanation of what you mean by ‘preferable.’

(b) **Generally speaking, do you find one method preferable to the other? This is an important pedagogical question: As a teacher, you may need to choose one method to emphasize.**

3.3. **Exercises.**

1. Is the system \( \{2x + 7y = 9, 5x - 10y = 3\} \) equivalent to the system \( \{55y = 39\} \)? Explain.

2. Solve the system \( \{3x - 5y = 4, 2x + 8y = 15\} \) by elimination. Be sure to write down a succession of equivalent systems of equations, and plot the lines corresponding to each system (see Figure 2).

3. Another common way to solve a system of two linear equations is to solve both equations for \( y \), and equate the two expressions. Use this technique to solve the system \( \{3x - 5y = 4, 2x + 8y = 15\} \).

4. Not all systems of equations possess just one single solution. To investigate this phenomenon, construct an example of a system of two linear equations that:

   (a) has no solution.

   (b) has infinitely many solutions.

5. **(This is a continuation of Exercise 4.)** Show that if a system of two equations (in two unknowns) does not possess infinitely many solutions, then it possesses either precisely one solution or it possesses no solution at all.

6. Given the system of linear equations \( \{Ax + By = C, A'x + B'y = C'\} \), suppose that \( AB' \neq BA' \). **Prove that the system has exactly one solution** by finding an explicit formula for \( x \) and \( y \) in terms of \( A, B, C, A', B', C' \). **Use your answer to find the solution** to the system \( \{3x - 5y = 4, 2x + 8y = 15\} \).
4. Parameterized Lines

In Section 1.1, we investigated how a geometric interpretation of a line (as a curve in the plane determined by two points) is closely linked to an algebraic interpretation as solution sets for $Ax + By = C$. In this section, we present a physical interpretation of lines in the plane that leads to a very different algebraic interpretation known as parameterizations of lines\(^3\).

4.1. Whirlwind review of vectors. Since our discussion relies on vectors in the plane, we pause for a refresher on vectors.

A vector in the plane is an equivalence class of directed line segments, where two directed segments are equivalent if they have the same length and direction (see Figure 3 and Exercise 15).

Of course, when we want to draw a vector in the plane, we can't draw all of the directed segments with the same length and direction. Instead, we choose a representative. The representative most often chosen is the one which begins at the origin, and is called the standard representative (see Figure 4). Also indicated in Figure 4 is the fact that endpoints of standard representatives provide a means of identifying vectors with

---

\(^3\)The idea of parameterization was discussed in Chapter 1, Section 7.1.
ordered pairs. Thus, when we refer to ‘the vector \((a,b)\)’ we mean the vector whose standard representative ends at the point \((a,b)\).

**Your Turn 10.** Consider the directed segment beginning at \((x_1,y_1)\) and ending at \((x_2,y_2)\).

(a) Draw a figure illustrating the original segment and the corresponding standard representative.

(b) Give a geometric description of how one obtains the standard representative from the original segment.

(c) List the coordinates of the endpoint of the standard representative.

We may perform arithmetic with vectors. Specifically:

**Definition 3.** Let \((a,b)\), \((c,d)\) be vectors and \(r\) a real number.

(a) The sum of \((a,b)\) and \((c,d)\) is the vector \((a + c, b + d)\).

(b) The difference of \((a,b)\) and \((c,d)\) is the vector \((a - c, b - d)\).

(c) Scaling the vector \((a,b)\) by a factor of \(r\) gives the vector \((ra,rb)\). (We often call \((ra,rb)\) a scalar multiple of \((a,b)\).)

(d) The length of \((a,b)\), denoted \(||(a,b)||\), is the real number \(\sqrt{a^2 + b^2}\).

All of the items of Definition 3 have important geometric/physical interpretations\(^4\), as shown in Figure 5.

\(^4\)The question of whether Definition 3 is well defined will not be addressed.
4.2. Moving with constant velocity: Parameterizations of lines. In describing parameterizations of lines (which model motion in a fixed direction) it will be helpful to have the notion of a direction vector for a line. To obtain a direction vector for a line, we pick two distinct points \((x_1, y_1)\) and \((x_2, y_2)\) on the line, and form the vector \((x_2 - x_1, y_2 - y_1)\). Any two direction vectors for a given line are parallel; that is, each direction vector is a scalar multiple of the other. (See Exercise 13 for an algebraic proof of this fact.)

Now we are ready to discuss parameterizations of lines. Physically, we may think of a line as the path of an object moving in the plane at constant velocity (i.e., constant direction of motion and constant speed). Using the language of vectors, we can produce an algebraic representation of this ‘physical’ approach to lines. Let \(p(t)\) denote the location of the object at time \(t\). To derive an algebraic expression for \(p(t)\), there are two important vectors to consider. The first is \(p(0)\), the position of the object at time \(t = 0\). The second is the vector \(v = p(1) - p(0)\). We should regard \(v\) as the velocity vector for the object, since

- \(v\) is a direction vector for the line, pointing in the direction of motion as \(t\) increases, and
- the length of \(v\) is the speed of the object. (After all, the length of \(v\) is precisely the distance the object travels in one unit of time.)

With this in mind, the position of the particle at time \(t\) is obtained by starting at \(p(0)\) and adding \(tv\). For example, the point \(p(3)\) is obtained by starting at \(p(0)\) and adding \(3v\) (see Figure 6). Therefore

\[ p(t) = p(0) + tv. \]  

The equation given in (2) is called a parametric equation for our line, and \(t\) is called the parameter.

\(5\)Parameterized lines with a nonuniform speed are discussed in the exercises.
One can also write the parametric equation in a way that emphasizes $x$ and $y$ coordinates. If $p(0) = (d, f)$ and $v = (c, e)$, then we have $p(t) = (d, f) + t(c, e) = (d + tc, f + te)$, which we might write as

$$x = d + tc \quad y = f + te$$

We have simply identified the $x$- and $y$-coordinates of the vector $p(t)$ as functions of $t$.

To summarize in the context of a specific example, consider a parametric equation for a line satisfying $p(0) = (4, -7)$ and $p(1) = (3, 2)$. We have that $v = (3, 2) - (4, -7) = (-1, 9)$. Thus the parametric equation for the line is $p(t) = (4, -7) + t(-1, 9)$. Alternatively, we could write $x = 4 - t$, $y = -7 + 9t$. The graph of this line is given in Figure 7.
4. PARAMETERIZED LINES

Your Turn 11. Look up the words ‘parameter’ and ‘parametric’ in the dictionary. How are these definitions related to parametric equations?

Your Turn 12. Suppose a path traced out by an object moving with constant velocity in the plane is described by both a parametric equation \( p(t) = p(0) + tv \) and a nontrivial linear equation \( Ax + By = C \). What important information about the object (if any) is provided by the parametric equation which is not provided by the linear equation? (Cryptic hint: Parametric equations not only describe paths but also describe journeys).

4.3. Different parameterizations for the same line. The parameterization of a line in the plane is not at all unique. Many choices must be made in attaching a parameterization to a given line with parameter \( t \). We must pick the point on the line that will be the ‘starting point’ (that is, the point on the line corresponding to \( t = 0 \)). This already involves a choice with infinitely many possible outcomes, since any point on the line can correspond to \( t = 0 \). Also, we need to pick a velocity vector for the line. This vector will point in the direction of motion along the line as \( t \) increases, and the length of the vector will determine the speed of motion. Here we also have infinitely many choices (two choices for direction, infinitely many for length). Figure 8 illustrates three different parameterizations for the same line.
Figure 9. Are we doomed?

**Your Turn 13.** Consider the line with the parameterization \( p(t) = (2, 4) + t(3, -2) \).

(i) Find a parameterization for this same line but with a different starting point, a different direction of motion as the parameter increases, and a different speed.

(ii) Explain why \( p(t) = (2, 4) + t(3, -2) \) and your parameterization from part (i) actually produce the same line in the plane.

### 4.4. Intersecting parameterized lines.

The following Your Turn introduces us to intersection:

**Your Turn 14.** Figure 9 shows a small piece of Earth’s orbit together with that of Comet Armadillo.

(i) The orbits are drawn as line segments. Is this reasonable, considering the real shape of these orbits?

(ii) The figure shows that the orbits intersect. Should we send up a spaceship to blow up Comet Armadillo and prevent a collision with Earth?

(iii) How might parametric descriptions of the lines in Figure 9 help in answering part (ii)?

We have already discussed methods for finding the intersection of two lines, but what if the lines are expressed in parametric form? For instance, suppose we have lines with parameterizations

\[
p(t) = (3, 4) + t(5, 6) \quad \text{and} \quad q(s) = (2, -9) + s(3, -8).
\]

At a point of intersection, \( x \)- and \( y \)-coordinates must agree, so we must have \( 3 + 5t = 2 + 3s \) and \( 4 + 6t = -9 - 8s \).

This leads to the system

\[
\begin{align*}
5t - 3s &= -1 \\
6t + 8s &= -13.
\end{align*}
\]
Solving this system yields \( s = -\frac{59}{58} \) and \( t = -\frac{47}{58} \). In fact, knowing either \( t \) or \( s \) is alone enough to find the point of intersection. For example, using \( t = -\frac{47}{58} \), we have that the point of intersection is\
\[
 p(-47/58) = (3, 4) - \frac{47}{58} (5, 6) = (-61/58, -50/58).
\]
We would obtain the same point by computing \( q(-59/58) \).

**Your Turn** 15. Consider the example given above, in which we found the intersection of the lines \( p(t) = (3, 4) + t(5, 6) \) and \( q(s) = (2, -9) + s(3, -8) \).

(a) To find a point of intersection, in equating \( x \)- and \( y \)-coordinates why is it not sufficient to consider the equations \( 3 + 5t = 2 + 3t \) and \( 4 + 6t = -9 - 8t \)?

(b) What question is being answered by simultaneously solving the equations \( 3 + 5t = 2 + 3t \) and \( 4 + 6t = -9 - 8t \) for \( t \)? (Is there a solution in this case?)

**Your Turn** 16. In finding the point of intersection for the lines \( p(t) = (-2, -1) + t(1, 1) \) and \( q(s) = (5, 3) + s(4, 1) \), the solutions manual for your students’ textbook says the following:

Equate \( x \)- and \( y \)-coordinates for each line gives \(-2 + t = 5 + 4s \) and \(-1 + t = 3 + s \), giving rise to the system
\[
\begin{align*}
 t - 4s &= 7 \\
 t - s &= 4.
\end{align*}
\]
Solving the system gives \( t = 3 \) and \( s = -1 \), so the solution is \( (3, -1) \).

Are there any mistakes? Explain.

**4.5. Converting between parametric and non-parametric equations.** We have two different descriptions of lines in the plane: as solutions to linear equations in two variables \((Ax + By = C)\) and also as points corresponding to parametric equations \((p(t) = p(0) + tv)\). If our goal involves expressing \( y \) as a function of \( x \), then the first setting is preferable. If we want to describe motion in the plane, we prefer the second. It is useful to know how to translate between these two descriptions.

Beginning with a parametric equation, say \( p(t) = (2 + 3t, 4 - 2t) \), how do we find a standard equation in the form \( Ax + By = C \)? Perhaps the simplest approach is to use the parametric equation to find two points
on the line, say \( p(0) = (2, 4) \) and \( p(1) = (5, 2) \), then determine that the slope is \( \frac{2 - 4}{5 - 2} = -\frac{2}{3} \), and finally apply the point-slope form to obtain \( y - 4 = -\frac{2}{3}(x - 2) \).

Here is another approach: We are given \( x = 2 + 3t \) and \( y = 4 - 2t \). We must find constants \( A, B, \) and \( C \) such that \( Ax + By = C \) for all points along the line. Now \( C \) is a constant, so \( Ax + By \) must also be constant for all points \((x, y)\) on the line. This means that \( t \) must not appear in \( Ax + By \), after we substitute \( x = 2 + 3t \) and \( y = 4 - 2t \). To “get the \( t \)’s to cancel,” we can take \( A = 2 \) and \( B = 3 \). Then \( Ax + By = 2(2 + 3t) + 3(4 - 2t) = 16 + 0t = 16 \), so points on the line satisfy \( 2x + 3y = 16 \), and this is an equation of the line. This method is known as eliminating the parameter.

In a third approach, we can obtain a standard equation of the line \( p(t) = (2 + 3t, 4 - 2t) \) from its slope-intercept form. The slope of the line is \(-\frac{2}{3}\) (why?), and to find the \( y \)-intercept, we check that \( x = 0 \) exactly when \( 2 + 3t = 0 \), or \( t = -\frac{2}{3} \). So the \( y \)-intercept is \( 4 - 2\left(-\frac{2}{3}\right) = \frac{16}{3} \). Hence an equation of the line is

\[
y = -\frac{2}{3}x + \frac{16}{3} \Leftrightarrow 2x + 3y = 16
\]

Your Turn 17. Consider the process of converting from parametric to standard form.

(a) Generalize the first approach given above by finding a standard equation corresponding to the parametric equation \( p(t) = (d + tc, f + te) \). Describe how one is to find \( A, B, \) and \( C \).

(b) Are there any kinds of parameterized lines for which the previous strategy will fail to produce a result? Explain.

Conversely, given an equation for a line (in \( Ax + By = C \) form), how do we find a parameterization of the line? To begin, we should recall that a single line has many different parameterizations. For now, let us content ourselves with finding just one parameterization. Suppose the line has equation \( 3x + 4y = 10 \). We can simply use \( x \) itself as a parameter. Since \( y = -3/4 x + 10/4 \), we have \( x = t \) and \( y = -3/4 t + 10/4 \), and thus the parameterization is \( p(t) = (t, -3/4 t + 10/4) \). For confirmation, note that this implies that \((0,10/4)\) is a point on the line, and the vector \( p(1) - p(0) = (1, -3/4) \) points in the direction of the line. Both of these facts are consistent with the original equation \( 3x + 4y = 10 \).
4.6. Exercises.

1. Construct a word problem whose solution is best represented by a parametric description for a line. Also, construct a word problem whose solution is best represented by a linear equation of the form \(Ax + By = C\).

2. Graph the line \(2x + 3y = 1\). Find two points on the line, compute the direction vector determined by these two points, and plot the direction vector.

3. For the line with parameterization \(x = 4 + 9t, y = -2 + 3t\), find \(p(0)\) and \(v\).

4. Find the parametric equation for the line with:
   (i) \(p(0) = (-2, -7)\) and \(p(1) = (3, -9)\).
   (ii) \(p(3) = (4, 2)\) and \(p(10) = (18, 10)\).

5. For a moving object whose position is given by \(p(t) = p(0) + tv\), verify that the average velocity of the object over any time interval \([t_1, t_2]\) is indeed constant.

6. Which of the following pairs of parametric equations actually determine the same line in the plane?
   (i) \(p(t) = (2, 5) + t(3, 8)\) and \(q(t) = (3, 1) + t(2, 7)\)
   (ii) \(p(t) = (2, 5) + t(3, 8)\) and \(q(t) = (8, 21) + t(9, 24)\)
   (iii) \(p(t) = (2, 5) + t(3, 8)\) and \(q(t) = (8, 20) + t(9, 24)\)

7. A fox has position \(x = -4t + 9, y = 8t - 9\) and a rabbit has position \(x = 3t + 2, y = 3t + 3\). At what point do their paths cross, and who reaches that point sooner?

8. Find a standard equation for the line with parametric equation \(p(t) = (-7 + 5t, 2 + 11t)\) using two different methods.

9. For the line \(p(t) = (ct + d, ct + f)\), assume that neither \(c\) nor \(e\) is zero. Find the \(x\)- and \(y\)-intercepts of the line, and deduce the double-intercept form for the equation of the line.
10. Find two parameterizations of the line $7x - 3y = 6$, first using $x$ as the parameter and then using $y$.

11. Find a parameterization of the horizontal line $y = 5$. Find a parameterization of the vertical line $x = -2$.

12. Find the point of intersection of the line with equation $3x + 6y = 2$ and the line with parametric equation $\mathbf{p}(t) = (2, 7) + t(1, 4)$.

13. This exercise explores parallelism.
   (a) Verify that two nonzero vectors $(A, B)$ and $(A', B')$ are parallel (i.e., each vector is a scalar multiple of the other) if and only if $AB' = BA'$.
   (b) Prove that any two direction vectors for a given line are parallel.

14. Prove that parallelism of nonzero vectors in the plane is an equivalence relation.

15. Formulate an equivalence relation on the set of directed line segments so that two segments with the same length and direction are equivalent. Verify that your relation is an equivalence relation. (Hint: How should the standard representatives of two equivalent segments compare?)

16. We know that $\mathbf{p}(t) = (1, 1) + t(1, 2)$ describes a line, but
   (i) What geometric object is described by $\mathbf{q}(t) = (1, 1) + t^2(1, 2)$? By $\mathbf{r}(t) = (1, 1) + \sin t(1, 2)$?
   (ii) Write a short story about the journey described by $\mathbf{r}(t) = (1, 1) + \sin t(1, 2)$ with $0 \leq t \leq \pi$.
   (iii) Find a cubic polynomial function $f$ so that $(1, 1) + f(t)(1, 2)$ describes the following journey for $0 \leq t \leq 2$:
      I begin at $(1, 1)$ and travel toward $(2, 3)$. I slow down and stop at time $t = 1$ at the point $(2, 3)$, and then continue on to the point $(3, 5)$ at time $t = 2$.

The following two problems investigate the medians and centroid of a triangle.

17. Consider a triangle with vertices $(x_1, y_1)$, $(x_2, y_2)$, and $(x_3, y_3)$. We consider a line containing one vertex of the triangle and also the midpoint of the opposite side of the triangle. The portion of the line within the solid triangle is called a median of the triangle.
(1) Using the parameter $t_1$, parameterize the line $L_1$ containing $(x_1, y_1)$ and the midpoint of the side containing $(x_2, y_2)$ and $(x_3, y_3)$. Arrange the parameterization so that $t_1 = 0$ gives $(x_1, y_1)$ and $t_1 = 1$ gives the midpoint of the side containing $(x_2, y_2)$ and $(x_3, y_3)$. (The corresponding median consists of all points on the line where $0 \leq t_1 \leq 1$.)

(2) In a similar fashion, parameterize the other two medians of the triangle.

(3) Show that the three medians intersect at a single point (called the centroid of the triangle). Find the coordinates of this point in terms of the coordinates of the vertices. What values of $t_1$, $t_2$, and $t_3$ occur at the centroid?

(4) Along a median, find the ratio $\frac{\text{distance from centroid to vertex}}{\text{distance from centroid to midpoint of opposite side}}$.

(5) The three medians of a triangle split the solid triangle into six small triangles. Show that all six triangles have equal area (one-sixth of the area of the original triangle).

Let $T$ be the triangle with vertices $(x_1, y_1)$, $(x_2, y_2)$, and $(x_3, y_3)$. Let $T^{(1)}$ be the triangle whose vertices are the midpoints of the sides of $T$, let $T^{(2)}$ be the triangle whose vertices are the midpoints of the sides of $T^{(1)}$, etc. Using coordinates, show that $T$ and $T^{(2)}$ have the same median. Conclude that $T \cap T^{(2)} \cap T^{(4)} \cap \ldots$ is the median of $T$.

5. Parallel Lines, Perpendicular Lines, and Distance

In this section, we investigate the notion of perpendicularity and use it to investigate certain familiar formulas involving the slopes of parallel and perpendicular lines and to determine the distance between two lines.

Let’s Go 3. The following questions concern parallel and perpendicular lines.

(a) Given equations for two lines, how can you tell whether the lines are parallel? Whether the lines are perpendicular? Give as many conditions as possible.

(b) Create equations for a pair of lines that are:

(i) parallel, nonvertical, nonhorizontal, and do not pass through the origin.

(ii) perpendicular, nonvertical, nonhorizontal, and do not pass through the origin.
(c) How might one see intuitively that the slopes of perpendicular lines, neither of which are vertical, are ‘opposite reciprocals’ of one another? (Try drawing a few pictures.)

5.1. Perpendicular vectors. Two nonzero vectors are said to be perpendicular if their standard representatives form a right angle at the origin. This description of perpendicularity is intuitive and easily pictured (see Figure 10). However, there is one rather large problem: This ‘right angle’ description is hard to verify visually. After all, an angle which appears to be 90° may only measure 89°. What we need is an algebraic definition of perpendicularity which both captures the geometric description given above and is easily verified:

**Definition 4.** Two vectors \((A, B)\) and \((A', B')\) are said to be perpendicular if \(AA' + BB' = 0\).

**Your Turn 18.** Use Definition 4 to find two vectors perpendicular to \((5, 3)\).

We want to be sure that Definition 4 captures the geometric essence of perpendicularity illustrated in Figure 10. According to the Pythagorean theorem and its converse,

\[
AA' + BB' = 0 \iff A^2 + B^2 + (A')^2 + (B')^2 = (A - A')^2 + (B - B')^2)
\]

\[\iff \left(\sqrt{A^2 + B^2}\right)^2 + \left(\sqrt{(A')^2 + (B')^2}\right)^2 = \left(\sqrt{(A - A')^2 + (B - B')^2}\right)^2\]

\[\iff \text{The triangle implied by Figure 10 is a right triangle.}\]

---

\[^6\text{You may recognize } AA' + BB' \text{ as the dot product of the two vectors in question.}\]
5.2. “Perpendicular form” for the equation of a line. Perpendicularity gives another perspective on linear equations. Suppose, for example, that we want to find the equation of the line that is perpendicular to the vector \((4, 7)\) and contains the point \((1, 5)\). Let \((x, y)\) be a point on this line. The vector with representative beginning at \((1, 5)\) and ending at \((x, y)\) is given by \((x - 1, y - 5)\) (See Figure 11). This vector is a direction vector for the line, so it is perpendicular to the vector \((4, 7)\). Therefore, appealing to Definition 4, we have \(0 = 4(x - 1) + 7(y - 5)\). Simplifying, we conclude that \(4x + 7y = 39\) is an equation for the line.

Conversely, one can see, for example, that the vector \((5, 6)\) is perpendicular to any direction vector for the line \(5x + 6y = 2\) (see Exercise 3). More generally,

Proposition 5. The vector \((A, B)\) gives a perpendicular vector to the line \(Ax + By = C\).

Figure 11. Lines and perpendicular vectors

Your Turn 19. For those of you familiar with equations of planes in three-space: How is the concept of determining a line from a point on the line and a perpendicular vector similar to the concept of determining an equation of a plane in three-space? Discuss.

5.3. Parallel lines, perpendicular lines, and familiar formulas. We define lines \(L\) and \(L'\) to be parallel if they have direction vectors that are parallel\(^7\). The lines will be perpendicular if their direction vectors are perpendicular.

\(^7\)Recall that two vectors \((A, B)\) and \((A', B')\) are parallel if one is a nonzero scalar multiple of the other. Equivalently, Exercise 13 says that the vectors are parallel if and only if \(AB' = BA'\).
Your Turn 20. An alternate definition of parallel lines states that two lines are parallel if they do not intersect in a single point. Compare this alternate definition of parallelism to to definition given above. Please include a figure.

Your Turn 21. Given parameterized lines $p(t) = p(0) + tv$ and $q(s) = q(0) + sw$, what conditions on $v$ and $w$ will insure that the lines are parallel? That the lines are perpendicular? Draw convincing figures illustrating your answers.

By defining perpendicularity of lines in terms of vectors (as we did above), we find that it is very easy to tell whether two lines are parallel or perpendicular, using only their equations (for proof details, see Exercise 19):

**Proposition 6.** The lines given by the equations $Ax + By = C$ and $A'x + B'y = C'$ are:

(i) parallel if and only if the vectors $(A, B)$ and $(A', B')$ are parallel.

(ii) perpendicular if and only if the vectors $(A, B)$ and $(A', B')$ are perpendicular.

Your Turn 22. You are grading the following proof of Proposition 6: “Since $(A, B)$ and $(A', B')$ are direction vectors for the lines $Ax + By = C$ and $A'x + B'y = C'$, respectively, the lines are parallel if and only if the vectors $(A, B)$ and $(A', B')$ are parallel.” Describe any mistakes in the proof.

Proposition 6 should vaguely remind us of some high-school formulas concerning parallel and perpendicular lines which we have all encountered at one time or another. For example, recall that:

- two non-vertical lines are parallel if and only if they have the same slope, and
- two non-vertical lines are perpendicular if and only if the product of their slopes is $-1$.

The first result is certainly believable—we expect lines with the same ‘steepness’ to be parallel (see Exercise 20). On the other hand, the second result is not immediately obvious, yet we can tackle it in short order using Proposition 6: Let $Ax + By = C$ and $A'x + B'y = C'$ be two non-vertical lines. According to Proposition 6, these lines are perpendicular if and only if $(A, B)$ is perpendicular to $(A', B')$, meaning that $AA' + BB' = 0$. Since $B$ and $B'$ are non-zero (why?), we have

$$AA' + BB' = 0 \iff AA' = -BB' \iff \frac{-A}{B} \cdot \frac{-A'}{B'} = -1,$$
5. PARALLEL LINES, PERPENDICULAR LINES, AND DISTANCE

5.4. The distance between parallel lines; the distance between a line and a point. In this section, we will derive a formula for the distance between two parallel lines, as well as a formula for the distance between a line and a point. These formulas become useful when we approximate certain data sets by lines (see Section 6 of this chapter).

To begin, suppose we have two parallel lines $L_1$ and $L_2$ given by equations $Ax + By = C_1$ and $Ax + By = C_2$, respectively. Consider the line $L_\perp$ through the origin that is perpendicular to $L_1$ and $L_2$. Since the vector $(A, B)$ is perpendicular to both lines, we have a parametric equation for $L_\perp$, simply $p(t) = (At, Bt)$ (Figure 12).

Next, we find the points where $L_\perp$ intersects $L_1$ and $L_2$. Beginning with $L_1$, we see that the point $p(t_1)$ is on $L_1$ exactly when it satisfies the equation for $L_1$, that is $C_1 = Ax + By = A(At_1) + B(Bt_1) = t_1(A^2 + B^2)$. Thus, the intersection point of $L_\perp$ with $L_1$ is the point on $L_\perp$ where the parameter $t_1$ equals $C_1/(A^2 + B^2)$. Likewise, $L_\perp$ and $L_2$ intersect at the point where the parameter $t_2$ equals $C_2/(A^2 + B^2)$.

Finally, we see that the ‘distance’ between $L_1$ and $L_2$ (essentially a ‘minimum’ distance between the lines) is simply the distance between these two points of intersection:
\[
\text{dist}(L_1, L_2) = \text{dist} \left( t_1(A, B), t_2(A, B) \right) \\
= |t_1 - t_2| ||(A, B)|| \\
= \frac{|C_1 - C_2|}{\sqrt{A^2 + B^2}} \sqrt{A^2 + B^2} \\
= \frac{|C_1 - C_2|}{\sqrt{A^2 + B^2}}.
\]

We have just proved:

**Proposition 7.** The distance between the parallel lines given by the equations \(Ax + By = C_1\) and \(Ax + By = C_2\) is \(\frac{|C_1 - C_2|}{\sqrt{A^2 + B^2}}\).

**Your Turn 23.** Use the Pythagorean Theorem together with an illustrative drawing to produce a convincing argument that the segment of \(L_\perp\) joining lines \(L_1\) and \(L_2\) is the shortest possible segment joining the two parallel lines.

We can use Proposition 7 to find the (minimum) distance between a point \((x_0, y_0)\) and a line \(L\) given by \(Ax + By = C\). The parallel line that contains \((x_0, y_0)\) is \(Ax + By = C'\), where \(C' = Ax_0 + By_0\). Thus

**Proposition 8.** The distance between a point \((x_0, y_0)\) and a line \(L\) with equation \(Ax + By = C\) is

\[
\text{dist} \left( (x_0, y_0), L \right) = \frac{|C - (Ax_0 + By_0)|}{\sqrt{A^2 + B^2}}.
\]

**Your Turn 24.** There may be many perfectly valid ways to measure the distance between two objects. Suggest two alternate ways to measure the distance between a point and a line, and explain how they might be useful.

5.5. Exercises.

1. Graph the lines \(2x + 3y = 1\) and \(4x + 6y = 5\). Then, find a direction vector for each line (by picking a pair of points on each line and subtracting), and verify algebraically that the two direction vectors are parallel.
2. Find three vectors perpendicular to the vector \((-2, 5)\). Plot them in the \(xy\)-plane.

3. Find a direction vector for the line \(5x + 6y = 2\). Show that the direction vector is perpendicular to the vector \((5, 6)\). Draw a figure illustrating the situation.

4. Find the equation of the line containing the point \((4, -2)\) and perpendicular to the vector \((5, -3)\).
Then find the equation of the line containing the vector \((4, -2)\) and parallel to the vector \((5, -3)\).

5. For each of the following lines, find a vector parallel to the line and a vector perpendicular to the line.
   
   (a) \(-4x + 7y = 10\)
   (b) \(y = 3x - 8\)

6. Find an equation for the line \(L\) satisfying the given conditions.
   
   (a) \(L\) is perpendicular to the line \(y = \frac{1}{2}x + \pi\) and passes through the origin.
   (b) \(L\) intersects the line \(2x + 5y = 9\) at the point \((2, 1)\) at right angles.
   (c) \(L\) is perpendicular to the segment connecting \((2, 5)\) and \((4, -1)\) and contains the midpoint of the segment.

7. Suppose the line \(L\) has \(x\)-intercept \(a \neq 0\) and \(y\)-intercept \(b \neq 0\).
   
   (a) Find an equation of the line through the origin that is perpendicular to \(L\).
   (b) Find an equation of the line perpendicular to \(L\) that has the same \(x\)-intercept as \(L\).
   (c) Find an equation of the line that is perpendicular to \(L\) and intersects \(L\) at the midpoint of the segment joining the two intercepts of \(L\).
   (d) Prove that the line \(L'\) with \(x\)-intercept \(b\) and \(y\)-intercept \(-a\) is perpendicular to \(L\).

8. Find two parametric equations \(\mathbf{q}(t) = \mathbf{q}(0) + t\mathbf{v}\) each satisfying both of the following conditions:
   
   (a) The resulting line is perpendicular to the line \(\mathbf{p}(t) = (3, 4) + t(5, 6)\),
   (b) The parametric equation describes the motion of an object that is located at \((8, -1)\) at \(t = 0\), and has the same speed as the object described by \(\mathbf{p}(t) = (3, 4) + t(5, 6)\).
9. Find the parametric equations for the two parameterized lines that are parallel to the parameterized line \( p(t) = (3, 4) + t(5, 6) \), pass through the point \((8, -1)\) at time 0, and have the same speed as \( p(t) \).

10. Find the distance between:

   (a) the parallel lines \( 4x - 2y = 20 \) and \( 4x - 2y = -10 \).

   (b) the parallel lines \( y = 5x + 2 \) and \( y = 5x + 4 \).

   (c) the parallel lines \( p(t) = (3 + 4t, 5 + 6t) \) and \( q(t) = (-2 + 2t, 45 + 3t) \).

   (d) the point \((3, -7)\) and the line \( 2x - 5y = 30 \).

11. Find equations for:

   (a) the two lines with slope \( \frac{5}{3} \) that are of distance 2 from the point \((8, -4)\).

   (b) the two lines parallel to the line \( 2x + 7y = 3 \) that have distance 12 from this line.

   (c) the line parallel to \( 3x - 7y = 4 \) which has the same distance to \((2, 6)\) as does \( 3x - 7y = 4 \).

12. Let \( S \) be the line segment \( \{(x, 0) : 0 \leq x \leq 1\} \). Describe the collection of all points in the plane whose distance to \( S \) is \( \frac{1}{2} \). Do the same for the line segment with endpoints \((0, 0)\) and \((2, 5)\).

13. Consider \( p(t) = (ct + d, et + f) \). Values for \( c, d, e, f \) give important information about the line parameterized by \( p \):

   (i) Find a vector parallel to the line and a vector perpendicular to the line.

   (ii) Express the slope of the line in terms of \( c, d, e, f \).

   (iii) Find necessary and sufficient conditions on \( c, d, e, f \) for the line to be vertical, and conditions for the line to be horizontal.

   (iv) Find necessary and sufficient conditions on \( c, d, e, f \) for the line to be the \( x \)-axis or the \( y \)-axis.

   (v) Find the \( x \)-intercept of the line. (When does it not exist, and when is it not unique?) Do the same for the \( y \)-intercept.

   (vi) Find necessary and sufficient conditions on \( c, d, e, f \) for the line to pass through the origin.
5. PARALLEL LINES, PERPENDICULAR LINES, AND DISTANCE

Figure 13. Proving the Pythagorean Theorem

14. The Pythagorean theorem\(^8\) has been a key tool throughout this section. Carefully draft a statement of the Pythagorean theorem. Then, use Figure 13 to prove the Pythagorean Theorem.

15. If two lines are parallel, what can their intersection be? State and prove your answer.

16. Let \(a\) and \(b\) be positive. The line \(x/a + y/b = 1\) bounds a triangular region in the first quadrant. Find the equation of the parallel line that bounds a triangle of half the area.

17. Is perpendicularity of vectors an equivalence relation? Prove or disprove.

18. Prove Proposition 5. Be sure to draw an illustrative figure to accompany your proof. (Hint: Show that any direction vector determined by two points on the line is perpendicular to \((A, B)\).

19. Here is a proof for part (i) of Proposition 6: “The lines are parallel if and only if the vectors \((-B, A)\) and \((-B', A')\) are parallel, and the vectors \((-B, A)\) and \((-B', A')\) are parallel if and only if \((A, B)\) and \((A', B')\) are parallel.” Fill in the details by showing that

(a) the vector \((-B, A)\) is a direction vector for the line \(Ax + By = C\), and

(b) the vectors \((-B, A)\) and \((-B', A')\) are parallel if and only if \((A, B)\) and \((A', B')\) are parallel.

20. Use Proposition 6 to prove that two non-vertical lines are parallel if and only if they have the same slopes.

---

\(^8\)There are many, many proofs for the Pythagorean theorem: some from antiquity (e.g., Euclid), and some from modern times and unlikely sources (e.g., President James Garfield). It is not known whether Pythagorus (Greek, ca 600 b.c.e.) actually proved the theorem that bears his name.
21. Suppose the line \( L \) has equation \( Ax + By = C \). Find the point on \( L \) closest to the origin.

22. *Altitudes and orthocenter of a triangle.* (Note: Computer algebra software may be helpful in this exercise.) Consider a triangle with vertices \((x_1, y_1), (x_2, y_2), \text{ and } (x_3, y_3)\). An *altitude of the triangle* is a line passing through a vertex that is also perpendicular to the opposite side. A triangle has three altitudes, one for each side.

   (1) Find a vector parallel to the side opposite the vertex \((x_1, y_1)\). Use this and the “perpendicular form of the equation of a line” to find the equation of the altitude through \((x_1, y_1)\).

   (2) Write down the equations of the other two altitudes.

   (3) Show that the three altitudes meet at one point. Find the coordinates of this point (called the *orthocenter of the triangle*).

23. *Perpendicular bisectors and circumcenter of a triangle.* (Note: Computer algebra software may be helpful in this exercise.) Consider a triangle with vertices \((x_1, y_1), (x_2, y_2), \text{ and } (x_3, y_3)\). A *perpendicular bisector of a side* is a line passing through the midpoint of the side that is also perpendicular to the side. A triangle has three perpendicular bisectors, one for each side.

   (1) Find the midpoint of the side containing \((x_1, y_1)\) and \((x_2, y_2)\). Find a vector parallel to the side. Then use the “perpendicular form of the equation of a line” to find the equation of the perpendicular bisector.

   (2) Write down the equations of the perpendicular bisectors of the other two sides.

   (3) Show that the three perpendicular bisectors meet at one point. Find the coordinates of this point (called the *circumcenter of the triangle*).

   (4) By computation, show that the vertices of the triangle all have the same distance to the circumcenter. (This shows that the circumcenter is the center of a circle that contains the three vertices.)

   (5) Show that the distance from a vertex to the circumcenter is greater than or equal to the distance from any point on any side of the triangle to the circumcenter. (This shows that the circle circumscribes the triangle.)
24. Let $T$ be the triangle with vertices $(x_1, y_1)$, $(x_2, y_2)$, and $(x_3, y_3)$. Let $T'$ be the triangle whose vertices are the midpoints of the sides of $T$. Let $T''$ be the triangle whose vertices are the midpoints of the sides of $T'$.

(1) Using coordinates, show that each side of $T'$ is parallel to a side of $T$.

(2) Show that each altitude of $T'$ is a perpendicular bisector of $T$. (Hence the circumcenter of $T$ is the orthocenter of $T'$.)

(3) Explain why the circumcenter of $T'$ is the orthocenter of $T''$.

(4) Find the vertices of a triangle $S$ such that the midpoints of $S$ give $T$.

6. An Application of Distance: Using Lines to Fit Data

Mathematicians often use simple, easily understandable objects to model complicated phenomena. We know that lines are the simplest plane curves, so it is no surprise that lines are often called to service in the modeling business.

One garden-variety way that lines are used in modeling is to ‘fit’ discrete sets of points in the plane. For example, consider Figure 14, showing a set of points in the plane. By ‘fitting a line’ to this data, we mean finding a single line which in some sense comes as close as possible to all the points in our data set. (Reasons for wanting to do this are numerous, and some will be indicated below.) Of course, in order to

Figure 14. Some points in the plane

---

9 This method crops up all the time in calculus, where, for example, one uses rectangles (simple objects) to approximate the area of various regions. You should try to think of several other similar examples from calculus.

10 One imprecise (but surprisingly effective) way to come up with a line of best fit is simply to draw the line which looks best to your eye.
fit lines *precisely*, we will need to clarify what is meant by ‘coming as close as possible to all the points in our data set.’ As you might guess, this ultimately boils down to measuring ‘distances’ between points in our data set and a given line. We shall discuss two common ways to measure these distances: the method of *perpendicular offsets* and the method of *vertical offsets*. From Figure 15, we can see that measuring a perpendicular offset is nothing more than computing the classical distance between a point and a line (see Proposition 8), whereas measuring a vertical offset involves computing the distance to the line in a vertical direction (perhaps you came up with this method in Your Turn 24).

![Figure 15. Fitting lines by perpendicular and vertical offsets](image)

6.1. **Pipelines and perpendicular offsets.** In this section, we explore the following problem (see Figure 16):

Kleenblu Water Company is to build a pipeline originating from the local water treatment plant (situated at the origin) and servicing towns with locations \((a_1, b_1), \ldots, (a_n, b_n)\).

Propose a linear path for the pipeline that comes closest to the towns.

Suppose the path of the pipeline is the line \(y = mx\). According to Proposition 8, the minimum distance from a town \((a_i, b_i)\) to the pipeline (i.e., the perpendicular offset) is \(\frac{|b_i - ma_i|}{\sqrt{m^2 + 1}}\). To solve the problem, we want to find \(m\) so that the totality of these offsets is as small as possible. One way to approach the problem is to minimize the sum of the offsets, but this involves doing calculus with unsavory things like absolute
values and square roots. So, to simplify our lives we minimize the sum $S(m)$ of the squared offsets\textsuperscript{11}, where

$$S(m) = \sum_{i=1}^{n} \frac{(b_i - ma_i)^2}{m^2 + 1}.$$ 

Of course, this is a calculus problem, and upon computing $S'(m)$ we find that the critical values for $S$ are precisely the zeroes of $C_1m^2 + C_2m - C_1$ (see Exercise 1), where

$$C_1 = \sum_{i=1}^{n} a_i b_i \quad \text{and} \quad C_2 = \sum_{i=1}^{n} (a_i^2 - b_i^2).$$

The quadratic formula then tells us that in case $C_1 \neq 0$, the critical values for $S$ are

$$m = \frac{-C_2 \pm \sqrt{C_2^2 + 4C_1^2}}{2C_1}.$$ 

Finally, these two values of $m$ must be tested (e.g., First Derivative Test) to see which one corresponds to a global minimum for $S$, and this solves the problem.

**Your Turn 25.** Suppose instead we are to find the path of a linear pipeline which originates at town $(a_1, b_1)$ (not necessarily the origin) and comes as close as possible to towns $(a_2, b_2), \ldots, (a_n, b_n)$. In terms of difficulty, how does this problem compare to the one we just solved? Discuss.

**Your Turn 26.** Solve the pipeline problem in the case that there are two towns located at $(1, 0)$ and $(0, 1)$. Compute $S(m)$ explicitly in this case. What interesting thing do you observe? Propose criteria on the towns that imply a unique solution to the pipeline problem.

\textsuperscript{11}There are also theoretical reasons for working with squared offsets: These heighten the influence of data points (in this case, towns) which lie 'far away' from the rest of the data without allowing them to dominate the approximation. Simply summing the (nonsquared) offsets does not give enough weight to data points 'out of line' with the others. Another reason for using squared offsets stems from the statistical distribution of error. For details on these issues one can consult a numerical analysis text (e.g., Burdon and Faires) and/or a text on statistical inference (e.g., Larson).
Another level of complexity is added to the pipeline problem if we do not require the pipeline to pass through the origin. In this case the path of the pipeline is $y = mx + b$ and the perpendicular offset from town $(a_i, b_i)$ to the pipeline is $\frac{|b + ma_i - b_i|}{\sqrt{m^2 + 1}}$, so the sum of the squared perpendicular offsets is

$$S(m, b) = \sum_{i=1}^{n} \frac{(b + ma_i - b_i)^2}{m^2 + 1}.$$ 

Using multivariable calculus, we find that the critical points for $S$ are of the form $(m, b)$, with $m$ a solution to the quadratic equation

$$(C_1 - nAB)m^2 + (C_2 + n(A^2 + B^2))m + (nAB - C_1) = 0,$$

and $b = B - mA$, where $C_1, C_2$ are as in (3) and

$$A = \frac{1}{n} \sum_{i=1}^{n} a_i \quad \text{and} \quad B = \frac{1}{n} \sum_{i=1}^{n} b_i.$$ 

These critical points must then be tested to determine which one determines a global minimum for $S$.

### 6.2. Vertical offsets, magnetic fields, and linear regression.

Have you thought about the earth’s magnetic field lately? If not, you’re probably not alone. Yet the earth’s magnetic field is an important part of nature. Birds, whales, and turtles seem to use the field to navigate, and the field draws harmful solar radiation toward the poles. (This channeling of radiation is what causes the aurora borealis in northern latitudes.) So, we might be concerned that the earth’s magnetic field has showed signs of marked dissipation over the last century. Our goal for this section is to obtain a reasonable linear model for the intensity of the field over the last several decades.

To see the decline in the intensity of the earth’s magnetic field, we consider some data obtained from the magnetic observatory in Teoloyucan, Mexico:

<table>
<thead>
<tr>
<th>Intensity (nT)</th>
<th>46076</th>
<th>45594</th>
<th>45176</th>
<th>44753</th>
<th>44178</th>
<th>43694</th>
<th>43108</th>
<th>42559</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (decades)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

Here, the first row of the table contains annual average intensity readings (in nanoTeslas) at the observatory for the years 1923, 1933, …, 1993. The second row gives time in decades since 1923. We can clearly see that the intensity of the field is declining. Also, a plot of the data (see Figure 17) shows that the data is nearly linear, so it makes sense to model the data using a line in the plane. We might then use this line to make predictions about the future strength of the magnetic field.
How are we to come up with a line modeling the magnetic field data? We could use the method of perpendicular offsets (described above), but it turns out that the method of vertical offsets might better suit the problem at hand. To see why ‘vertical is better’ in this case, we observe that there is an underlying function implied by the nature of the data set. Specifically, the (annual average) intensity of the magnetic field is naturally a function of time (measured in decades since 1923). Denoting this function \( f \), the table above indicates \( f(0) = 46076, f(1) = 45594, \) et cetera. Understanding the behavior of the function \( f \) is crucial. If we had very detailed information about the function \( f \), then we could make good predictions about the future intensity of the magnetic field. Therefore our primary goal should not merely be to find a line passing as close as possible to the data points (as in the previous section) but rather to find a linear function \( \ell(t) = mt + b \) whose output values come as close as possible to those for \( f \), insofar as is evidenced by the data. Thus, for \( t = 0, 1, 2, \ldots, 7 \), we want to minimize the totality of differences \( |\ell(t) - f(t)| \). Finally, as we can see from Figure 18, these differences \( |\ell(t) - f(t)| \) correspond to vertical offsets.

Now, as in the previous section, we are faced with a calculus minimization problem involving absolute values, which we avoid as before by minimizing the sum of the squared vertical offsets instead. That is, we are to minimize the two-variable function

\[
S(m, b) = \sum_{i=1}^{n} ((mt_i + b) - f(t_i))^2,
\]
where, in this case, $n = 8$ and $t_1 = 0$, $t_2 = 1$, et cetera. After a little work, we find that $S(m, b)$ will be minimized when

$$m = \frac{n (\sum_{i=1}^{n} t_i \cdot f(t_i)) - (\sum_{i=1}^{n} t_i) (\sum_{i=1}^{n} f(t_i))}{n (\sum_{i=1}^{n} t_i^2) - (\sum_{i=1}^{n} t_i)^2},$$

and

$$b = \frac{(\sum_{i=1}^{n} t_i^2) (\sum_{i=1}^{n} f(t_i)) - (\sum_{i=1}^{n} t_i \cdot f(t_i)) (\sum_{i=1}^{n} t_i)}{n (\sum_{i=1}^{n} t_i^2) - (\sum_{i=1}^{n} t_i)^2}.$$

Substituting in our specific information gives $m = -500.833$ and $b = 46145.2$. So, our linear approximation of the function $f$ (often called a ‘least squares’ approximation) is

$$\ell(t) = 46145.2 - 500.833t.$$

Figure 19 shows how well $\ell$ fits the data.
Your Turn 27. Using Fathom or a handheld calculator, enter the magnetic field data and fit a line to the data. Do you obtain the same line as given in (6)? Does your calculator use the method of vertical offsets or the method of perpendicular offsets?

Your Turn 28. Using the function \( \ell \) given in Equation (6), predict when the intensity of the magnetic field at Teoloyucan will hit 0. What common sense argument might you offer against this prediction?

6.3. Exercises.

1. Given

\[
S(m) = \sum_{i=1}^{n} \frac{(b_i - ma_i)^2}{m^2 + 1},
\]

find \( S'(m) \) and compute the critical values for \( S \).

2. We saw above that the critical values for \( S \) (Exercise 1) are precisely the zeroes of \( C_1 m^2 + C_2 m - C_1 \), where

\[
C_1 = \sum_{i=1}^{n} a_i b_i \quad \text{and} \quad C_2 = \sum_{i=1}^{n} (a_i^2 - b_i^2).
\]

Describe the nature of the critical values in the case \( C_1 = 0 \).

3. For the data set \{ (1, 2), (2, 1.5), (3, 4) \},

(a) Find \( m \) so that the line \( y = mx \) minimizes the sum of the squared perpendicular offsets.

(b) Find \( m \) so that the line \( y = mx \) minimizes the sum of the perpendicular offsets (not squared).

(c) Find \( m \) so that the line \( y = mx \) minimizes the sum of the squared vertical offsets.

(d) Find \( m \) so that the line \( y = mx \) minimizes the sum of the vertical offsets (not squared).

4. PerpetuGas Inc. is to build a linear gas pipeline originating at Town A and passing as close as possible to Towns B, C, and D. Propose a path for the pipeline given that the locations of the towns are (1, 2), (2, 5), (2, 2), and (4, 3), respectively. (Simplify the problem by translating (1, 2) to (0, 0) and moving the rest of the points left one unit and down two units.)
5. Crickets are known to chirp faster as temperature rises. For example, consider the following cricket data:

<table>
<thead>
<tr>
<th>Chirps per 10 seconds</th>
<th>Temperature (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>58</td>
</tr>
<tr>
<td>15</td>
<td>62</td>
</tr>
<tr>
<td>16</td>
<td>67</td>
</tr>
<tr>
<td>17</td>
<td>67</td>
</tr>
<tr>
<td>18</td>
<td>65</td>
</tr>
<tr>
<td>19</td>
<td>72</td>
</tr>
<tr>
<td>20</td>
<td>70</td>
</tr>
</tbody>
</table>

Use the discussion above to fit a line to this data, and use your line to:

(a) predict the temperature if the cricket chirps 9 times per 10 seconds, and

(b) predict the temperature at which crickets will stop chirping.

(Note: You will have to choose between the methods of perpendicular offsets and vertical offsets. Which method is more appropriate here?)

For the following exercises, we recommend you use the statistical software Fathom to plot data and produce a ‘best fit’ (least squares) line.

6. For this exercise you will need a large paper cup, a clear plastic or glass measuring cup, and a timer. Fill the paper cup with water, punch a small hole in the bottom with a sharp object, and then record the amount of water in the measuring cup in ten-second intervals over a period of two minutes.

(a) Make a scatterplot of the data you’ve collected.
(b) Fit a line to the data.
(c) Use your line to predict the amount of water lost after ten minutes.
(d) Imagine the hole in the cup is a leak in one of your home water pipes. Predict how much water will be wasted in one month.

7. (For those who have had multivariable calculus) Verify that the critical points for $S$ are of the form $(m, b)$, with $m$ a solution to the quadratic equation

$$(C_1 - nAB)m^2 + (C_2 + n(A^2 + B^2))m + (nAB - C_1) = 0,$$

and $b = B - mA$, where $C_1, C_2$ are as in (3) and

$$A = \frac{1}{n} \sum_{i=1}^{n} a_i \quad \text{and} \quad B = \frac{1}{n} \sum_{i=1}^{n} b_i.$$  

8. (For those who have had multi-variable calculus.) Verify Equations (4) and (5).
7. The Collection of All Lines in the Plane

Up to this point, we have mostly considered lines “one at a time.” We have, for instance, taken a single line and considered its graph, its slope, and its intercepts. When finding intersections, we have also considered two lines at once. Now however, we will try to find a convenient geometric description for “all” (or at least, “most”) lines at the same time. Such a description of a large group of objects is often called a moduli space. Just as the graph of a function helps us visualize the behavior of the function, a moduli space for lines can help us visualize various large collections of lines which might interest us (e.g., the collection of all lines with positive slope).

7.1. All non-vertical lines and the mb-plane. We all know that a line is made up of infinitely many points. So, it may be surprising that the key to visualizing a large set of lines is to represent each line by a single point. Let’s practice with this idea:

Let’s Go 4. Consider Figure 20, which shows a collection of points on the left, and a collection of lines on the right.

Figure 20. Pairing points and lines

(a) Looking carefully at Figure 20, describe a natural way to pair a point (on the left) with a line (on the right). Then, list your pairings.
(b) Find an equation for the line which would be paired with the point \((-7,2.5)\).
(c) Is there any point that can be paired with a vertical line? Why or why not?
(d) Can a given non-vertical line be paired with two different points? Why or why not?
Let’s Go 4 suggests that any non-vertical line can be paired with a point \((m, b)\) (and vice versa), where \(m\) is the slope of the line and \(b\) is the \(y\)-intercept. Therefore, the collection of all non-vertical lines is in one-to-one correspondence with \(\mathbb{R}^2\), thought of as the \(mb\)-plane.

If a single point in the \(mb\)-plane determines a single non-vertical line, then a collection of points in the \(mb\)-plane (such as a curve or region in the \(mb\)-plane) represents a collection of non-vertical lines. For example, consider the set of all non-vertical lines in the plane whose slopes are equal to their \(y\)-intercepts (see left side of Figure 21). These lines correspond to points in the \(mb\)-plane for which \(b = m\) (why?). If we plot these points in the \(mb\)-plane, we obtain a line passing through the origin (see Figure 21). Therefore,

\[\text{Figure 21. Lines with slopes equal to } y\text{-intercepts}\]

the points on the line \(m = b\) in the \(mb\)-plane represent all non-vertical lines whose slopes are equal to their \(y\)-intercepts.

**Your Turn 29.** Is it coincidence that all of the lines in the left side of Figure 21 pass through the point \((-1, 0)\)? Discuss.

**Your Turn 30.** Give a description in the \(mb\)-plane for each of the following sets of lines in the \(xy\)-plane:

(a) All lines containing the point \((0, 4)\).

(b) Lines with \(y\)-intercept strictly between 2 and 3.

(c) Lines whose slope is an integer.
Your Turn 31. When giving examples of lines in the plane, my algebra teacher, Mrs. D., has a tendency to give only examples that pass through the origin. How might you use the mb-plane to produce a pictorial argument convincing Mrs. D. that her policy is causing her to miss most lines in the plane?

7.2. Non-vertical, non-horizontal lines and the ab-plane. In the previous section we were able to describe large collections of lines by pairing each line with a point in the mb-plane. It turns out that there are other ways to produce pairings between lines and points. For example,

Let’s Go 5. Consider Figure 22, which shows a collection of points on the left, and a collection of lines on the right.

![Figure 22. Pairing points and lines, again](image)

(a) Looking carefully at Figure 22, describe a natural way to pair a point (on the left) with a line (on the right). Then, list your pairings.

(b) Find an equation for the line which would be paired with the point (−7, 2.5).

(c) Can a vertical line be paired with a unique point? (Don’t forget to consider the y-axis.)

(d) Can a horizontal line be paired with a unique point? (Don’t forget to consider the x-axis).

(e) Can the point (0, 0) be paired with a unique line?

Let’s Go 5 suggests that any line not passing through the origin with unique x- and y-intercepts can be paired with a single point \((a, b)\), where \(a\) and \(b\) are the (nonzero) x-and y-intercepts of the line, respectively. Thus the collection of all lines with unique x- and y-intercepts not going through the origin is in one-to-one correspondence with \(\mathbb{R}^2\) with the axes removed, thought of as the ab-plane.
We return to our example from the previous section. Consider all non-horizontal lines whose slope and $y$-intercept are equal. Recall from Your Turn 29 that, regardless of $y$-intercept, all of these lines have $x$-intercept $-1$. So, each of these lines is represented in the $ab$-plane by a point of the form $(-1, b)$. This gives rise to a (punctured) vertical line $a = -1$ in the $ab$-plane (see Figure 23).

![Figure 23. Lines with slopes equal to their $y$-intercepts, again](image)

**Your Turn 32.** Give a description in the $ab$-plane for each set of lines in the $xy$-plane.

(a) Non-vertical, non-horizontal lines containing the point $(0,4)$.

(b) Non-vertical, non-horizontal lines with $y$-intercept strictly between 2 and 3.

(c) Lines whose slope is a nonzero integer.

**Your Turn 33.** Consider the sequence $(1,1), (1,1/2), (1,1/3), (1,1/4)\ldots$ of points in the $ab$-plane. The limit of the sequence is $(1,0)$. Is it reasonable to think of this point as representing some line in the $xy$-plane? Discuss.

7.3. **The collection of all lines in the $xy$ plane.** The models we have considered thus far, the $mb$- and $ab$-planes, have fallen short of describing all lines in the plane. (The first only describes non-vertical lines, and the second describes non-vertical, non-horizontal lines not passing through the origin). We can describe all lines in the plane, but the picture is a bit more complicated.

Consider the unit sphere $S$ in $\mathbb{R}^3$, centered at the origin. We define an equivalence relation on the sphere by saying that $(x, y, z)$ is equivalent to itself and to its negative $(-x, -y, -z)$. Thus every equivalence class contains exactly two points.
Theorem 9. The collection of all lines in the $xy$-plane is in one-to-one correspondence with the equivalence classes of points on the sphere $S$, minus the north and south poles\textsuperscript{12}.

Proof. Figure 24 will be a handy reference throughout. Consider the $xy$-plane embedded in $\mathbb{R}^3$ as the plane $z = 1$, not $z = 0$. Given a line $L$ in this plane, it determines a plane $P$ through the origin in $\mathbb{R}^3$; this is the plane through the origin whose intersection with the plane $z = 1$ is the line $L$. Let $L'$ be the line through the origin that is perpendicular to $P$. It intersects $S$ in exactly two points, which are equivalent points. Note that one cannot obtain the north and south poles: this would correspond to $L'$ being a vertical line and $P$ being the plane $z = 0$, which doesn’t intersect the plane $z = 1$.

Conversely, starting with a pair of equivalent points on the sphere $S$ (except the poles), we get a non-vertical line $L'$ through the origin in $\mathbb{R}^3$ (namely, the line connecting the two points). We let $P$ be the plane through the origin in $\mathbb{R}^3$ that is perpendicular to $L'$ (it is not horizontal since $L'$ is not vertical). Finally we obtain a line $L$ in the plane $z = 1$, namely, $L$ is the intersection of the plane $P$ and the plane $z = 1$. □

7.4. Exercises.

1. Sketch the region in the $mb$-plane corresponding to each family of (nonvertical) lines.

   (1) All lines through the origin.

   (2) All horizontal lines.

   (3) All lines with slope 1.

\textsuperscript{12}In other words, the collection of all lines in the plane may be identified with the real projective plane, less a point.
(4) All lines with slope 2 whose $y$-intercept is an integer.

(5) All lines with slope 2 whose $y$-intercept is positive.

(6) All lines with positive slope.

(7) All lines with $x$-intercept equal to 4.

(8) All lines with $x$-intercept between 8 and 11.

(9) All lines with $x$-intercept between $-1$ and 2.

(10) All lines through the point $(5, 7)$.

(11) All lines whose $x$- and $y$-intercepts are the same number.

(12) All lines tangent to the unit circle.

(13) All lines with nonempty intersection with the unit disk $\{(x, y) : x^2 + y^2 \leq 1\}$.

2. Sketch the region in the $ab$-plane corresponding to each family of lines given in Exercise 1. (Only consider lines with unique $x$- and $y$-intercepts not going through the origin!)

3. Describe all the lines in the $xy$-plane that correspond to the following equivalence classes on the unit sphere (as described in the proof of Theorem 9):

   (1) $\{(1, 0, 0), (-1, 0, 0)\}$

   (2) the equator

   (3) $\{\pm(1/\sqrt{3},1/\sqrt{3},1/\sqrt{3})\}$

   (4) the circle $x^2 + z^2 = 1$

4. In the proof of Theorem 9 we identified all lines in the plane with equivalence classes of points on the unit sphere.

   (a) What portion of the unit sphere corresponds to points in the $ab$-plane? (Hint: Identify a point $(a_0, b_0)$ in the $ab$-plane with a line in the ‘lifted’ $xy$-plane $\{(x, y, 1) \mid x, y \in \mathbb{R}\}$. Find two points on this line, then find a vector normal to the plane determined by these two points together with $(0,0,0)$.)

   (b) What portion of the unit sphere corresponds to points in the $mb$-plane? (Hint: Operate as in part (a).)
8. Exercises Involving Student Work

This section contains student solutions to various problems involving lines in the plane. We recommend that these solutions be analyzed prior to certain sections in the chapter (indicated below). For each problem and corresponding set of solutions, complete the following tasks:

1. Take 5 minutes in an attempt to solve the problem (without appealing to resources such as your text, etc.).
2. Discuss each student’s solution strategy. Indicate ways in which the student was “on-track”. Also, identify any incorrect assumptions, misconceptions, and errors revealed in each student’s work.
3. Try to determine whether the student’s numerical answer is correct.
4. Rate each solution on a scale from 1 to 5 with 5 as the highest rating and 1 as the lowest.
5. Compare your ratings in part 4 with another student (or group), and discuss your criteria for giving these ratings. Attempt to reach consensus with your partner (or group).
6. Carefully construct at least two solutions to the problem which employ different strategies. (You may use your text as a resource if you wish.)

5. (To be considered prior to Section 5.3) Find an equation for the line which is parallel to \(2x + 3y = 7\) and contains the point (4, 10).

6. (To be considered prior to Section 5.4) Find the distance between the line \(y = -4x + 3\) and the point (0, 0).
CHAPTER 3

Quadratic Polynomials

Polynomials\(^1\) are useful for approximation in calculus because we know exactly how to integrate and
differentiate them.\(^2\) This, among other considerations, underscores the importance of understanding polyno-
mials, and partially explains the presence of polynomials in the secondary curriculum. Yet algebra involving
polynomials is far from simple, and has been at the root of some of the most famous longstanding open
problems in mathematics.\(^3\) There is no sure-fire way to factor polynomials or to solve polynomial equations.

One way to approach the study of polynomials is to investigate low-degree polynomials first. Linear
(that is, degree-one) polynomials have graphs that are lines, and were the subject of the last chapter. The
next logical step is to examine quadratic (that is, degree-two) polynomials, which have the form
\[ ax^2 + bx + c, \]
where \(a, b, c\) are fixed real numbers with \(a \neq 0\). In this chapter, we will see that quadratic equations can be
readily solved, that quadratic polynomials have important applications (e.g., splines), and that the graphs of
quadratic polynomials are parabolas, thus distinguishing quadratic polynomials from other functions having
‘U-shaped’ graphs.

What is it about quadratic polynomials that make them so approachable? Consider the following Let’s
Go:

\section*{Let’s Go} 1. A graphing utility may be handy for this task.

\footnote{A \textit{polynomial} is an expression of the form \(a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0\), where \(x\) is an indeterminate and \(a_n, \ldots, a_0\)
are constants (for us, usually real numbers, occasionally complex numbers). If the \(a_i\)s are real numbers, we obtain a \textit{polynomial function}
with domain and codomain equal to the real numbers; the function assigns the output \(a_nr^n + a_{n-1}r^{n-1} + \cdots + a_1r + a_0\)
to the input \(r\).}

\footnote{This realization fueled the development of calculus. For example, due to the geometric series expansion \(\frac{1}{1-x} = 1 + x + x^2 + \cdots\) for \(|x| < 1\), Isaac Newton observed that one could obtain approximations of the natural logarithm by integrating the
series term by term, just as one would do with a polynomial.}

\footnote{For over two thousand years, it was unknown whether one could construct a square with area equal to that of a given
circle using only a compass and unmarked straightedge. This construction, known as \textit{squaring the circle}, was finally shown to
be impossible in 1882 when F. Lindemann showed that \(\pi\) cannot be a root of any polynomial with rational coefficients.}
(a) Graph \( y = x^2 \) and \( y = -2x^2 + 4x - 9 \). Describe how by graphing transformations (translating, reflecting and stretching), you can put one graph atop the other.

(b) Based on part (a), do you think that the graph of every quadratic polynomial can be obtained from the graph of \( y = x^2 \) by graphing transformations?

(c) Can the graph of every cubic (degree-three) polynomial be obtained from the graph of \( y = x^3 \) by graphing transformations? Discuss.

Let's Go 2. In this exercise you are asked to explore your prior knowledge of quadratic polynomials.

(a) List as many mathematical topics, concepts, and computational techniques as you can that are connected to quadratic polynomials.

(b) What should secondary students learn and know about quadratic polynomials?

(c) List aspects of quadratic polynomials that are likely to be difficult or confusing for secondary students.

Perhaps surprisingly, one can understand virtually everything about a quadratic polynomial \( ax^2 + bx + c \) by understanding \( x^2 \). This happens through the algebraic technique called completing the square, which we will explore in detail. In terms of the graphs, the technique implies that the graph of \( ax^2 + bx + c \) is obtained from the graph of \( x^2 \) merely by a stretching in the vertical direction, followed by shifts in the vertical and horizontal directions. Further, completing the square gives an easy way of finding the roots of any degree-two polynomial, and yields the quadratic formula.

1. The Squaring Function and Parabolas

In this section we discover that the squaring function from \( \mathbb{R} \) to \( \mathbb{R} \), defined by the simple algebraic formula \( x \mapsto x^2 \), has special geometric characteristics.

1.1. The graph of the squaring function: more than just a U-shaped curve. Given that many functions have graphs that are roughly U-shaped, what is unique about the shape of the squaring function’s graph? To begin sorting this out, consider the following Let’s Go:

Let’s Go 3. Refer to Figure 1.
1. THE SQUARING FUNCTION AND PARABOLAS

(a) Make a list of words describing the graph of the squaring function. (Feel free to use some of the words you learned in calculus.)

(b) Using a graphing utility, examine the graphs of \( y = |x^2|, y = x^4, y = \frac{3x^2}{1-x^2}, \) and \( y = \frac{\sqrt{1+x^2} - 1}{\sqrt{2}-1}, \) for \(-1.2 < x < 1.2\). Which words from part (a) can also be used to describe these graphs? What happens if you plot the graphs for \(-10 < x < 10\)?

(c) Are there any words from part (a) that do not serve to describe the graphs in part (b)?

Like the curves in Let’s Go 3, a parabola is a special type of U-shaped curve. However, while the curves in Let’s Go 3 are defined algebraically, a parabola is defined geometrically, using fundamental geometric concepts such as line, point, and distance. Namely, given a line \( L \) and a point \( F \) in the plane, the parabola determined by \( L \) and \( F \) is the collection of points in the plane equidistant from \( L \) and \( F \). The line \( L \) is called the directrix of the parabola and the point \( F \) is the focus of the parabola (see Figure 2).

Your Turn 1. Sketch the axis of symmetry of the parabola in Figure 2. Why, in fact, do we expect the figure to be symmetric?

At first glance, the process of finding points equidistant from a given line and point in the plane seems to have little to do with the process of squaring numbers, but there is a connection: The graph of \( y = x^2 \)

---

Footnote: You may recall that parabolas, along with ellipses and hyperbolas, are (non-degenerate) conic sections. That is, they can be obtained by slicing a cone with a plane. The names for these curves were first coined by Apollonius in the 3rd century BCE.
is the parabola with focus $(0, \frac{1}{4})$ and directrix $y = -\frac{1}{4}$. To see why this is true, let $(x, y)$ be a point in the plane, as shown in Figure 3. Its distance to the point $(0, \frac{1}{4})$ is $\sqrt{x^2 + \left(y - \frac{1}{4}\right)^2}$, while its distance to the line $y = -\frac{1}{4}$ is $|y + \frac{1}{4}|$. Thus $(x, y)$ is on the parabola $\iff \sqrt{x^2 + \left(y - \frac{1}{4}\right)^2} = \left|y + \frac{1}{4}\right|$

$\iff x^2 + \left(y - \frac{1}{4}\right)^2 = \left(y + \frac{1}{4}\right)^2$

$\iff x^2 + y^2 - \frac{y}{2} + \frac{1}{16} = y^2 + \frac{y}{2} + \frac{1}{16}$

$\iff x^2 = y$. 

Figure 2. A parabola, with focus and directrix

Figure 3. $y = x^2$ as a parabola
So \((x, y)\) lies on the graph of \(y = x^2\) if and only if it lies on the parabola with focus \((0, \frac{1}{4})\) and directrix \(y = -\frac{1}{4}\). The same reasoning is used to show that the graph of \(y = ax^2\), obtained from \(y = x^2\) by vertical stretching, is a parabola with focus \((0, \frac{1}{4a})\) and directrix \(y = -\frac{1}{4a}\) (see Exercise 3).

**Your Turn 2.** Using a ruler, convince yourself that the graph given in Figure 2 is really a parabola having the indicated focus and directrix.

**Your Turn 3.** While teaching a pre-calculus class, you ask: “What is a parabola?” One student answers, “Basically, it’s a U.” Another student answers, “It’s \(y = x^2\).” Give a response to your students that is helpful and mathematically accurate.

1.2. **The reflection property.** When objects (photons, marbles, etc.) strike a surface, they bounce off in accordance with the *equal angles principle*, where the ‘equal angles’ are determined by the path of the object together with the tangent line to the surface at the point of contact (see part (a) of Figure 4). Depending on the shape of the surface, the total effect of this principle can be interesting. For example, the principle demands that *any object which bounces off of a parabola upon arriving from the focus must travel ‘out of’ the parabola on a path parallel to the axis of symmetry* (see part (b) of Figure 4).\(^5\) This is called the *reflection property* of the parabola. Specifically:

---

\(^5\)The effect on an ellipse is striking as well: The equal angles principal implies that objects travel from one focus to the other. This has interesting applications, such as the whispering room in the U.S. Capitol building (sound travels from one focus to another) and dental lamps (a light source at one focus illuminates your teeth at the other focus).
Theorem 1. (Reflection Property) Let $a$ be a real number and consider the parabola with graph $y = ax^2$. Take the line segment from the focus $(0, \frac{1}{4a})$ to a point $(x_0, ax_0^2)$ on the graph, and consider a segment parallel to the y-axis with lower endpoint $(x_0, ax_0^2)$, as shown in Figure 5. The two angles $\theta_1$ and $\theta_2$ formed by these segments together with the tangent line at $(x_0, ax_0^2)$ are equal.

**Figure 5.** The reflection property

To verify the reflection property (Theorem 1), let $(x_0, ax_0^2)$ be the point on the graph. Our goal is to show that $\theta_1$ and $\theta_2$ are equal. Since $\theta_1$ and $\theta_3$ are opposite angles, they are equal; hence we are done if we can show that $\theta_2 = \theta_3$.

The triangle with vertices at $(x_0, ax_0^2)$, $(0, \frac{1}{4a})$, and $(x_0, -\frac{1}{4a})$ is isosceles because $(x_0, ax_0^2)$, as a point on the parabola, is equidistant from the focus and the directrix. By calculus, the slope of the tangent line to the graph of $y = ax^2$ at $(x_0, ax_0^2)$ is $2ax_0$. On the other hand, the slope of the line joining $(0, \frac{1}{4a})$ and $(x_0, -\frac{1}{4a})$ is $-\frac{1}{2ax_0}$. Since the product of the slopes is $-1$, the two lines are perpendicular. Thus the tangent line divides the isosceles triangle with vertices at $(x_0, ax_0^2)$, $(0, \frac{1}{4a})$, and $(x_0, -\frac{1}{4a})$ into two congruent halves. It follows that $\theta_2 = \theta_3$, proving the reflection property.

**Your Turn 4.** Explain how the parabolic reflection property could be used to heat a hot dog outdoors on a sunny day. Then produce a list of everyday objects that exploit the parabolic reflection property described above.

1.3. Exercises.
1. THE SQUARING FUNCTION AND PARABOLAS

1. (a) Draw a line and a point not on the line, then construct five points on the parabola determined by the point and line. Explain your strategy.

(b) Using straightedge and compass, how could you construct points on the parabola?

2. Graph the parabola with focus \((0, -0.001)\) and directrix \(y = 0.001\).

3. We consider \(y = ax^2\).

(a) Compare the graph of \(y = ax^2\) with that of \(y = x^2\) when \(a = 2\), when \(a = -\frac{1}{2}\), and when \(a = -4\). (Graph these on a single set of axes.)

(b) Show that the graph of \(y = ax^2\) is the parabola with focus \((0, \frac{1}{4a})\) and directrix \(y = -\frac{1}{4a}\) by showing that each point on the graph of \(y = ax^2\) is equidistant from the focus and the directrix.

4. Use Exercise 3 to find the equation of the parabola with focus \((0, 7/3)\) and directrix \(y = -7/3\).

5. What happens to the focus of the parabola \(y = ax^2\) as \(a \to \infty\)? As \(a \to 0\)? Construct a few graphs illustrating these two situations. (Use Exercise 3.)

6. We explore \(y = x^2\) in the context of inverse functions (see Section 6.3 of Chapter 1).

(a) Let \(f_1 : [0, \infty) \to [0, \infty)\) be the function \(f_1(x) = x^2\). Graph \(f_1\). Find a formula for \(g_1(x)\), the inverse function of \(f_1(x)\). What are the domain and range of \(g_1\)? Explicitly show that \(f_1 \circ g_1\) and \(g_1 \circ f_1\) are the identity functions on their domains. Graph \(g_1\) on the same axes as \(f_1\). How are their graphs related to each other?

(b) Let \(f_2 : (-\infty, 0] \to [0, \infty)\) be the function \(f_2(x) = x^2\). Repeat part (1) for \(f_2(x)\).

(c) Let \(f : (-\infty, \infty) \to [0, \infty)\) be defined by \(f(x) = x^2\). Does \(f(x)\) have an inverse function? Explain.

7. We use the method of ‘finite differences’ to investigate rates of change of the squaring function.

(i) Make a table with the numbers 1, 2, 3, \ldots, 10 down the left column.

(ii) In the second column, give the squares of the entries in the first column (1, 4, 9, \ldots, 100).

(iii) In the third column, give the differences between the entries in the second column (4 − 1, 9 − 4 \ldots).
(iv) In the fourth column, give the differences between the entries in the third column.

Now respond to the following:

(a) What resemblance do you see between the table, and the first and second derivatives of \( x^2 \)?

(b) How can one explain the results of the table algebraically? (Hint: call three successive entries in the first column \( n \), \( n + 1 \), and \( n + 2 \), then try to fill in the entries in the remaining columns.)

(c) What do you expect would happen if you used the function \( x^3 \) instead of \( x^2 \)? Try to prove your conjecture.

8. \( Constructing \text{ tangents to a parabola.} \) Consider the parabola with equation \( y = ax^2 \), and let \( Q \) be a point on the parabola other than the vertex. Let \( P \) be the point on the \( y \)-axis obtained in the following way: from \( Q \), move horizontally to the \( y \)-axis and call this point on the \( y \)-axis \( Q' \). Then let \( P \) be the reflection of \( Q' \) across the \( x \)-axis.

(a) Illustrate this construction on paper or with a graphing utility, being sure to construct the line \( PQ \).

Does this line appear to be tangent to the parabola?

(b) Show that the line \( PQ \) must be tangent to the parabola at \( Q \).

9. In this exercise, we consider parabolas whose axis of symmetry need not be parallel to the \( y \)-axis.

(a) Find the equation of the parabola with focus \( \left( \frac{1}{4}, 0 \right) \) and directrix \( x = -\frac{1}{4} \).

(b) Find the equation of the parabola with focus \( (2, 7) \) and directrix \( 3x + 4y = 1 \). (You will need the formula for the distance between a point and a line. Also, by squaring, try to write your answer in the form \( ax^2 + bx + cy^2 + dy + exy = f \).)

(c) Let \( L \) be the line with equation \( Ax + By = C \), and let \( (x_0, y_0) \) be a point not on the line. Find the equation for the parabola determined by this line and point. (Again, try to write your answer in the form \( ax^2 + bx + cy^2 + dy + exy = f \). The constants \( a, \ldots, f \) will depend on \( A, B, C, x_0, y_0 \).)

10. Show that the graph of \( y = x^4 \) is not a parabola. \( One strategy is to use Exercise 9. \)

11. We can regard a quadratic polynomial \( ax^2 + bx + c \) as a function that assigns the output \( ax^2 + bx + c \) to the input \( x \).
(a) Let \(a_1, a_2, b_1, b_2, c_1, c_2 \in \mathbb{R}\). Prove that if \(a_1x^2 + b_1x + c_1 = a_2x^2 + b_2x + c_2\) for all \(x \in \mathbb{R}\), then \(a_1 = a_2, b_1 = b_2,\) and \(c_1 = c_2\). (This shows that distinct quadratic polynomials produce distinct functions from \(\mathbb{R}\) to \(\mathbb{R}\).)

(b) The analogue of (a) may be false when \(\mathbb{R}\) is replaced with a different number system. Illustrate this by finding two polynomials (each of degree two or less) with coefficients in \(\mathbb{Z}/2\) which yield the same function from \(\mathbb{Z}/2\) to \(\mathbb{Z}/2\).

2. Completing the Square and Solving Quadratic Equations

We shall see shortly that aspects of general quadratic polynomials \(ax^2 + bx + c\) (e.g., graphs and roots) reduce to aspects of the squaring function.\(^6\) The key technique is known as completing the square, highlighted in this section.

**Let’s Go 4.** Based on what you recall from high school algebra, perform the operation called “completing the square” on the polynomials \(x^2 + 7x + 5\) and \(3x^2 + 11x + 2\). Show your work. What goal is accomplished by completing the square?

The following Let’s Go introduces us to the main ideas involved:

**Let’s Go 5.** Consider the quadratic equation \(x^2 + 7x + 5 = 0\).

(a) Substitute \(z = x + 7/2\) in the equation above (that is, replace \(x\) by \(z - 7/2\)) and simplify. What is special about the resulting quadratic equation?

(b) Solve the quadratic equation you obtained in part (a). Was it easy? Why?

(c) Use your solution to part (b) to find a solution to \(x^2 + 7x + 5 = 0\).

(d) Factor the expression you obtained in part (a). How does this help you factor \(x^2 + 7x + 5\)?

2.1. Algebraic viewpoint. You likely observed in Let’s Go 5 that the absence of a ‘middle term’ (the term \(bx\)) makes the business of solving quadratic equations and factoring quadratic expressions much easier.

For instance, we solve \(x^2 - 10 = 0\) easily by setting \(x^2 = 10\) and then finding the square roots of 10, and we

---

\(^6\)This very special behavior does not apply to higher degree polynomials in general, as you may have observed in Let’s Go 1.
quickly factor \( x^2 - 10 \) into \( (x - \sqrt{10})(x + \sqrt{10}) \) as a difference of squares. So, we should be able to solve and to factor easily, provided that we can sensibly hide away the middle term of a quadratic polynomial. This ‘hiding’ of middle terms (hinted at in Let’s Go 5) is known as completing the square.

Acting on insight gained in Let’s Go 5, to ‘complete the square’ for \( x^2 + bx + c \) we substitute \( x = \frac{b}{2} \), simplify the corresponding expression, and then return to an expression in \( x \) by substituting \( z = x + \frac{b}{2} \):

\[
x^2 + bx + c = \left( z - \frac{b}{2} \right)^2 + b \left( z - \frac{b}{2} \right) + c
\]
\[
= \left( z^2 - bz + \frac{b^2}{4} \right) + \left( bz - \frac{b^2}{2} \right) + c = z^2 + \left( c - \frac{b^2}{4} \right)
\]
\[
= \left( x + \frac{b}{2} \right)^2 + \left( c - \frac{b^2}{4} \right).
\]

These steps are summarized in Equation (1):

\[
(1) \quad x^2 + bx + c = \left( x + \frac{b}{2} \right)^2 + \left( c - \frac{b^2}{4} \right).
\]

Then, if we are asked to solve \( x^2 + bx + c = 0 \), we may use Equation (1) to obtain \( (x + \frac{b}{2})^2 = -\left( c - \frac{b^2}{4} \right) \), which can be solved by taking square roots of both sides in the same way that you would solve an equation like \( x^2 = 10 \).

**Your Turn 5.** A student claims that you can’t solve \( (x + \frac{b}{2})^2 = -\left( c - \frac{b^2}{4} \right) \) because the right hand side is negative. Give a helpful response.

**Your Turn 6.** Use Equation (1) to factor \( x^2 + bx + c \). (Interpret the right side of the equation as the difference of two squares.)

**Your Turn 7.**

(a) Many students learn a method for completing the square as follows:

\[
x^2 + 10x + 17 = (x^2 + 10x) + 17 = (x^2 + 10x + 25) + 17 - 25 = (x + 5)^2 - 8.
\]

At first glance this appears very different from the point of view given in in Equation (1) and Let’s Go 5. Use this method to complete the square in \( x^2 + bx + c \). Why do you get the same result as Equation 1?

---

7As we will see in a later chapter, the substitution method is not only useful for completing the square, but is also useful for ‘hiding’ certain middle terms of higher degree polynomials. Using substitution to simplify quadratic equations originates with the Babylonians (ca. 1700 BCE), who were able to use this method to solve quadratic equations in generality.
2. Completing the Square and Solving Quadratic Equations

(b) Why do you think “completing the square” is called “completing the square”?

The general case is not much harder (see Exercise 5); using the substitution \( z = x + \frac{b}{2a} \) allows us to complete the square for \( ax^2 + bx + c \):

**Theorem 2.** If \( a, b, c \) are real numbers with \( a \neq 0 \), then

\[
ax^2 + bx + c = a \left( x + \frac{b}{2a} \right)^2 + \left( c - \frac{b^2}{4a} \right).
\]

**Your Turn 8.** Use Theorem 2 to solve the equation \( ax^2 + bx + c = 0 \) (assume \( a \neq 0 \)). Derive the quadratic formula from your answer.

### 2.2. Geometric viewpoint.

The prefix *quad* in the word quadratic means *square*. Multiplying a number by itself is called *squaring*. Our method of solving quadratic equations (explored above) is called completing the square. With all of these ‘squares’ conjuring up thoughts of the geometric squares, perhaps it is not surprising that square figures provided the original motivation for completing the square.

To see how square figures fit into the picture, we examine the quadratic \( x^2 + 10x \). (The constant term has been set to zero intentionally because it plays no role in completing the square.) We view the quantity \( x^2 + 10x \) as representing the area of a region in the plane, where \( x^2 \) represents the area of a square with side length \( x \), and \( 10x \) represents the sum of the areas of two \( 5 \times x \) rectangles (each with area \( 5x \)). Gluing the square and the two rectangles gives the \( L \)-shaped region (called a *gnomon*) shown in part (a) of Figure 6. We then ‘complete the square’\(^8\) by tacking on the \( 5 \times 5 \) square in the corner, as shown in part (b) of Figure 6.

From the picture, the original area \( x^2 + 10x \) is what remains when the \( 5 \times 5 \) rectangle is removed from the corner of the ‘completed’ \( (x + 5) \times (x + 5) \) square\(^9\)

\[
x^2 + 10x = (x + 5)^2 - 25.
\]

\(^8\) *Completing the square* was first described in the 9th century by Mohammed ibn-Musa al-Khwarizmi, working in Baghdad, who used it to solve quadratic equations. For example, to solve \( x^2 + 10x = 30 \), the equation becomes \( (x + 5)^2 - 25 = 30 \), or \( (x + 5)^2 = 55 \). Arab mathematicians of this era were particularly inspired by the geometric algebra of the ancient Greeks, hence the geometric flavor of their method.

\(^9\) Manipulatives known as *algebra tiles* are often used in middle school to explore concepts such as these. Students create figures much like those given in Figure 6.
Your Turn 9. Give a geometric picture and description showing how to complete the square for $x^2 - 10x$.
(Since the coefficient of $x$ is negative, you will need a slightly different picture than the one for $x^2 + 10x$.)

2.3. Exercises.

1. Consider $x^2 + 10x - 30$.
   (a) Complete the square using any method.
   (b) Use your answer to part (a) to solve the equation $x^2 + 10x = 30$.
   (c) Use your answer to part (a) to factor $x^2 + 10x - 30$. (View the result of part (a) as a difference of two squares. Factor accordingly.)

2. Complete the square. In parts (a) and (c), use the substitution $z = x + \frac{b}{2a}$. In the other parts, use a method of your choice.
   (a) $x^2 - 2x - 35$
   (b) $-x^2 + x + 12$
   (c) $4x^2 - 40x + 91$
   (d) $x^2 - 4x + 1$
   (e) $x^2 - 14x + 52$
   (f) $4x^2 + 12x + 9$
   (g) $x^2 - 10$
   (h) $x^2 + 14x$
3. Use your solutions to Exercise 2 to find the solutions (in \( \mathbb{C} \)) of the equation \( p(x) = 0 \), for each polynomial \( p(x) \) in Exercise 2.

4. In learning certain calculational techniques (such as completing the square), often students do not exactly follow the template that the teacher presents. It is then the teacher’s job to decide whether a student’s approach is acceptable. Below, some hypothetical student work related to completing the square is presented. Comment on the degree to which each student’s work is correct and list any suggestions you would make to the student.

(a) \[
3x^2 + 10x - 8 = 3(x^2 + 10x - 8) = 3((x + 5)^2 - 8) - 25 = 3(x + 5)^2 - 8 - 25
= 3(x + 5)^2 - 33
\]

(b) \[
3x^2 + 10x - 8 = \left( x^2 + \frac{10}{3}x - \frac{8}{3} \right) = \left( x^2 + \frac{10}{3}x + \left( \frac{5}{3} \right)^2 - \left( \frac{5}{3} \right)^2 - \frac{8}{3} \right)
= \left( x^2 + \frac{10}{3}x + \left( \frac{5}{3} \right)^2 - \frac{25}{9} - \frac{24}{9} \right) = \left( (x + \frac{5}{3})^2 - \frac{49}{9} \right) = 3 \left( (x + \frac{5}{3})^2 - \frac{49}{9} \right)
= 3 \left( x + \frac{5}{3} \right)^2 - \frac{49}{3}
\]

(c) \[
3x^2 + 10x - 8
= \left( x^2 + \frac{10}{3}x - \frac{8}{3} \right)
= \left( x^2 + \frac{10}{3}x + \left( \frac{5}{3} \right)^2 - \left( \frac{5}{3} \right)^2 - \frac{8}{3} \right)
= \left( x^2 + \frac{10}{3}x + \left( \frac{5}{3} \right)^2 - \frac{25}{9} - \frac{24}{9} \right)
= \left( (x + \frac{5}{3})^2 - \frac{49}{9} \right)
= 3 \left( (x + \frac{5}{3})^2 - \frac{49}{9} \right)
= 3 \left( x + \frac{5}{3} \right)^2 - \frac{49}{3}
\]
5. Obtain the equation \( ax^2 + bx + c = a \left( x + \frac{b}{2a} \right)^2 + \left( c - \frac{b^2}{4a} \right) \) by making the substitution \( z = x + \frac{b}{2a} \). (This generalizes Equation (1).)

6. Completing the square is used in many contexts in analytic geometry, for example, in many calculations involving conic sections such as ellipses, hyperbolas, and circles. Here, we concentrate on circles. Given a point \((x_0, y_0)\) in the plane and a positive number \(r\), the circle with center \((x_0, y_0)\) and radius \(r\) is defined to be the collection of all points in the plane whose distance to \((x_0, y_0)\) is \(r\).

   (a) Find the equation of the circle with center \((1, 3)\) and radius 7. Your equation should be of the form \( x^2 + ax + y^2 + by = c \).

   (b) Find the equation of the circle with center \((x_0, y_0)\) and radius \(r\).

   (c) By completing the square, find the center and radius of the circle with equation \( x^2 - 6x + y^2 + 3y = 100 \).

   (d) By completing the square, find the center and radius of the circle with equation \( x^2 + ax + y^2 + by = c \).

   (e) What must be true about \(a, b, c\) for the equation \( x^2 + ax + y^2 + by = c \) to have any solutions? (Look carefully at your answer to the previous part.)

   (f) Prove that if \((x_0, y_0)\) is a point on the circle \( x^2 + y^2 = r^2 \), then the radial segment from \((0, 0)\) to \((x_0, y_0)\) is perpendicular to the tangent line to the circle at \((x_0, y_0)\).

3. Quadratic Polynomials: Graphs, Roots, and Factors

In this section, we investigate graphing quadratic polynomials, and explore the relationships between roots and factors of a quadratic polynomial.

3.1. Graphing transformations and parabolas.

**Let’s Go 6.** As part of a task in your algebra class, students are required to describe the main features of the graph of \( y = 2x^2 + 44x + 193 \). One student produces the graph shown in Figure 7 and says it clearly shows a line. Describe at least three ways you might help him discover his error.

In Let’s Go 6, some of you may have pointed out that all quadratic polynomials have parabolic graphs. But how do we know this? After all, in Let’s Go 3 we discovered that not all \(U\)-shaped graphs are necessarily
parabolas. Fortunately, the issue can be resolved by completing the square: Recall from Theorem 2 that graphing \( y = ax^2 + bx + c \) is the same (via completing the square) as graphing

\[
y = a \left( x + \frac{b}{2a} \right)^2 + \left( c - \frac{b^2}{4a} \right),
\]

which is obtained from the graph of \( y = ax^2 \) by the following sequence of standard graphing transformations:

- Shift horizontally by \( \left| \frac{b}{2a} \right| \) units.
- Shift vertically by \( \left| c - \frac{b^2}{4a} \right| \) units.\(^{10}\)

Therefore the graph of \( y = ax^2 + bx + c \) must be a parabola since it is obtained by sliding the known parabola \( y = ax^2 \) elsewhere in the plane.\(^{11}\) Moreover, one can show (see Your Turn 10, Exercise 21, and Exercise 25) that the new parabola has focus \( \left( -\frac{b}{2a}, c - \frac{b^2}{4a} \right) \), directrix \( y = -\frac{1}{4a} + c - \frac{b^2}{4a} \), and vertex \( \left( -\frac{b}{2a}, c - \frac{b^2}{4a} \right) \).

**Your Turn 10.** Let \( a, b, c \) be real numbers, with \( a \neq 0 \).

(a) Use completing the square together with graphing transformations to find the vertex of \( y = 2x^2 + 44x + 193 \). (Hint: What is the vertex of \( y = 2x^2 \)?)

(b) Use completing the square together with graphing transformations to show that \( \left( -\frac{b}{2a}, c - \frac{b^2}{4a} \right) \) is the vertex of \( y = ax^2 + bx + c \).

(c) Use completing the square to find the \( x \)-intercepts of \( y = 2x^2 + 44x + 193 \).

\(^{10}\)Of course, the direction of any horizontal or vertical shift will depend on the signs of \( \frac{b}{2a} \) and \( c - \frac{b^2}{4a} \), respectively.

\(^{11}\)Translation in the plane is an example of an *isometry*—a mapping that preserves angles and distances.
3.2. Quadratic formula, the nature of roots, and the be-plane. In Your Turn 8, you discovered the quadratic formula by completing the square. The formula says that the solutions \( r_1 \) and \( r_2 \) of \( ax^2 + bx + c = 0 \) (i.e., the roots of \( ax^2 + bx + c \)) are

\[
r_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad r_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.
\]

Since the expressions for \( r_1 \) and \( r_2 \) involve square roots of \( b^2 - 4ac \), it follows that the nature of \( r_1, r_2 \) will be determined by \( b^2 - 4ac \), which is usually called the discriminant of \( ax^2 + bx + c \). Assuming (as usual) that \( a, b, c \) are real numbers, the following table summarizes the situation:

<table>
<thead>
<tr>
<th>Sign of discriminant</th>
<th>Result</th>
<th>Terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b^2 - 4ac &gt; 0 )</td>
<td>( r_1, r_2 ) are real</td>
<td>( ax^2 + bx + c ) has two distinct real roots</td>
</tr>
<tr>
<td>( b^2 - 4ac = 0 )</td>
<td>( r_1, r_2 ) are real and equal</td>
<td>( ax^2 + bx + c ) has a root of multiplicity two</td>
</tr>
<tr>
<td>( b^2 - 4ac &lt; 0 )</td>
<td>( r_1, r_2 ) are nonreal complex conjugates</td>
<td>( ax^2 + bx + c ) has conjugate roots</td>
</tr>
</tbody>
</table>

For monic quadratic polynomials (that is, those that can be written as \( x^2 + bx + c \)) the situation can be visualized in the plane by identifying \( x^2 + bx + c \) with the point \((b,c)\). (See part (a) of Figure 8.)\(^\text{12}\) Using this identification together with the conditions on the \( b^2 - 4ac \) described above, the nature of the roots can be plainly seen: those with real roots lie below the parabola \( b^2 = 4c \), those with non-real complex roots lie above \( b^2 = 4c \), and those with a root of multiplicity two (sometimes called double roots) lie on \( b^2 = 4c \). (See part (b) of Figure 8.)

\(^\text{12}\)This is similar to the identification of non-vertical lines \( y = mx + b \) with points \((m,b)\) given in Section 7 of Chapter 2.
3. QUADRATIC POLYNOMIALS: GRAPHS, ROOTS, AND FACTORS

(a) What information do the solutions of $ax^2 + bx + c = 0$ provide about the graph of $y = ax^2 + bx + c$?

(b) Construct and graph three specific quadratic polynomials possessing negative, zero, and positive discriminant, respectively.

(c) What does the sign of the discriminant tell us about the graph of $y = ax^2 + bx + c$? (Part (b) will be helpful.)

(d) In the case where the discriminant $b^2 - 4ac$ is negative, why are the solutions of $ax^2 + bx + c = 0$ complex conjugates?

3.3. Factoring quadratic polynomials. In school, before we mastered the quadratic formula or completing the square, we learned how to factor certain quadratic polynomials, that is, to write a quadratic polynomial as a product of two degree-one polynomials. This was generally a trial-and-error exercise. For example, if we were asked to factor $x^2 - 7x + 12$, we experimented with pairs of integers whose product was 12 and were happy to discover that $(x - 4)(x - 3)$ “worked,” meaning that the product of $(x - 4)$ and $(x - 3)$ indeed is $x^2 - 7x + 12$.

Most quadratics can’t be factored by this method. For example, we can check that $x^2 + 6x + 2 = (x + (3 + \sqrt{7})) (x + (3 - \sqrt{7}))$ by multiplying out the right hand side, but it’s unlikely that we could obtain such a factorization by trial-and-error. Only later in school, after we learned about completing the square and the quadratic formula, did we reach the point where we could factor polynomials like $x^2 + 6x + 2$.

To understand why, we need to explore the relationship between the roots and the factors of a quadratic polynomial.13

Your Turn 12. Consider a quadratic polynomial that can be factored as $a(x - r_1)(x - r_2)$.

(a) Explain why $r_1$ and $r_2$ are roots. (What is the definition of root?)

(b) Explain why if $s \neq r_1$ and $s \neq r_2$, then $s$ cannot be a root of the polynomial.

Your Turn 12 tells us that if a quadratic polynomial can be written in factored form, then the roots can be identified immediately. We now consider the converse: if we know the roots (or merely one root) of a quadratic polynomial, can we obtain a factorization of the polynomial? The answer is yes:

13Many of the ideas in this section will be explored in greater generality in Chapter 12.
Your Turn 13. Consider a quadratic polynomial $ax^2 + bx + c$ (with $a \neq 0$) and suppose that $r_1$ is known to be a root of the polynomial. We can perform the algorithm called long division of polynomials to $ax^2 + bx + c$ and $x - r_1$, obtaining a quotient $Q(x)$ and a remainder $R(x)$, which satisfy $ax^2 + bx + c = (x - r_1) \cdot Q(x) + R(x)$.

(a) In the case we have described (division of a degree-two polynomial by a degree-one polynomial), explain why $R(x)$ must be a constant.

(b) Perform the long division. (You should discover that $Q(x) = ax + (ar_1 + b)$ and $R(x) = ar_1^2 + br_1 + c$.)

(c) Given that $r_1$ is a root of $ax^2 + bx + c$, explain why $R(x)$ must be zero. (Give two explanations, one using the equation $ax^2 + bx + c = (x - r_1) \cdot Q(x) + R(x)$, and the other using the equation $R(x) = ar_1^2 + br_1 + c$.)

(d) Since $R(x) = 0$, we obtain $ax^2 + bx + c = (x - r_1) \cdot Q(x) = (x - r_1) \cdot (ax + (ar_1 + b))$. Show that this can be rewritten in the form $a(x - r_1)(x - r_2)$. (Your answer will give you an expression for $r_2$. Note that by Your Turn 12, $r_2$ is also a root of $ax^2 + bx + c$.)

Your Turns 12 and 13 give a complete description of the relationship between roots and factors of a quadratic polynomial.

Now let’s consider how the coefficients of the polynomial depend on the roots:

Your Turn 14.

(a) Find expressions$^{14}$ for the coefficients of the quadratic polynomial $a(x - r_1)(x - r_2)$ in terms of $a$, $r_1$, and $r_2$.

(b) Consider the first example from this section: factoring $x^2 - 7x + 12$. According to (a), what pair of conditions has to be true about the roots of the polynomial? Is this what you were thinking when you first learned to factor quadratics?

(c) In writing an exam for your algebra class, you want to include a quadratic polynomial $f$ with roots $2, -7$ and $f(0) = 28$. How can you use part (a) to simplify your task?

---

$^{14}$This is Viète’s Theorem in the case of quadratic polynomials. In the general case, Viète’s Theorem tells us that the coefficients of any polynomial can be expressed as a function of its roots. Francois Viète (1540-1603) was a pioneer in symbolic algebra (using symbols and formulas to express ideas and make computations as opposed to using ordinary language).
Conversely, how do the roots of a quadratic polynomial depend on the coefficients? This is precisely the content of the Quadratic Formula, which produces the roots \( r_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \) and \( r_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \) in terms of the coefficients!

**Your Turn 15.** Take the expressions for \( r_1 \) and \( r_2 \) from the quadratic formula, and substitute them into the “expressions for the coefficients” that you obtained in Your Turn 14(a). Is the answer surprising?

Finally, we consider an issue that may confuse many high school students:

**Your Turn 16.**

(a) My math teacher has a magic algebra problem: Factor \( x^2 - 2x + 2 \). The problem is magic because the answer depends on the time of year the question is being asked! In the fall the answer is ‘cannot be factored.’ In the spring, the answer is \( (x - (1 + i))(x - (1 - i)) \). Can you explain what is happening? How can you adjust the statement of the magic problem so that it has precisely one correct answer?

(b) How is the problem of factoring \( x^2 + 6x + 2 \) also a “magic problem”? How is it related to the issue you discussed in (a)? (This time, the key issue has nothing to do with complex numbers. What is the key issue?)

3.4. Exercises. Completing the square, finding roots and vertex, factoring, and graphing

1. Without using technology, for each of the following quadratic polynomials, labeled 1–10 and given below:
   (a) Factor the polynomial, if you can do so quickly.
   (b) Complete the square.
   (c) Use your work in completing the square to find the vertex of the parabola (x-and y-coordinates) and the roots.
   (d) Use \(-\frac{b}{2a}\) to double-check the x-coordinate of the vertex.
   (e) Use the quadratic formula to double-check the roots.
   (f) Using the roots of the polynomial, factor the polynomial.
(g) Multiply out your factorization and check that you obtain the original polynomial.

(h) Find the $y$-intercept of the graph.

(i) Graph the polynomial.

1. $x^2 + 5x$
2. $x^2 - 8$
3. $4x^2 + 20x$
4. $x^2 - 3x - 70$
5. $12x^2 + 8x - 15$
6. $x^2 - 6x + 26$
7. $16x^2 - 40x + 33$
8. $x^2 + 8$
9. $x^2 + 6x + 58$
10. $25x^2 - 40x + 23$

2. Examine your solution to the previous exercise. What differences emerged among the examples?

*Complex, real, rational, and integer roots*

3. Without computing roots, determine whether each polynomial has two distinct real roots, a root of multiplicity two, or two conjugate roots.

1. $x^2 + 2x - 63$
2. $6x^2 - 31x + 18$
3. $x^2 + 4x - 6$
4. $x^2 - 6x + 58$
5. $5x^2 + 9x$
6. $7x^2 - 3$
7. $\pi x^2 + ex + 7$
8. $x^2 + 2\sqrt{10}x + 10$
4. We consider \( ax^2 + bx + c \), with \( a, b, c \in \mathbb{R} \) and \( a \neq 0 \).

(a) Is every \( x \)-intercept of the graph a root of the polynomial?

(b) Is every root of the polynomial an \( x \)-intercept of the graph?

5. The purpose of this exercise is to give an alternate approach to studying the nature of the roots of \( ax^2 + bx + c \).

(a) Suppose that \( r_1 \) and \( r_2 \) are real numbers. Prove (by explicit computation) that the coefficients \( a, b, c \) of the polynomial \( a(x - r_1)(x - r_2) \) satisfy \( b^2 - 4ac \geq 0 \). \( \text{(This shows that if the roots of the polynomial are real, then } b^2 - 4ac \geq 0. \text{)} \)

(b) Suppose that \( r \) and \( s \) are real numbers. Prove (by explicit computation) that the coefficients \( a, b, c \) of the polynomial \( a(x - (r + si))(x - (r - si)) \) satisfy \( b^2 - 4ac \leq 0 \). \( \text{(This shows that if the roots of the polynomial are a pair of complex conjugates, then } b^2 - 4ac \leq 0. \text{)} \)

(c) Suppose that \( r \) is any number. Prove (by explicit computation) that the coefficients \( a, b, c \) of the polynomial \( a(x - r)(x - r) \) satisfy \( b^2 - 4ac = 0 \). \( \text{(This shows that if polynomial has a real root of multiplicity two, then } b^2 - 4ac = 0. \text{)} \)

6. Consider the quadratic polynomial \( x^2 + bx + 12 \).

(a) For what choices of an integer \( b \) does this polynomial have integer roots? Justify your answer. \( \text{(This should help convince you that integer roots of polynomials are uncommon.)} \)

(b) Assuming that \( b \) is an integer, show that a rational root of \( x^2 + bx + 12 \) is necessarily an integer. \( \text{(This should convince you that rational roots of polynomials are uncommon. You will need some facts from number theory to solve this problem.)} \)

7. Let \( a, b, c \) be integers, with \( a \neq 0 \). Let \( r_1 \) and \( r_2 \) be the roots of \( ax^2 + bx + c \).

(a) Show that if \( r_1 \) is rational, then so is \( r_2 \).

(b) Show that if a root is rational, then it can be written as \( \frac{p}{q} \), where \( p, q \) are integers, \( q \) divides \( a \), and \( p \) divides \( c \). \( \text{(This is the Rational Root Theorem for quadratic polynomials. You will need some facts from number theory to solve this problem.)} \)
Computing roots of quadratic polynomials

8. Find the roots of $8x^2 - 20x + 37$ in simplest possible form by:

(a) using the quadratic formula

(b) completing the square.

Which method provides the easiest path to the simplified roots? Is this surprising?

9. Find all the solutions in the real numbers of the equation $7x^6 + 8x^3 + 2 = 0$ for $x$. (How did this problem sneak into a chapter about quadratic polynomials?)

10. Consider $ax^2 + bx + c$, with $a \neq 0$, but allowing $a, b, c$ to be complex numbers. Does the quadratic formula give roots of the polynomial? (You should interpret this question to mean, “Can the expressions for the roots, coming from the quadratic formula, be interpreted as complex numbers?”) Discuss.

The relationship between roots and factors

11. Find the monic quadratic polynomial with roots $r_1 = \frac{-4 + \sqrt{15}}{3}$ and $r_2 = \frac{-4 - \sqrt{15}}{3}$.

12. Find a quadratic polynomial with integer coefficients, whose roots are

(a) 1 and $-5$

(b) $2 \pm 3i$

(c) $-5 \pm \sqrt{7}$

(d) $\frac{1}{2}$ and $\frac{5}{4}$

(e) $\frac{3}{2} \pm \sqrt{\frac{2}{11}}$

13. Find the monic quadratic polynomial with real coefficients that has $5 - 7i$ as one of its roots.

14. Find the unique quadratic polynomial $p(x)$ such that all three of the following conditions are satisfied:

- The coefficient of $x^2$ is 4.
- All of the coefficients are integers.
- $7 + \sqrt{6}$ is one of the roots.
15. Suppose $r_1 = i$ and $r_2 = 3$. What monic ‘quadratic polynomial’ has these roots? Does this violate the fact that non-real roots of quadratic polynomials must appear in conjugate pairs? Explain

16. An algebra student responds to the questions “Factor $x^2 - 5x + 6$” with the answer “$x = 2$ or $x = 3$.” Has the student correctly answered the question? Discuss.

17. In solving $x^2 - 5x + 6 = 0$ by factoring, we use the fact that $(x - 2)(x - 3) = 0$ implies $x - 2 = 0$ or $x - 3 = 0$.

(a) What property of complex numbers ensures this?

(b) Does this property hold for the set of $2 \times 2$ matrices with real entries?

The graph of a quadratic polynomial

18. Show that the graph of $ax^2 + bx + c$ is symmetric about the vertical line $x = -\frac{b}{2a}$.

19. Assume that $a > 0$. Let $f(x)$ be the function whose graph is the “left half” of the parabola $y = ax^2 + bx + c$ (the points on the graph with $x \leq -\frac{b}{2a}$). In terms of $a$, $h$, and $k$, give the following:

(a) the domain of $f$

(b) the range of $f$

(c) an expression for the inverse of $f$.

20. Theorem 2 may be interpreted as

$$ax^2 + bx + c = a(x + h)^2 + k.$$ 

Give conditions on $a$, $h$ and $k$ that are equivalent to the statement that the polynomial has real roots. (How is the graph of $a(x + h)^2 + k$ obtained from the graph of $ax^2$ by graphing transformations involving $h$ and $k$?)

21. Let $f(x) = ax^2 + bx + c$ with $a \neq 0$. As discussed in the text, the graph of $f$ is a parabola. Show that the focus is the point $\left( -\frac{b}{2a}, \frac{1}{4a} + \left( c - \frac{b^2}{4a} \right) \right)$, and the directrix is the line with equation $y = -\frac{1}{4a} + c - \frac{b^2}{4a}$.

(Use Equation 2 from this section, along with Exercise 3 from Section 1.)
22. Using Geometer’s Sketchpad, define a parameter $t$. Then animate the graph of $f(x) = 0.5x^2 + tx + 2$ and watch the path of the vertex.

(a) What kind of curve do you think is being traced out by the vertex of $y = f(x)$ as $t$ changes?
(b) Prove or disprove your guess from part (a) by finding a formula for the curve traced out by the vertex of $y = f(x)$ as the parameter $t$ changes.
(c) Find a formula for the curve traced out by the vertex of $g(x) = ax^2 + tx + c$ as the parameter $t$ changes.

Families of quadratic polynomials

23. We may visualize various families of monic quadratic polynomials by identifying each quadratic polynomial with a point in the $bc$-plane. Describe (in words or by a sketch in the $bc$-plane) the collection of all monic, degree-two polynomials $x^2 + bx + c$ satisfying the given conditions. (Your answers will be in terms of $b$ and $c$.)

(a) The $y$-intercept is positive.
(b) The vertex is on the $x$-axis.
(c) The vertex is on the line $x = 3$.
(d) The vertex is on the line $y = 7$.
(e) The vertex is $(3, 7)$.
(f) The roots are 5 and 8.
(g) One of the roots is 5.
(h) One of the roots is zero.
(i) The roots are real, and both are between 6 and 7.
(j) The roots are real, and their distance apart is no more than 5.
(k) The imaginary part of each root is between $-3$ and 3.
(l) The vertex is in the first quadrant.

Calculus and quadratic polynomials

15 A similar idea was explored in Section 7 of Chapter 2.
24. Let \( f(x) = ax^2 + bx + c \) be a quadratic polynomial. Show that the discriminant \( b^2 - 4ac \) is equal to zero if and only if \( f(-\frac{b}{2a}) = f'(-\frac{b}{2a}) = 0 \).

25. Let \( a > 0 \). Show that the extremum of \( f(x) = ax^2 + bx + c \) occurs at \( x = -\frac{b}{2a} \) by

(a) Using calculus.

(b) Completing the square and finding the value of \( x \) that maximizes (or minimizes) the function. (Think about adding or subtracting a squared expression, which must be nonnegative.)

(c) Averaging the roots of \( ax^2 + bx + c \).

(d) By averaging any two values of \( x \) for which the values of \( f(x) \) are equal.

26. Let \( f(x) = ax^2 + bx + c \) with \( a \neq 0 \). Use calculus to find the intervals on which \( f \) is increasing/decreasing. Then, use calculus to discuss the concavity of the graph of \( f \). (Your answer will depend on the sign of \( a \).)

27. From calculus, recall the Mean Value Theorem: Suppose that \( f(x) \) is continuous on the closed interval \([x_1, x_2]\) and differentiable on the open interval \((x_1, x_2)\). Then there exists \( x_3 \in (x_1, x_2) \) such that

\[
\frac{f(x_2) - f(x_1)}{x_2 - x_1} = f'(x_3).
\]

(a) What is the geometric interpretation of the Mean Value Theorem? (Think about slope.)

(b) Suppose that \( f(x) = ax^2 + bx + c \) with \( a \neq 0 \). Show that for any \( x_1 \) and \( x_2 \), the unique value of \( x_3 \) that satisfies the Mean Value Theorem is \( \frac{x_1 + x_2}{2} \) (the midpoint of \([x_1, x_2]\)).

28. Let \( f(x) = ax^2 + bx + c \) with \( a \neq 0 \). Find \( \lim_{x \to \infty} f(x) \) and \( \lim_{x \to -\infty} f(x) \). (Your answer will depend on the sign of \( a \).)

4. Application: Quadratic Splines

Some knowledge of linear algebra is a prerequisite for this section.

The smooth lines of a modern dashboard may be modeled mathematically by “smoothly” joining together pieces of simple curves, such as graphs of polynomial functions. The end result of this joining is a curve
called a spline. In this section, we will consider quadratic splines, which are formed by “smoothly joining” graphs of quadratic polynomials.

4.1. Fitting a quadratic to three points. Building a quadratic spline requires us to ‘fit’ the graphs of quadratic polynomials to existing data points in the plane. We begin with a warm-up problem: finding a quadratic polynomial $ax^2 + bx + c$ whose graph passes through three data points $(x_1, y_1), (x_2, y_2), (x_3, y_3)$, no two of which lie on the same vertical line (see Figure 9). To solve this problem, we must determine values for $a$, $b$, and $c$.

Since each data point is on the graph of $ax^2 + bx + c$, the following equations hold:

$$ax_1^2 + bx_1 + c = y_1$$
$$ax_2^2 + bx_2 + c = y_2$$
$$ax_3^2 + bx_3 + c = y_3$$

This system of equations can then be translated into the matrix equation

$$
\begin{pmatrix}
  x_1^2 & x_1 & 1 \\
  x_2^2 & x_2 & 1 \\
  x_3^2 & x_3 & 1 \\
\end{pmatrix}
\begin{pmatrix}
  a \\
  b \\
  c \\
\end{pmatrix} =
\begin{pmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
\end{pmatrix},
$$

which we want to solve for $a$, $b$, and $c$. From linear algebra, we know that the equation (3) has a unique solution exactly when the determinant of the matrix in the lefthand side of Equation (3) is nonzero. A computation (Exercise 2) shows that this determinant is equal to $(x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$, and this is

\[\text{In this section, we must allow the possibility that } a = 0 \text{ when we consider polynomials } ax^2 + bx + c.\]
nonzero since it was assumed that no two of the points \((x_1, y_1), (x_2, y_2), (x_3, y_3)\) lie on the same vertical line. So, we have proved the following theorem:

**Theorem 3.** Let \((x_1, y_1), (x_2, y_2), (x_3, y_3)\) be points in the plane, no two of them on the same vertical line. Then there exist unique real numbers \(a, b, c\), such that the points \((x_i, y_i)\) all lie on the graph of \(y = ax^2 + bx + c\).

The notion of a *spline*, however, requires a slightly different notion of “fitting a polynomial to points.” The following Your Turn introduces the key idea:

**Your Turn 17.** Let \((x_1, y_1)\) and \((x_2, y_2)\) be points in the plane with \(x_1 \neq x_2\).

(a) How many quadratic polynomials have graphs passing through both \((x_1, y_1)\) and \((x_2, y_2)\)? Give algebraic as well as geometric illustrations.

(b) How many quadratic polynomials \(f\) have \(f'(x_1) = 5\) and have graphs passing through the two points?

### 4.2. Fitting Quadratics to More Points: Quadratic Splines

To begin, consider a collection of data points \((x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\) in the plane with \(x_0 < x_1 < \cdots < x_n\). On each interval \([x_{i-1}, x_i]\) for \(1 \leq i \leq n\), construct a quadratic polynomial \(f_i\) with \(f_i(x_{i-1}) = y_{i-1}\) and \(f_i(x_i) = y_i\). The end result is a “piecewise quadratic function” \(F : [x_0, x_n] \to \mathbb{R}\) defined by \(F(x) = f_i(x)\) whenever \(x \in [x_{i-1}, x_i]\). The graph of \(F\), which is a compilation of the graphs of the \(f_i\), passes through each data point\(^{17}\) (see Figure 10).

Of course, when we look at Figure 10, we realize that the function \(F(x)\) we have constructed may exhibit some sharp corners and thus may fail to be “smooth”. From calculus, we know that one way to impose ‘smoothness’ on a function is to require it to be differentiable at all points.\(^{18}\) Since quadratic functions are differentiable on their domains, the only differentiability problems occur at places where we join two quadratic functions together. Thus, by requiring that \(f'_{i-1}(x_{i-1}) = f'_i(x_{i-1})\) for \(2 \leq i \leq n\), we force \(F\) to be differentiable at all points of its domain (see Figure 11). In fact, as you may have discovered in Your Turn 17, this condition nearly forces \(F\) to be unique:

\(^{17}\)Your Turn 17 tells us that the polynomials \(f_i\) are not uniquely determined, so there are many ways to form the function \(F\). This is a good thing, since we will need to adjust the \(f_i\)’s to make the graph of \(F\) “smooth.”

\(^{18}\)To most mathematicians, a function is defined to be smooth if it is infinitely differentiable. For us, however, we are using the word smooth as an informal descriptor of a differentiable function.
Theorem 4. Let data points $(x_0, y_0), \ldots, (x_n, y_n)$ and $\gamma \in \mathbb{R}$ be given, with $x_0 < x_1 < \cdots < x_n$. There is a unique set of quadratic polynomials $\{f_1, \ldots, f_n\}$ satisfying the following conditions:

- $f_i : [x_{i-1}, x_i] \rightarrow \mathbb{R}$.
- $f_i(x_{i-1}) = y_{i-1}$ and $f_i(x_i) = y_i$ for $1 \leq i \leq n$. (This ensures the spline will pass through the data points.)
- $f'_1(x_0) = \gamma$. (This specifies the initial “direction” of the spline.)
- $f'_{i-1}(x_{i-1}) = f'_i(x_{i-1})$ for $2 \leq i \leq n$. (This ensures that the spline will be ‘smooth.’)

The polynomials $f_1, \ldots, f_n$ determine a ‘smooth’ quadratic spline $F : [x_0, x_n] \rightarrow \mathbb{R}$ with $F(x) = f_i(x)$ for $x \in [x_{i-1}, x_i]$.
We verify Theorem 4 using a finite induction argument. First we show that $f_1$ is uniquely determined by the hypotheses, and then we show that for $2 \leq i \leq n$, if $f_{i-1}$ exists and satisfies all of the necessary conditions, then $f_i$ is uniquely determined by $f_{i-1}$.

We first handle $f_1(x)$, which we may write as $a_1x^2 + b_1x + c_1$ for some real numbers $a_1, b_1, c_1$. Since $f_1(x_0) = y_0$, $f_1(x_1) = y_1$, and $f'_1(x_0) = \gamma$, we have

$$a_1x_0^2 + b_1x_0 + c_1 = y_0$$
$$a_1x_1^2 + b_1x_1 + c_1 = y_1$$
$$2a_1x_0 + b_1 = \gamma,$$

which yields the matrix equation

$$\begin{pmatrix}
x_0^2 & x_0 & 1
\end{pmatrix}
\begin{pmatrix}
a_1 \\
b_1 \\
c_1
\end{pmatrix}
= 
\begin{pmatrix}
y_0 \\
y_1 \\
\gamma
\end{pmatrix}.$$

This equation will have a unique solution (i.e., will yield unique values for $a_1, b_1, c_1$ determining $f_1(x)$) if and only if

$$\det \begin{pmatrix}
x_0^2 & x_0 & 1 \\
x_1^2 & x_1 & 1 \\
2x_0 & 1 & 0
\end{pmatrix} \neq 0.$$

A quick computation (see Exercise 4) verifies this, so $f_1$ is uniquely determined. A similar argument will show that for $2 \leq i \leq n$, $f_i(x)$ is uniquely determined given that $f_{i-1}$ exists and satisfies all of the hypotheses of Theorem 4 (see Exercise 10).

4.3. Exercises.

1. In the notation of Theorem 3, explain what is special about the configuration of the points $(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$ when $a = 0$. What is special about the configuration of the three points when both $a$ and $b$ are zero?

2. Show that $\det \begin{pmatrix}
x_3^2 & x_1 & 1 \\
x_3^2 & x_2 & 1 \\
x_3^2 & x_3 & 1
\end{pmatrix} = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$.

3. Find the quadratic polynomial $ax^2 + bx + c$ whose graph contains the points $(1, 9)$, $(-2, 27)$, and $(4, 3)$. Then graph the three points and the polynomial.
4. Suppose $x_0 \neq x_1$. Compute

$$\det \begin{pmatrix} x_0^2 & x_0 & 1 \\ x_1^2 & x_1 & 1 \\ 2x_0 & 1 & 0 \end{pmatrix}$$

and show it is nonzero.

5. Find a smooth quadratic spline $F(x)$ passing through the points $(1, 1)$, $(2, 4)$, $(3, 5)$, and $(4, 5)$ satisfying $F'(1) = 1$. (This will involve solving three systems of three linear equations in three unknowns—so be careful! You might want to make your calculator or or a computer algebra system to do some of the work for you.)

6. Let $(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$ be three non-colinear points in the plane. Prove that there exist unique real numbers $a, b, c$ such that each point $(x_i, y_i)$ satisfies the equation $x^2 + ax + y^2 + by = c$. (The equation is the equation for a circle. Thus, this exercise shows that any three non-colinear points in the plane lie on exactly one circle.)

7. As an illustration of Exercise 6, find the equation of the unique circle containing the points $(0, 0)$, $(1, 0)$, and $(0, 2)$.

8. Consider the data points and quadratic polynomials described in Theorem 4, with $f'_{i-1}(x_{i-1}) = f'_i(x_{i-1})$ for $2 \leq i \leq n$. Verify that $F'(x_i)$ exists. (Hint: Consider the limit of \( F(x_i + h) - F(x_i) \)

9. Revisit Simpson’s Rule from your calculus book. Explain what Simpson’s Rule is used for, and where Theorem 3 is used in the proof of Simpson’s Rule.

10. Complete the proof of Theorem 4 by showing that $f_i(x)$ is uniquely determined by the existence of $f_{i-1}(x)$ satisfying the hypotheses of the theorem.

11. Find a smooth quadratic spline $F(x)$ passing through the points $(1, 2)$, $(2, -1)$, and $(3, 3)$ satisfying $F'(2) = 0$. (This doesn’t quite fit Theorem 4. See what you can do!)
12. Can you extend the statement of Theorem 3 to the case when there are four data points? To \( n \) data points? (Hint: To establish a pattern, it may be helpful to recall that there is a unique line \( y = bx + c \) passing through any two points points in the plane with distinct \( x \)-coordinates.)

13. It is possible to generalize Theorem 4 to the case of cubic (degree-three) polynomials (and even to polynomials of arbitrary degree \( n \)). Make a conjecture about how cubic polynomials give a further way to “fit the spline to the points.” Verify your conjecture.

5. Application: Problems Involving Tangency

Tangency, which arises in many geometric problems, is a concept that can be appreciated from the point of view of algebra, calculus and geometry. In this section we investigate two tangency problems involving quadratic polynomials from a variety of perspectives. Both problems admit algebraic solutions in which the quadratic formula is prominently featured.\(^{19}\) Geometric solutions depend on the fact that the line tangent to a circle at a point \( P \) is always perpendicular to the radial segment joining the center of the circle to \( P \) (see the first part of Figure 12, and Exercises 3 and 4).

---

\(^{19}\)This is why these problems are placed in this chapter.
5.1. A circle tangent to a parabola. We investigate two methods of solving the following problem (see Figure 13):

PROBLEM 1. What is the radius of the circle, centered at $(0, 1)$, that is tangent to the graph of $y = x^2$?

First, there is a solution to Problem 1 using calculus: Let $(x_0, x_0^2)$ be the point in the first quadrant where the desired circle is tangent to the parabola, as shown in Figure 13. On one hand, the slope of the tangent line to the parabola at this point is $2x_0$ (since $\frac{d}{dx}x^2 = 2x$). On the other hand, the tangent line is also tangent to the circle, and therefore must be perpendicular to the radial segment joining $(0, 1)$ and $(x_0, x_0^2)$. Due to perpendicularity, we conclude that product of the slope of the radial segment and the slope of the tangent line must equal $-1$, giving $2x_0 \cdot \frac{x_0^2 - 1}{x_0} = -1$. Solving yields $x_0 = \frac{1}{\sqrt{2}}$, so the point of tangency is $\left(\frac{1}{\sqrt{2}}, \frac{1}{2}\right)$. Finally, the desired radius is the distance between $(0, 1)$ and $\left(\frac{1}{\sqrt{2}}, \frac{1}{2}\right)$, namely $\sqrt{\frac{3}{2}}$.

Second, we consider a non-calculus algebraic solution: Imagine a family of circles centered at $(0, 1)$ of various radii (see Figure 14). Geometrically, we see that if the radius is “too small” (smaller than the circle that is tangent), then the circle and parabola do not intersect. On the other hand, if the circle is a little “too big,” then the circle intersects the parabola in four points. When the circle is “just right” (tangent to the parabola), then there are precisely two points of intersection, which are $(x_0, x_0^2)$ and $(-x_0, x_0^2)$. This agrees with what we find when we actually try to compute the points of intersection. In considering the

\[\text{If we take much bigger circles, suddenly we have three points of intersection (the vertex and two points high up on the parabola), and then just two points of intersection (very high up on the parabola).}\]
system \{y = x^2, x^2 + (y - 1)^2 = r^2\}, we can eliminate \(y\) and obtain \(x^2 + (x^2 - 1) = r^2\), which simplifies to \(x^4 - x^2 + (1 - r^2) = 0\). This is a quadratic in \(x^2\), so we can apply the quadratic formula to obtain

\[
x^2 = \frac{1 \pm \sqrt{1 - 4(1 - r^2)}}{2} = \frac{1 \pm \sqrt{4r^2 - 3}}{2}.
\]

From this, we conclude that

\[
(4)\quad x = \pm \sqrt{\frac{1 \pm \sqrt{4r^2 - 3}}{2}}.
\]

Your Turn 18. How does Equation (4) give the solution to Problem 1?

Your Turn 19. Upon a first glance at Equation (4), it looks as if there are four \(x\)-values, corresponding to four points of intersection between the circle and parabola. Geometrically, we know that this does not always happen—it depends on how big the radius is. How is this reflected algebraically in Equation (4)?

5.2. Dual circles. We investigate three methods of solving the following problem:

Problem 2. Consider the two circles of radius 1 centered at \((\pm 2, 0)\). Find equations for all lines that are tangent to both circles (see Figure 15).

From Figure 15 there seem to be four such lines: Two horizontal lines \(y = \pm 1\) and two other lines passing through the origin. In the following solutions, we principally focus our attention on the latter two lines.

First, there is a geometric solution: Let \((x_0, y_0)\) be the point of tangency indicated in Figure 15. The segment from \((2, 0)\) to \((x_0, y_0)\) is perpendicular to the segment from \((0, 0)\) to \((x_0, y_0)\), and so the product
of their slopes is $-1$. In summary, we have $\frac{y_0}{x_0-2} \cdot \frac{y_0}{x_0} = -1$, therefore $y_0^2 = 2x_0 - x_0^2$. On the other hand, since $(x_0, y_0)$ lies on the indicated circle, we have $(x_0 - 2)^2 + y_0^2 = 1$. We can now eliminate $y_0$, obtaining $2x_0 - x_0^2 = 1 - (x_0 - 2)^2$, which gives $x_0 = \frac{3}{2}$. Substituting into a previous equation (and observing $y_0 > 0$), we learn that $y_0 = \sqrt{3}$. Thus the equation of the line through $(x_0, y_0)$ is $y = x\sqrt{3}$. By symmetry, the other line is $y = -x\sqrt{3}$.

Second, for a calculus solution we realize the slope of the line through $(0, 0)$ and $(x_0, y_0)$ must equal the slope of the tangent line to the circle $(x - 2)^2 + y^2 = 1$ at $(x_0, y_0)$. The slope of the first line is $\frac{y_0}{x_0}$. By implicit differentiation, the slope of the tangent line is $2 - \frac{x_0 y_0}{2}$ (see Exercise 5). If we equate $\frac{y_0}{x_0}$ and $2 - \frac{x_0 y_0}{2}$, we obtain the equation $y_0^2 = 2x_0 - x_0^2$, and then we can solve for $x_0$ and $y_0$ as in the first (geometric) solution.

Finally, there is a completely different approach to the problem using algebra: To begin, we observe that a line is tangent to a circle exactly when they intersect in a single point. Now suppose that we have the line $y = mx + b$ and the circle $(x - 2)^2 + y^2 = 1$. To find where they intersect, we eliminate $y$ to obtain $(x - 2)^2 + (mx + b)^2 = 1$, which we can rewrite as the quadratic equation

\begin{equation}
(m^2 + 1)x^2 + (2mb - 4)x + (b^2 + 3) = 0.
\end{equation}

To achieve exactly one point of intersection, we want Equation 5 to have exactly one solution. Fortunately, we know that a quadratic equation $Ax^2 + By^2 + C = 0$ has exactly one solution when $B^2 = 4AC$, which, when applied to Equation (5), gives $(2mb - 4)^2 = 4(m^2 + 1)(b^2 + 3)$. This simplifies to

\begin{equation}
b^2 + 4mb + 3m^2 = 1.
\end{equation}
In summary, we have learned that the line \( y = mx + b \) is tangent to the circle \((x - 2)^2 + y^2 = 1\) when \( b^2 + 4mb + 3m^2 = 1 \). Applying the same process, we discover that \( y = mx + b \) is tangent to the other circle \((x + 2)^2 + y^2 = 1\) when \( b^2 - 4mb + 3m^2 = 1 \).

Your Turn 20. Consider the unfinished algebraic solution to Problem 2 above.

(a) What equations must be satisfied for \( y = mx + b \) to be tangent to both circles?

(b) Use the results of part (a) to find values for \( m \) and \( b \), and state the equations of the four tangent lines.

(c) The algebraic solution to the problem (which you just completed in part (b)) is superior to the other solutions because no a priori assumptions are made about the tangent lines. What assumptions about the tangent lines are being made in the geometric and calculus solutions?

5.3. Exercises. Exercises 1–1 discuss the concept of tangency, or address aspects of the Problems discussed in the text. Exercises 7–10 are new “Problems Involving Tangency,” similar to the Problems discussed in the text.

1. Let \( I \subset \mathbb{R} \) be an open interval, and suppose \( f, g : I \to \mathbb{R} \) are both differentiable at \( c \in I \). We say the graphs of \( f \) and \( g \) are tangent at \( c \) if \( f(c) = g(c) \) and \( f'(c) = g'(c) \). (Another way to say this is that the tangent lines (as defined in calculus) for \( f \) and \( g \) at \( c \) coincide.) Use this definition to show that \( f(x) = \frac{2}{\pi} \sin(\pi x) \) and \( g(x) = (x - 2)^2 - 1 \) are tangent at \( x = 1 \).

2. One may think that in order for a line \( \ell \) to be tangent to a curve \( C \) at a point \( P \), there must be some ‘neighborhood’ of \( P \) in which the only intersection of \( \ell \) and \( C \) occurs at \( P \). We show this is false by considering the curve

\[
f(x) = \begin{cases} 
  x^2 \sin(1/x) & x \neq 0, \\
  0 & x = 0
\end{cases}
\]

and the line \( g(x) = 0 \) at the origin.

(a) Show that the graphs of \( f \) and \( g \) are tangent at \( x = 0 \). (Use the definition of tangency given in Exercise 1. You’ll need to use the definition of the derivative to compute \( f'(0) \).)
(b) Show that in any open interval containing $x = 0$, there exists $t \neq 0$ with $f(t) = g(t) = 0$.

3. The definition of tangency given in Exercise 1 is rather limited for several reasons, including the fact that it doesn’t account for vertical tangents. A broader definition can be formulated in terms of vector functions: Let $I \subset \mathbb{R}$ be an interval and $f, g : \mathbb{R} \to \mathbb{R}^2$ be differentiable vector functions. The curves given by the images of $f$ and $g$ are said to be tangent at a point $P$ if there exists $t \in I$ such that $P = f(t) = g(t)$ and $f'(t) = g'(t)$. Use this definition to prove the statement at the beginning of the section about tangents to circles. (Hint: by translation, assume the circle is given by $f(t) = (r \cos t, r \sin t)$, that the point of tangency is $(r \cos t_0, r \sin t_0)$, and that the tangent line is $g(t) = (r \cos t_0, r \sin t_0) + t(a, b)$. Show that $g$ being tangent to $f$ at $P$ implies that $(a, b)$ is orthogonal to $(r \cos t_0, r \sin t_0)$.)

4. Many geometry texts state that a line is tangent to a circle at a point $P$ if the line intersects the circle at $P$ and only at $P$. Use this geometric definition to convince yourself of the truth of the statement “The line tangent to a circle at a point $P$ is always perpendicular to the radial segment joining the center of the circle to $P$,” by considering what might happen if the tangent line and the radial segment are not perpendicular.

5. Suppose $y_0 \neq 0$. Use implicit differentiation to verify that the slope of the tangent line to the circle $(x - 2)^2 + y^2 = 1$ at $(x_0, y_0)$ is $\frac{2 - x_0}{y_0}$.

6. Give another geometric solution to Problem 2, starting with the fact that the triangle with vertices at $(0, 0)$, $(2, 0)$, and $(x_0, y_0)$ is a right triangle. (Hint: Use the Pythagorean Theorem.)

7. Find the equations of the four lines that are tangent to the circle $(x + 2)^2 + y^2 = 1$ and to the circle $(x - 5)^2 + y^2 = 4$.

8. Find the equations of the circles that are tangent to the parabola $y = x^2$ at two points, and also tangent to the line $y = 2x + 20$. (You should find that there are two such circles.)

9. Find the equations of the two circles that pass through $(0, 0)$ and are tangent in two places to the ellipse $\frac{x^2}{4} + \frac{y^2}{9} = 1$. 
10. Let \( c > 0 \). For what values of \( c \) can one find a circle centered at \((0, c)\) that is tangent to the parabola \( y = x^2 \) at two points?

6. Exercises Involving Student Work

This section contains student solutions to various problems involving quadratic polynomials. We recommend that these solutions be analyzed prior to certain sections in the chapter (indicated below). For each problem and corresponding set of solutions, complete the following tasks:

1. Take 5 minutes in an attempt to solve the problem (without appealing to resources such as your text, etc.).

2. Discuss each student’s solution strategy. Indicate ways in which the student was “on-track”. Also, identify any incorrect assumptions, misconceptions, and errors revealed in each student’s work.

3. Try to determine whether the student’s numerical answer is correct.

4. Rate each solution on a scale from 1 to 5 with 5 as the highest rating and 1 as the lowest.

5. Compare your ratings in part 4 with another student (or group), and discuss your criteria for giving these ratings. Attempt to reach consensus with your partner (or group).

6. Carefully construct at least two solutions to the problem which employ different strategies. (You may use your text as a resource if you wish.)

11. Complete the square for the quadratic polynomial \( ax^2 + bx + c \). Show your work.

12. Find (the unique) quadratic polynomial such that all three of the following are true:

   - All the coefficients are integers.
   - The coefficient of \( x^2 \) is 4.
   - \( 7 + \sqrt{6} \) is one of the roots.
One reasoner in the party, bolder than the others, and shocked that someone doubted he had a soul, observed the interlocutor through the eyepiece of quadrant from two stations, and at the third spoke thus:

“You believe, then, sir, that because you are a thousand fathoms from head to foot, you are a ...”

“A thousand fathoms!” cried the dwarf. “Good heavens, how can he know my height? A thousand fathoms! He’s not an inch off. What? This atom has measured me? He is a geometrician, he knows my size; and I, who see him only through a microscope, I do not yet know his?”

“Yes, I have measured you,” said the physicist, “and I shall certainly measure your big friend, too.”

The proposition was accepted, His Excellency stretched out full length; for if he had remained standing, his head would have been too far above the clouds...Then, by a series of interrelated triangles, the physicist concluded that what they saw was indeed a handsome young man one hundred twenty thousand royal feet tall.

Voltaire (from Micromegas)

The quote above refers to triangulation, the method of computing a long distance (like the height of the handsome young royal) by means of measuring angles and distances from a few observation points. It is the
basis of surveying, including such feats as measuring the circumference of the earth. Triangulation is just one of the many applications of trigonometry (literally, *triangle measure*), a branch of mathematics based on the relationships among the measures of angles and sides in a triangle. In this chapter we will investigate some important aspects and applications of trigonometry, including the notions of angles and angle measure, the behavior of trigonometric functions, and the role of trigonometric identities and their historical importance in the development of trigonometric tables. We conclude the chapter with an introduction to trigonometric polynomials, providing an application of trigonometry beyond measuring triangles.

1. Angles

We are aware that angles play an important role in trigonometry, but what exactly *is* an angle? This surprisingly difficult question has been answered in several different ways throughout history. In this section, we explore the meaning and measurement of angles.

1.1. The meaning of angle: An exploration. We supposedly “know an angle when we see one,” but translating our common sense knowledge into a clear, concise definition can be a difficult process. The following Let’s Go exercises will help us focus our intuition concerning the definition and measurement of angles.

**Let’s Go 1.** Consider the pairs of ‘angles’ given in Figures 1 and 2.

![Figure 1. Pairs of ‘Angles’](image)

(a) For each pair angles shown in Figure 1, discuss both similarities and differences between the ‘angles.’
(b) From your observations in part (a), construct a definition of angle.

(c) For each pair of angles given in Figure 2, discuss both similarities and differences between the ‘angles.’

(d) In light of your observations in part (c), is it necessary to make revisions to the definition you gave in part (b)? If so, make them.

Let’s Go 2. Identify any misconceptions that arise in the following statements, and discuss their possible sources. Each statement refers to a portion of Figure 3.

(a) In part (a) the lefthandmost angle is smallest, and the size of the angles increases as we move to the right.

---

1These statements, motivated by Jane Keiser’s article in *Mathematical Thinking and Learning* (6 no. 3 (2004), 285-306), are representative of those given by 6th grade students.
4. TRIGONOMETRY

(b) The measure of an angle is determined by the ‘space’ between the two rays, so, in part (b) the lefthandmost angle is smallest, and the angles increase in size as we move to the right.

(c) As we add more sides to a regular polygon (as shown in part (c)), the angles (corners) are not as ‘sharp’, so the angles are getting smaller as we add more sides.

(d) In part (d) of the figure, the angle labeled $60^\circ$ is larger than the angle labeled $90^\circ$.

Let’s Go 3. When measuring an angle, what exactly are we measuring? Discuss. (Your observations in Let’s Go 2 may be helpful.)

1.2. Settling on an interpretation of the word ‘angle’. It is possible to interpret the word angle in various ways. However, since the measurement of angles is important in trigonometry, we settle on an interpretation of ‘angle’ that is most suitable for measurement. Informally, we declare an angle in the plane with vertex $P$ to be a ‘process of rotation’ centered at the point $P$. Further, an angle is positive if the movement is counterclockwise, negative if the movement is clockwise, and zero if there is no movement at all.

It is evident that motion is fundamental to our definition of angle. In order to capture angles on paper, we need to be able to represent this motion using stationary figures. One common way to represent angles involves rays, as shown in part (a) of Figure 4. Given an angle with vertex $P$, we rotate a ray emanating from $P$ (the initial side) onto a second ray (the terminal side). A circular arrow is used to indicate the motion from the initial side toward the terminal side. (In the event that no circular arrow is given, we usually assume that the angle is positive and that it consists of less than half of a complete revolution.)

Circular arcs provide another way to represent angles (see part (b) of Figure 4). Given an angle with vertex $P$, we apply the rotation to a point. Thinking of the point as the tip of a pencil, a circular arc centered at $P$ is drawn as we apply the process of rotation. Placing an arrowhead on the arc indicates whether the angle is positive or negative. Often, rays and arcs are used together to represent an angle, as shown in part

---

2Why do we say ‘process of rotation’ as opposed to ‘rotation’ in our interpretation of ‘angle’? Strictly speaking, a rotation in the plane is a linear transformation of determinant one that preserves distances between points. This fails to capture the orientation of an angle (i.e., whether it is positive or negative) and it also fails to describe situations in which the angle corresponds to more than one complete revolution. However, one can use rotations in the following formal definition: An angle is a triple $(P,T,n)$ where $P$ is a point in the plane, $T$ is a rotation, and $n$ is a nonnegative integer (here sign $(n)$ describes orientation and $|n| − 1$ the number of complete revolutions).
(c) of Figure 4. Finally, we must use care when representing angles corresponding to more than one complete revolution. The number of complete revolutions should be evident in the representation, as shown in part (d) of Figure 4.

Certain angle representations, known as standard representations, are particularly useful in trigonometry (as we shall see in Section 2). A standard representation of an angle is one whose initial side is the positive $x$-axis and whose vertex is the origin (as in part (d) of Figure 4).

Finally, in common mathematical usage the meaning of the word angle must often be deduced from context, since it can be used to describe the process of rotation (e.g., a twisting motion), a representation (e.g., two rays), or a measurement (e.g., number of radians). As we proceed we will be using the word angle, as well as symbols denoting angles, in each of these three contexts.

**Your Turn 1.** Consider the following alternate historical definitions of angle:

An angle is the inclination to one another of two lines in a plane which meet one another and do not lie in a straight line. [Euclid]

An angle is a quantity, namely, a distance between the lines or surfaces containing it. [Carpus of Antioch]

(a) Draw figures illustrating these two definitions.

(b) Compare these two angle definitions with the informal definition of angle given above. Discuss similarities and differences among the definitions.
(c) Are any of the three angle definitions more suited to measurement than the other two? Discuss.

1.3. Angle measurement. If our chief concern were plane geometry, then we could get along quite well without ever measuring angles. In fact, Euclid’s *Elements*, with its strict adherence to the Platonic rules of compass and straightedge constructions, contains no numerical angle measurements whatsoever\(^3\). In trigonometry however, the measurement of segments, arcs, and angles is the heart of the subject.

1.3.1. Degrees. For the purposes of degree measure, a circle is divided into 360 equal arcs by taking 360 equally spaced points on the circle, with each arc standing for one degree. To make an angle measurement, one counts the number of degrees in any circular arc representing a given angle, being sure to take into account any complete revolutions. We declare the final measurement to be positive, negative, or zero depending on whether the angle itself is positive, negative, or zero. For example, a negative angle represented by a third of a circle is \(\frac{-360^\circ}{3} = -120^\circ\).

While there is no intrinsic reason to divide a circle into 360 degrees, the notion of degree measure has an interesting history, originating with the Babylonians. The Babylonians may have recognized that a regular hexagon can be easily constructed within any given circle (see Figure 5). The vertices of the hexagon produce a subdivision of the circle into six equal parts and, since the Babylonian numeration system was a place-value system using base 60 (instead of the base 10 with which we are familiar in Hindu-Arabic numerals), it would

\[^3\text{The Greeks viewed the results of the *Elements* to be ‘basic’ or ‘elemental’. Topics involving measurement, such as trigonometry, were viewed as ‘higher’ mathematics.}\]
have been natural for them to subdivide each of the six parts of the circle into 60 pieces, thus giving 360 pieces total\(^4\).

1.3.2. **Radians.** The radian angle measure of an angle is found by counting how many copies of the radius it takes to exhaust a circular arc representing the angle. Imagining that the radius is a flexible rod, we may bend the rod around the arc so that one may see how many rods it takes to make up the whole arc. As before, the final radian measurement of an angle may be positive, negative, or zero depending on the type of angle being measured.

**Your Turn 2.** Consider the angles, with corresponding circular arcs, given below. For each angle, cut a length of some flexible material to the length of the radius of the circle. Then, by wrapping the material around the arc, estimate the radian measure of the angle.

---

\(^4\)There has been a lot of speculation as to why the Babylonians chose 60 as the base for their number system. One possible reason is that 60 is rich in divisors, making it more likely that a given fraction will have a terminating base 60 ‘decimal’ expansion. Another possible reason is that the number system was based on an existing form of monetary currency, which was divided into 60’s. The choice of 360 for the number of parts in a circle also may be related to the number of days in a year (365.25 \(\sim\) 360).
Of course, circular arcs of many differing radii can be used to represent the same angle. So, for radian measurement to be valid for a given angle, we require the righthand side of (1) to produce the same number regardless of which circular arc we choose to represent the given angle. This relies upon the intrinsic fact that for circles the ratio of circumference to radius is always $2\pi$ (see Exercise 13).

**Your Turn 3.** Discuss the meaning of Equation (1) when the radius of the circle is 1 unit.

Finally, since $360^\circ$ and $2\pi$ radians both correspond to an angle consisting of exactly one counterclockwise revolution, we deduce that

$$1^\circ = \frac{\pi}{180} \text{ radians} \quad \text{and} \quad 1 \text{ radian} = \frac{180^\circ}{\pi}.$$  

1.4. Exercises.

1. In a single picture, plot and label rays corresponding to the standard angles with radian measures $2\pi/3$, $5\pi/6$, $7\pi/4$, $7\pi/8$, and $31\pi/4$.

2. In a single picture, plot and label rays corresponding to the standard angles with radian measures 1, 5, and $\sqrt{3}$.

3. Estimate the measures of the three standard angles illustrated in Figure 7 using both radian and degree angle measure.

4. Revisit Let’s Go 2. Construct explanations you could give to students who hold these misconceptions.

5. The diameter of a circle is 14 inches and the length of an arc of the circle is 5 inches.
   (a) Find the measure of the central angle that intercepts the arc.
   (b) Give the angle measure from part (a) in degrees.

6. Consider the following conversion problems:

---

5The fact that the ratio of circumference to radius is constant in circles was known by virtually all of the advanced ancient civilizations, including the Babylonians, Chinese, Egyptians, Indians, and Greeks. An important point: the ratio gives a definition of $\pi$. We should also be aware that we are relying upon a meaningful definition of circumference. Since the circle is curved (not merely a line segment), the notion of its length is actually rather subtle, and requires a “limiting process” that one sees in calculus.
1. **ANGLES**

7. Consider the following conversion problems:

(a) For angles with radian measure 0, $\frac{2\pi}{3}$, $\frac{\pi}{6}$, and 4, find the corresponding degree measure.

(b) For angles with degree measure $0^\circ$, $135^\circ$, $80^\circ$, and $\pi^\circ$, find the corresponding radian measure.

8. (1) Draw a circle in the plane with center at the origin.

(2) Draw points on the circle to represent standard angles with measures $0^\circ$, $30^\circ$, $60^\circ$, $90^\circ$, $120^\circ$, $150^\circ$, $180^\circ$, $210^\circ$, $240^\circ$, $270^\circ$, $300^\circ$, $330^\circ$ (all multiples of $30^\circ$). Label the points with these degree measures.

(3) Label each point with the corresponding radian measure of the angle. Actually, you should do this in two different ways. First, write each radian measure in the form $2\pi \cdot \frac{k}{12}$ where $k$ is a nonnegative integer. Then rewrite each radian measure as $\pi$ times a fraction in reduced form (lowest terms).

9. A student is asked to convert 3 radians to degree measure. The student calculates

$$3 \cdot \frac{\pi}{180} = \frac{\pi}{60}$$

and answers $(\frac{\pi}{60})^\circ$. Use a diagram to show the student why his answer is incorrect.

10. In this exercise we investigate how certain rotations affect coordinates of points.

(a) Plot the eight points with coordinates $(\pm3, \pm4)$ and $(\pm4, \pm3)$. (All lie on the circle of radius 5 centered at the origin.)
(b) Explain (by labeling) how each point was obtained from the point \((3, 4)\) either by rotation by \(90^\circ\), \(180^\circ\), or \(270^\circ\), or by reflection over one of the lines \(x = 0\), \(y = 0\), \(y = x\), or \(y = -x\).

(c) Based on what you have just done, explain what one has to do to the coordinates of a point to obtain each of the following new points: (i) its rotation by \(90^\circ\); (ii) its rotation by \(180^\circ\); (iii) its rotation by \(270^\circ\); (iv) its reflection over the \(x\)-axis; (v) its reflection over the \(y\)-axis; (vi) its reflection over the line \(y = x\); (vii) its reflection over the line \(y = -x\).

11. We will use circles, and in particular, the unit circle repeatedly in this chapter, so it’s important for us to be able to produce points on circles at will!

(a) List six points \((x, y)\) that lie on the unit circle (circle with center at the origin and radius 1) and also are in the first quadrant \((x > 0 \text{ and } y > 0)\).

(b) Plot the six points carefully.

(c) Use your answer to (a) to find six points in the first quadrant that lie on the circle centered at the origin with radius \(\frac{3}{2}\). Plot the points.

(d) The point \((3, 4)\) does not lie on the unit circle. Which scalar multiples of \((3, 4)\) do lie on the unit circle?

12. Radian angle measure employs arc length. We could also measure angles using area. Specifically, given the standard representation an angle with radian measure \(\theta\), let \(A\) be the area of the circular sector bounded by the initial side (the \(x\)-axis), the terminal side, and the unit circle (taking into account rotations consisting of more than one complete revolution). Find the constant \(c\) so that \(\theta = cA\). (Hint: think about the case where the angle consists of one complete revolution.)

13. How does the fact that

\[
\frac{\text{Circumference}}{\text{Radius}} = 2\pi
\]

for any circle imply that the righthand side of Equation (1) is constant for every representative arc? Pictures may be helpful.

14. Insphere a regular hexagon in a circle using only a compass and unmarked straightedge.
2. Triangles, Circles, and Trigonometric Functions

The word *trigon* refers to a three-sided figure, while *metry* means measurement. Thus trigonometry is the measurement of triangles, which is tantamount to studying the measurement of and relationships among side-lengths and angles. There is no doubt that trigonometry is useful. Human beings have been using triangles to make measurements (e.g., the height of Everest, the circumference of the earth, the distance from the earth to the sun) for thousands of years\(^6\). But why is trigonometry nontrivial? After all, we know how to measure angles and line segments, so what can be so hard about trigonometry?

**Let’s Go 4.** Consider Figure 8. If \(\theta_2 = 2\theta_1\), is it true that \(|AD| = 2|AC|\)? Justify your answer with an argument or a compelling figure.

**Figure 8**

In Let’s Go 4 we begin to perceive that the relationships among angles and side lengths of a triangle are difficult to determine\(^7\). Understanding these relationships is a fundamental goal of trigonometry.

2.1. Trigonometric functions and triangles. The fundamental relationships among angles and side lengths in a right triangle are encoded in the six trigonometric functions. To define these functions, we start with an angle \(\theta\) satisfying \(0 < \theta < \pi/2\) (more values of \(\theta\) will be considered later). This angle determines a collection of similar right triangles, each of which is formed by erecting a perpendicular from the initial side

\(^6\)Less transparently, trigonometry is the essential tool for understanding periodic phenomena, such as vibrating strings, bouncing springs, and rotating tops. This aspect of trigonometry will be discussed in Section 7.

\(^7\)In Let’s Go 4, we see that the relationship between angle and side length is nonlinear. In fact, it is not given by any simple algebraic function like a polynomial or a rational function, so the relationship is transcendental.
of $\theta$ to the terminal side (see part (a) of Figure 9). Given one of these triangles (part (b) of Figure 9), we define

$$\begin{align*}
\cos \theta &= x/h \\
\sin \theta &= y/h \\
\tan \theta &= y/x \\
\sec \theta &= h/x \\
\csc \theta &= h/y \\
\cot \theta &= x/y,
\end{align*}$$

where $x$, $y$, and $h$ are the lengths of the adjacent side, the opposite side, the hypotenuse of the triangle, respectively.

**Your Turn 4.** Every function takes an input (domain value) and produces an output (range value).

(a) Describe the input values and output values for the trigonometric functions qualitatively.

(b) What can you add to the qualitative description of the output values in the special case $h = 1$?

(c) In our present definition of the trigonometric functions, why are we limited to considering angles between 0 and $\pi/2$ radians?

**Your Turn 5.** An angle $\theta$ with $0 < \theta < \pi/2$ corresponds to infinitely many right triangles (as in part (a) of Figure 9). Do the values of the trigonometric functions at $\theta$ depend on which of these triangles we choose? Why or why not?

**Your Turn 6.** Many of us remember that $\sin \pi/6 = 1/2$. Why is this true? Evaluate all six trigonometric functions at $\pi/6$, $\pi/4$, and $\pi/3$, and justify your answers.

---

*The modern names for the trigonometric functions are steeped in history. For example, ancient Hindus were the first to consider sine as a ratio of side-lengths, which they called *jiva*, a shortened version of “chord-half.” Arabs then translated this phonetically to *jaib*, which means “curve” or “fold.” Finally, as Europeans of the middle ages were becoming aware of Arab scholarship, they translated jaib as *sinus*, the Latin word for curve.*
2.2. Trigonometric functions and circles. Using right triangles to define the trigonometric functions becomes a tortured process for angles with measures not strictly between 0 and \( \pi/2 \). As reluctant as we might be to force triangles onto the back burner, it is much more natural to define trigonometric functions in terms of circles. To define the trigonometric functions in this way, we start with an angle of radian measure \( \theta \) in standard position, and a circle centered at the origin of radius \( r \) (see Figure 10). Let \( P = (x, y) \) be the point where the terminal side of the angle intersects the circle. All of the trigonometric functions of \( \theta \) are defined using the coordinates of the point \( P \). We put

\[
\cos \theta = \frac{x}{r} \quad \sin \theta = \frac{y}{r} \quad \tan \theta = \frac{y}{x} \\
\sec \theta = \frac{r}{x} \quad \csc \theta = \frac{r}{y} \quad \cot \theta = \frac{x}{y},
\]

Your Turn 7. For angles strictly between 0 and \( \pi/2 \), we now have two definitions for the trigonometric functions: one involving triangles and one involving circles. Construct a diagram illustrating the equivalence of the two definitions.

Your Turn 8. Show that \( \sin^2 \theta + \cos^2 \theta = 1 \). Why do you think this is called the Pythagorean identity?

2.3. Graphs and periodicity. Graphs of the trigonometric functions have the virtue of conveying information at a glance, some of which may not be obvious in non-graphical form. In this section we investigate the graph of the sine function. The graphs of the other trigonometric functions are addressed in the exercises.
We begin by examining a table of values\(^9\) for the sine function (see Figure 11). Values of \(\theta\) are given in approximately equal increments.)

\[\text{Figure 11}\]

**Your Turn 9.** Consider the table in Figure 11.

(a) Mark the approximate locations of various ‘special angles’ in the table, such as \(\pi/4\), \(\pi/2\), \(2\pi/3\), etc.

(b) Where does the function appear to be increasing the fastest? The slowest? Use Figure 12 to explain your answers.

\[\text{Figure 12}\]

Using the table, we can sketch a graph of the sine function (see Figure 13). Probably the first thing that pops out at us is that the graph of \(\sin \theta\) begins to repeat itself after traveling \(2\pi\) units on the \(\theta\)-axis. More

\(^9\)It is natural to wonder how one computes such values. One approach will be discussed in Section 4.
precisely,

\[ \sin(\theta + 2\pi) = \sin \theta \]

for every value of \( \theta \). Functions that repeat themselves in this way are called periodic. Specifically, a function \( f \) is said to be periodic with period \( p > 0 \) if \( f(x + p) = f(x) \) for all \( x \). Thus the sine function is periodic with periods \( p = 2k\pi \) for positive integers \( k \). Among these periods, the least period for the sine function is \( 2\pi \). The other trigonometric functions are also periodic but not all of them have \( 2\pi \) as their least period. The periodic nature of the trigonometric functions makes them an ideal tool for modeling various periodic phenomena, such as radio signals and vibrating strings.

Using a similar table of values for the cosine function, we can produce a graph of the cosine function (see Figure 14). We can see that the graph of the cosine function is obtained from the graph of the sine function by a horizontal shift to the left by \( \frac{\pi}{2} \), so

\[ \cos \theta = \sin \left( \theta + \frac{\pi}{2} \right) = \sin \left( \frac{\pi}{2} - \theta \right) \]

**Your Turn 10.** How does Equation (3) follow from the fact that the graph of the cosine function is obtained from the graph of the sine function by a horizontal shift to the left by \( \frac{\pi}{2} \) units?

**Your Turn 11.** Here we investigate the fact that all trigonometric functions are periodic.

(a) Using the (circular) definition of the sine function, explain why \( \sin(\theta + 2\pi) = \sin \theta \) for all \( \theta \). Draw a picture supporting your explanation.
(b) Find the periods of the other five trigonometric functions. (You may use the definitions of the trigonometric functions as well as graphs.)

2.4. Inverse trigonometric functions. To define the trigonometric functions, one starts with an angle $\theta$ and finishes with some sort of information about the corresponding point $P$ on the unit circle. For example, a step-by-step procedure for computing $\cos \theta$ might look like:

$$
(4) \quad \text{angle measure} \rightarrow \text{standard representation of angle} \rightarrow \text{point on unit circle} \rightarrow x\text{-coordinate of point}
$$

To obtain the inverse trigonometric functions, we reverse this procedure. That is, starting with some information about a point $P$ on the unit circle (e.g., the $x$-coordinate of $P$), we would like to finish with an angle $\theta$ that corresponds to the point $P$. For example, suppose we start with the information that the $x$-coordinate of a point $P$ on the unit circle is $1/2$, and we want to find a single number $\theta$ that determines a standard angle corresponding to $P$ (that is, given $\cos \theta = 1/2$, we want to determine $\theta$). As we reverse the procedure given in (4), we see that there are two places where a ‘choice’ has to be made: First, there are two points on the unit circle whose $x$-coordinate is $1/2$, and second, there are infinitely many standard angles with terminal side passing through each of these points. Therefore, there are infinitely many angles $\theta$ from which to choose, such as $\pi/3$, $-\pi/3$, or $2\pi + \pi/3$. More generally,

$$
\theta = \pi/3 + 2k\pi \quad \text{or} \quad \theta = -\pi/3 + 2k\pi, \quad k \in \mathbb{Z}.
$$

One way to visualize this is given in Figure 15: Look for intersections of the line $y = 1/2$ with the curve $y = \cos \theta$. 

![Figure 14](image-url)
The main problem in defining an inverse cosine function is that the cosine function is not one-to-one. The only reasonable way to solve this problem is to restrict the values of \( \theta \) from which we can choose. The standard way to do this (see Figure 16) is to restrict the cosine function to the interval \([0, \pi]\), so that the resulting cosine function \(\cos : [0, \pi] \to [-1, 1]\) is both one-to-one and onto. We then define the inverse cosine function (denoted by \(\cos^{-1}\) or \(\arccos\)) so that

\[
\arccos : [-1, 1] \to [0, \pi], \quad \cos^{-1} y = \theta \iff \cos \theta = y.
\]

In other words, \(\arccos x\) is the unique angle between 0 and \(\pi\) whose cosine is \(x\). For example, \(\arccos(1/2) = \pi/3\). Using similar reasoning, we may define the other five inverse trigonometric functions:

1. \(\arcsin : [-1, 1] \to [-\pi/2, \pi/2]\),
   \[\text{arcsin } y = \theta \iff \sin \theta = y.\]
2. \(\arctan : \mathbb{R} \to (-\pi/2, \pi/2)\),
   \[\text{arctan } y = \theta \iff \tan \theta = y.\]
3. \(\text{arcsec} : (-\infty, -1] \cup [1, \infty) \to [0, \pi/2) \cup (\pi/2, \pi]\),
   \[\text{arcsec } y = \theta \iff \sec \theta = y.\]
4. \(\text{arccsc} : (-\infty, -1] \cup [1, \infty) \to [-\pi/2, 0) \cup (0, \pi/2]\),
   \[\text{arccsc } y = \theta \iff \csc \theta = y.\]
5. \(\text{arccot} : \mathbb{R} \to (0, \pi)\),
   \[\text{arccot } y = \theta \iff \cot \theta = y.\]

**Your Turn 12.** In defining \(\arccos : [-1, 1] \to [0, \pi]\) above, why is important that \(\cos : [0, \pi] \to [-1, 1]\) be both one-to-one and onto?
Your Turn 13. Suppose \( f(x) \) is a periodic real-valued function defined on the real numbers. Explain why \( f(x) \) can not be one-to-one. Support your explanation with a convincing graph.

Your Turn 14. Notation is not always our best friend: Sometimes \( \sin^{-1} \) is used to denote the inverse sine function. Discuss how the ‘2’ in \( \sin^2 \theta \) differs in meaning from the ‘-1’ in \( \sin^{-1} \). How could this cause confusion among students seeing this notation for the first time?

2.5. Exercises.

1. For the standard angle and point in Figure 17, use a ruler to estimate \( x \), \( y \), and \( r \), and from these, estimate the values of the six trigonometric functions.

2. Sarah thinks that \( \tan 2\theta = 2 \tan \theta \) for all values of \( \theta \).
   
   (a) Discuss possible sources of Sarah’s misconceptions.
   
   (b) Help Sarah see that \( \tan 2\theta \neq 2 \tan \theta \) in two ways: By making a calculation, and by drawing a picture. (Hint: For a compelling picture, revisit Let’s Go 4.)

3. In this exercise, you will use graph paper or geometry software to create a table of values of the trig functions.
   
   (a) Construct a large quarter-circle on a sheet of graph paper using a compass. You should be picturing this as the portion of the circle in the quadrant \( x \geq 0, y \geq 0 \). Choose the radius so that relative to
the graph paper, it will be easy to read off lengths along the $x$- and $y$-axes (for example, make the radius equal to ten units).

(b) Using a protractor, sketch rays for the angles $10^\circ$, $20^\circ$, $70^\circ$, $80^\circ$.

(c) Observe the intersection points between the circle and the rays. Estimate the $x$- and $y$-coordinates of each of these points.

(d) Use this information to construct a table of approximate values for $\cos \theta$, $\sin \theta$, and $\tan \theta$ for $\theta = 0^\circ$, $10^\circ$, $20^\circ$, $80^\circ$, $90^\circ$.

(e) Use a calculator to obtain accurate values of the trig functions at these angles. How accurate were your estimates?

4. Find expressions for $\sin \theta$, $\cos \theta$, $\cot \theta$, and $\csc \theta$ using only $\sec \theta$ and $\tan \theta$.

5. Give an expression for the slope of the ray forming an angle with measure $\theta$, in terms of one or more trigonometric functions.

6. Does a value for the cosine function completely determine a value for the sine function? Let’s find out:

   (a) If $\cos \theta = \frac{\sqrt{3}}{2}$, what can $\sin \theta$ be? Illustrate the possible angles $\theta$, and give the radian measure of each possible angle. How are the angles related to each other?

   (b) If $\sin \theta = -\frac{3}{5}$, what can $\cos \theta$ be? Illustrate the possible angles $\theta$, and give the radian measure of each possible angle. How are the angles related to each other?

7. Using triangles, form a table of exact values of $\sin \theta$, $\cos \theta$, $\tan \theta$, $\csc \theta$, $\sec \theta$, and $\cot \theta$ for the angle measures $30^\circ$, $45^\circ$, $60^\circ$, $135^\circ$, and $210^\circ$.

8. A student has drawn a triangle with side lengths 3, 7, and 8. Let $\theta$ denote the measure of the angle opposite the side of length 3. The student claims that $\cos \theta = \frac{7}{8}$. Why is the student incorrect? What error has she made?

9. Use the graph of $\sin \theta$ together with the fact that $\cos \theta = \sin(\theta + \pi/2)$ to obtain the graph of $\cos \theta$ quickly.
10. Using graphing transformations, deduce the graph of each of the following, from the graph of \( \sin \theta \).

(Do not use a graphing device.)

(a) \( 5 \sin \theta \)

(b) \( \sin(3\theta) \)

(c) \( \sin(\theta - \pi/3) + 1 \)

(d) \( 5 \sin(\theta - \pi/3) + 1 \)

(e) \( \sin(2\theta - \pi/3) = \sin(2(\theta - \pi/6)) \)

11. Repeat Exercise 10 with

(a) \( \sin \theta \) replaced by \( \cos \theta \).

(b) \( \sin \theta \) replaced by \( \tan \theta \).

(c) \( \sin \theta \) replaced by \( \sec \theta \).

12. Use the marked unit circle in Figure 18 to estimate the following:

(a) \( \arcsin(0.8) \)

(b) \( \arccos(0.6) \)

(c) \( \arctan(0.4) \)

(d) \( \arctan(3.0) \)

13. Graph all six of the inverse trigonometric functions.
14. The following table shows what one calculator knows (and doesn’t know) about inverse trigonometric functions.

<table>
<thead>
<tr>
<th>x</th>
<th>-2.00</th>
<th>-1.00</th>
<th>-0.50</th>
<th>-0.25</th>
<th>0.00</th>
<th>0.25</th>
<th>0.50</th>
<th>1.00</th>
<th>2.00</th>
<th>5.00</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>arcsin x</td>
<td>-1.57</td>
<td>-0.52</td>
<td>-0.25</td>
<td>0.00</td>
<td>0.25</td>
<td>0.52</td>
<td>1.57</td>
<td>1.31</td>
<td>1.04</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>arccos x</td>
<td>3.14</td>
<td>2.09</td>
<td>1.82</td>
<td>1.57</td>
<td>1.31</td>
<td>1.04</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arctan x</td>
<td>-0.79</td>
<td>-0.46</td>
<td>-0.24</td>
<td>0.00</td>
<td>0.24</td>
<td>0.46</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Can any of the gaps in the table be filled in? Which and why?

(b) In each column, the entries for arcsin x and arccos x add to 1.57. Why is this happening?

15. Plot the graph of sin[arcsin x] for -1 ≤ x ≤ 1. Then plot the graph of arcsin[sin x] for -2π ≤ x ≤ 2π.

Briefly explain the reason for the shape of each graph.

16. Is it true that arctan x = \frac{\text{arcsin } x}{\text{arccos } x}? Why or why not?

17. Why does the function arcsec x accept only values of x with |x| ≥ 1?

18. In the following problems, show your work and briefly explain your strategy:

   (a) Find cos(\frac{52}{6} π).

   (b) Find arcsin(sin(\frac{50}{7} π)).

   (c) Find arccos(cos(\frac{1109}{11} π)).

19. Write each of the following as an algebraic expression. For example, cos(arcsin x) = \sqrt{1 - x^2}.

   (a) sin(arccos x)

   (b) sin(arctan x)

   (c) tan(arcsin x)

   (d) cos(arctan x)

20. We know that the trigonometric functions are all periodic. Are the inverse trigonometric functions periodic? Why or why not?

21. Using the definitions of the trigonometric functions, explain why cosine is an even function (that is, cos(-θ) = cos θ for all θ) and sine is an odd function (that is, sin(-θ) = -sin θ for all θ).
22. In this exercise\textsuperscript{10}, we estimate the distance from the earth to the sun (as a multiple of the distance from the earth to the moon) following the method of the great astronomer Aristarchus of Samos\textsuperscript{11} (310-230 b.c.).

(a) Suppose $\alpha, \beta$ are angle measures with $0 \leq \alpha < \beta \leq \frac{\pi}{2}$. Explain why \( \tan \alpha \frac{\alpha}{\alpha} < \tan \beta \frac{\beta}{\beta} \) and \( \sin \alpha \frac{\alpha}{\alpha} > \sin \beta \frac{\beta}{\beta} \).

(Hint: Set $\beta = k\alpha$ for some $k > 1$, so that one may think of $\beta$ as $k$ increments of $\alpha$. Then \( \frac{\tan \alpha}{\alpha} = \frac{k\tan \alpha}{\beta} \). Now, how does $k\tan \alpha$ compare to $\tan \beta$? To see how they compare, look carefully at Figure 19. Use the same trick for the other inequality, using Figure 12.)

(b) At quarter moon, when exactly half of the moon is visible, the lines of sight from the earth observer (E) to the moon(M) and to the sun (S) form a right triangle (see Figure 19). Aristarchus observed that the angle $\gamma$ in the figure measures approximately $87^\circ$, so that $\alpha = 3^\circ$. Use the inequalities in part (a) to deduce\textsuperscript{12} that $\frac{1}{20} < \sin \alpha < \frac{1}{10\sqrt{3}}$. (Hint: Use $\beta = \pi/6$ for the inequalities in part (a).)

\begin{figure}[h]
\centering
\caption{}
\end{figure}

\textsuperscript{10}This exercise is motivated by Eli Maor’s book \textit{Trigonometric Delights}.

\textsuperscript{11}Samos is an island in the eastern Aegean Sea, just off the coast of Turkey. The mathematical heritage of Samos is literally inscribed in the landscape in the form of a remarkable tunnel. The tunnel, which passes through a large central mountain, was built in approximately 500 b.c.e. Using geometric principles, workers excavated the tunnel from both ends simultaneously, meeting in the middle with very little error.

\textsuperscript{12}Aristarchus lacked a robust trigonometric table. However, he was able to deduce this inequality bounding the possible values of $\sin 3^\circ$.  

(c) Use the result of part (b) to deduce that the distance from the earth to the sun, namely $|EM|/\sin \alpha$, is between $10\sqrt{3}$ and 20 times the distance from the earth to the moon.

(d) Aristarchus’ estimate in part (c) is inaccurate. The actual distance to the sun is about 400 times the distance to the moon. Where do you suspect the error in the approximation originated?

3. Laws of Sines and Cosines

Thus far we only know how to evaluate the trigonometric functions at a few special angles. Trigonometry would remain limited and pointless if matters stood there. However there are two important theorems, known as the laws of sines and cosines, that allow us to move forward with trigonometry, and from which the real power of trigonometry is derived. We will see that these theorems provide the following crucial connections:

- Information about the trigonometric functions translates to information about triangles (not just right triangles).
- There exist nontrivial relationships (i.e., identities) among the trigonometric functions that allow us to evaluate the trigonometric functions for many angles.

3.1. The laws of sines and cosines. We present the Laws of Sines and Cosines, beginning with the Law of Cosines:

**Theorem 1.** (Law of Cosines) *If a triangle has side lengths $A$, $B$, and $C$, and an angle of measure $\gamma$ opposite the side of length $C$, then $C^2 = A^2 + B^2 - 2AB \cos \gamma$.*

**Proof.** We assume that $0 \leq \gamma \leq \pi/2$. The case where $\pi/2 < \gamma \leq \pi$ is addressed in Exercise 17.

Draw an altitude as in Figure 20. Applying the Pythagorean Theorem to the left subtriangle, we have $A^2 = h^2 + (A \cos \gamma)^2$. Applying the Pythagorean Theorem to the other subtriangle, we have

$$C^2 = h^2 + (B - A \cos \gamma)^2$$

$$= (A^2 - (A \cos \gamma)^2) + (B - A \cos \gamma)^2$$

$$= A^2 - A^2 \cos^2 \gamma + B^2 - 2AB \cos \gamma + A^2 \cos^2 \gamma$$

$$= A^2 + B^2 - 2AB \cos \gamma.$$
The Law of Cosines is valid no matter how we label the three sides of the triangle. In particular, $C$ does not necessarily represent the longest side of the triangle.

Your Turn 15. Which famous theorem is implied by the Law of Cosines when $\gamma = \pi/2$?

Your Turn 16. For the triangle shown in Figure 21, measure the sides labeled $A$ and $B$, and measure the included angle $\gamma$. Use this information together with the Law of Cosines to estimate the length of side $C$. Then measure $C$, and compare your measurement with your estimate.

We now proceed to the Law of Sines, but first we present a fact about circles that is important in its own right.
Proposition 2. If \( P, Q, \) and \( R \) are points on a circle whose center is \( O \), then the measure of the angle \( \angle POR \) is twice the measure of the angle \( \angle PQR \).

**Proof.** There are three cases to consider, we address only the situation represented in Figure 22. (The other two cases are discussed in Exercise 4.) Observe isosceles triangles \( \triangle PQO \) and \( \triangle ROQ \). Since the sum of the measures of the angles in a triangle is \( \pi \), we see that \( \triangle PQO \) has angles with measures \( \alpha, \alpha, \) and \( \pi - 2\alpha \), and \( \triangle ROQ \) has angles with measures \( \beta, \beta, \) and \( \pi - 2\beta \). Finally, since the sum of the three angles at \( O \) is \( 2\pi \), we compute

\[
\gamma = 2\pi - (\pi - 2\alpha) - (\pi - 2\beta) = 2(\alpha + \beta).
\]

\( \square \)

Theorem 3. (Law of Sines) Let \( A, B, \) and \( C \) denote the lengths of the sides of a triangle, and let \( \alpha, \beta, \) and \( \gamma \) denote the opposite angles, respectively. Then

\[
\frac{A}{\sin \alpha} = \frac{B}{\sin \beta} = \frac{C}{\sin \gamma} = 2r,
\]

where \( r \) is the radius of the circle that circumscribes the triangle.

**Proof.** Let \( P, Q, R \) be points on a circle with center \( O \) and radius \( r \), forming a triangle as shown in Figure 23. Draw in radial segments \( OQ \) and \( OR \), and a segment joining \( O \) to the midpoint \( M \) of \( QR \). By Proposition 2, \( \angle QOR \) has measure \( 2\beta \), so \( \angle QOM \) has measure \( \beta \). The segment \( QM \) has length \( B/2 \), and
the segment $OQ$ has length $r$. From the triangle $\triangle QOM$, we have $\sin \beta = \frac{B/2}{r}$, so $\frac{B}{\sin \beta} = 2r$. The other equations in the statement of the theorem are proved in a similar fashion.

Your Turn 17. Suppose a triangle has side lengths $A$, $B$, and $C$, and opposite angles $\alpha$, $\beta$, and $\gamma$, respectively. When will it be true that $A = \sin \alpha$, $B = \sin \beta$, and $C = \sin \gamma$? Draw a figure illustrating the situation.

3.2. Exercises.


2. For a triangle with side lengths $A = \sqrt{2}$, $B = 3$, and $C = \sqrt{5}$, find $\cos \gamma$. Do the same for the triangle with side lengths $A = \sqrt{2}$, $B = 3$, and $C = \sqrt{17}$. Sketch the triangles.

3. Use the Law of Cosines to verify that $\cos \pi/3 = \frac{1}{2}$. (Hint: Consider equilateral triangles.)
4. Prove Proposition 2 for the two cases indicated in Figure 24. (Hint: The proof for the righthand figure uses supplementary angles. For the lefthand figure, use the fact that the dashed line $OQ$ will form part of an isosceles triangle and that the angles in the two triangles meeting at $A$ are equal.)

**Figure 24.** This figure is for Exercise 4.

5. Suppose $A = 2$, $B = 3$, and $C = 7$.

   (a) Explain why there is no triangle with these side lengths.

   (b) Calculate what the Law of Cosines would give for $\cos \gamma$. (As you might expect in this situation, the Law of Cosines gives a nonsensical result.)

   (c) More generally, given $A$, $B$, and $C$, suppose that $C \geq A + B$. According to the Law of Cosines, what would $\cos \gamma$ be? Why is this nonsensical?

6. In a triangle, suppose that the angle measures are $2\pi/5$, $3\pi/7$, and $6\pi/35$. Suppose that the side opposite the angle with measure $6\pi/35$ has length 4. Find the lengths of the other sides. Sketch the triangle.

7. Consider an equilateral triangle with sides of length $L$. What is the radius of the circle in which the triangle can be inscribed?

8. Interpret the Law of Cosines and the Law of Sines when $\gamma = 90^\circ$ and when $\gamma = 0^\circ$. 
9. Consider the system of triangles shown below in Figure 25. Suppose that the distance from station A to station B is 25 miles. Use the Law of Sines to approximate the distance from station F to station E. (This, more or less, is the process of triangulation used in surveying.)

**Connection with Triangle Congruence Theorems:** Given enough initial information about a triangle, the Laws of Sines and Cosines can be used to determine all of its side lengths and angle measures. In Exercises 10 through 13, we discover how much information is required to uniquely determine a triangle, and forge a connection with triangle congruence theorems.

10. Here are some introductory exercises about congruence theorems.

   (a) What does it mean for two triangles to be congruent?

   (b) You may assume that SSS, SAS, and ASA are all valid congruence theorems. Discuss and draw a picture illustrating the meaning of each of these theorems.

   (c) Is AAS a valid triangle congruence theorem? If so, briefly show how it follows from the congruence theorems in part (b). If not, produce a counterexample.

   (d) Repeat part (c) for AAA and ASS.

11. Here we connect the Law of Cosines with SSS.

   (a) Does the value of \( \cos \gamma \) uniquely determine an angle \( \gamma \) satisfying \( 0 \leq \gamma \leq \pi \)? Why?
(b) Use the Law of Cosines to show that if we know all three sides of a triangle (say, of lengths $A$, $B$, and $C$), then we can uniquely determine all three angles. How does part (a) play a role in this process?

(c) Construct a large triangle. Measure the three side lengths. Then, use the Law of Cosines to compute the cosine of each of the angles in the triangle.

(d) Using your figure from part (c), measure the angles and compute the cosines of the angle measures. How well do the cosines you just computed agree with the cosines from part (c)?

12. Here we connect the Law of Cosines with SAS.

(a) Use the Law of Cosines to show that if we know two side lengths (say $A$ and $B$) and the interior angle between them (say $\gamma$), then we can determine the other two angles and the remaining side.

(b) Use a protractor or geometry software to draw a $60^\circ$ angle, and measure the lengths of the ‘sides’ $A$ and $B$ of the angle. Draw a third side (of length $C$) to complete a triangle. Use the Law of Cosines to compute $C$. Finally, measure $C$ directly and compare with your calculated value of $C$.

13. Here we connect the Law of Sines with ASA. Suppose we know angles $\beta$ and $\gamma$ (and hence the third angle $\alpha$) and the common side $A$. Use the Law of Sines to show that we can determine the other two sides $B$ and $C$.

14. Given a triangle, suppose that $A = 3$, $B = 4$, and $\sin \gamma = \frac{1}{2}$. Show that there are two triangles that satisfy these conditions. Find $C$, $\sin \alpha$, $\cos \alpha$, $\sin \beta$, and $\cos \beta$ for both triangles. Sketch the triangles.

15. Suppose that the angle measures $\alpha$, $\beta$, and $\gamma$ in a triangle are known.

(a) Explain why it is not possible to know the side lengths of the triangle (only knowing the angle measures).

(b) Explain how it is possible to compute the ratio of any two side lengths of the triangle.

16. Use the Law of Sines to find the equation of the circle containing the three points $(0,0)$, $(3,0)$, and $(0,4)$. (First use the Law of Sines to find the radius of the circle. Then, either use a fact from geometry or solve a system of equations to find the center.)
17. Prove the Law of Cosines in the case that the angle with measure $\gamma$ is obtuse. Be sure to include a sketch.

18. Let $v = (v_1, v_2)$ and $w = (w_1, w_2)$ be nonzero vectors in the plane, and let $\theta$ be the measure of the angle between them. (We assume $0 \leq \theta \leq \pi$.) Recall the definition of the dot product of $v$ and $w$:

$$v \cdot w = v_1 w_1 + v_2 w_2.$$ 

Prove that $v \cdot w = \|v\|\|w\| \cos \theta$. (Hint: apply the Law of Cosines to the triangle with vertices $v$, $w$, and $0$.)

19. Using the previous exercise, prove that $v \cdot w > 0$ if the angle with measure $\theta$ is acute, $v \cdot w < 0$ if the angle with measure $\theta$ is obtuse, and $v \cdot w = 0$ if $\theta = \pi/2$.

4. Values of the Trigonometric Functions: Ptolemy’s Theorem, Identities, and Tables

The Greek astronomer Aristarchus needed a value for $\sin 3^\circ$ to compute the distance to the sun in terms of the distance to the moon (see Exercise 22). Unfortunately, Aristarchus lacked an accurate trigonometric table (that is, he didn’t know $\sin 3^\circ$ with great accuracy), and his computation suffered accordingly.

The trouble is that determining values of the trigonometric functions is non-trivial. There are only a few angles (for example, those of radian measure 0, $\pi/6$, $\pi/4$, $\pi/3$, and $\pi/2$) for which we can compute these values. However, in order for trigonometry to be useful, one needs to be able to compute values of the trigonometric functions for virtually every angle under the sun.

The oldest treatise on trigonometry that has survived to modern times is Claudius Ptolemy’s *Almagest*\(^\text{13}\) (ca. A.D. 100). The *Almagest* contains a table of sines that scores high marks for both accuracy and completeness. Computations for these tables were accomplished by exploiting trigonometric identities, which may be thought of as fundamental yet non-obvious relationships among the trigonometric functions. Trigonometric identities often have their origin in plane geometry. In this section we carry on in the spirit of Ptolemy by showing how the Law of Sines may be combined with plane geometry to produce powerful information about the trigonometric functions, including trigonometric tables.

---

\(^{13}\)Actually, the *Almagest* is a treatise on astronomy, and was the greatest work of its kind from the time of its writing until the 16th century. In fact ‘Almagest’ literally means ‘the greatest’ in Arabic.
4.1. Ptolemy’s Theorem. We explore a geometric result attributed to Ptolemy, now known as Ptolemy’s Theorem, which is a key ingredient in producing useful trigonometric identities. However, before arriving at Ptolemy’s Theorem, we must begin with a result from plane geometry.

To set the stage, recall that if $A, B, C$ are points on a given circle, then the arc of the circle intercepted by $\angle BAC$ is defined to be the circular arc determined by points $B, C$ that does not contain the point $A$. (see Figure 26).

**Figure 26**

**Proposition 4.** Suppose points $A, B, C, D, E, F$ are on a given circle. If the arc intercepted by $\angle ABC$ is equal in length to that intercepted by $\angle DEF$, then $\text{measure}(\angle ABC) = \text{measure}(\angle DEF)$.

**Figure 27**

**Proof.** Let $O$ be the center of the circle as in Figure 27. The hypotheses imply that the circular arcs corresponding to $\angle AOC$ and $\angle DOF$ have equal length. This implies that $\angle AOC$ and $\angle DOF$ have equal radian measure, and hence by Proposition 2, $\angle ABC$ and $\angle DEF$ have equal measure. $\square$
Theorem 5. (Ptolemy’s Theorem) Let $ABCD$ be a cyclic quadrilateral\footnote{a quadrilateral in which the order of the vertices is $ABCD$, either clockwise or counterclockwise.} inscribed in a circle. Then

$$|AC||BD| = |AB||DC| + |BC||AD|. \quad (Ptolemy’s Theorem)$$

Proof. Note that $\angle BCA$ and $\angle BDA$ have the same measure, since they intercept the same arc on the circle (see Proposition 4). Now, select $E$ on $AC$ such that $\angle ABE$ and $\angle DBC$ have the same measure (see Figure 28). It follows that triangles $ABD$ and $EBC$ are similar, as are triangles $ABE$ and $DBC$, therefore

\[
\frac{|BC|}{|EC|} = \frac{|BD|}{|AD|} \quad \text{and} \quad \frac{|AB|}{|AE|} = \frac{|BD|}{|DC|}. \]

We conclude that

$$|BD||AC| = |BD|(|EC| + |AE|) = |AD||BC| + |AB||DC|. \quad \blacksquare$$

Your Turn 18. Paraphrase Ptolemy’s Theorem in terms of the sides and diagonals of the cyclic quadrilateral $ABCD$.

Your Turn 19. Suppose the inscribed quadrilateral $ABCD$ is a rectangle. What famous theorem does Ptolemy’s Theorem imply in this case?
4.2. Useful identities. Ptolemy’s Theorem may be used to give geometric justification for the following identities, which are arguably the most important identities in all of trigonometry:

**Theorem 6.** (Sum formulas) If \( \alpha \) and \( \beta \) are angle measures, then:

(i) \( \sin(\alpha + \beta) = \sin \alpha \cos \beta + \cos \alpha \sin \beta \).

(ii) \( \sin(\alpha - \beta) = \sin \alpha \cos \beta - \cos \alpha \sin \beta \).

**Proof.** The theorem holds for all angles (as we shall see in Section 5.1), but for now we only consider angles with radian measure satisfying \( 0 \leq \beta \leq \alpha \leq \pi/2 \).

For part (i), consider a circle of diameter 1 containing a cyclic quadrilateral \( ABCD \) such that

- \( AC \) is a diameter of the circle,
- measure(\( \angle BAC \)) = \( \alpha \), and
- measure(\( \angle DAC \)) = \( \beta \).

The situation is summarized in Figure 29. By Your Turn 17, we see that \( |DB| = \sin(\alpha + \beta) \). Further, from Proposition 2 we know that \( \angle ADC \) and \( \angle ABC \) are right angles, and putting this together with the fact that \( |AC| = 1 \) gives

\[
|AD| = \cos \beta \quad |DC| = \sin \beta \quad |BC| = \sin \alpha \quad |AB| = \cos \alpha.
\]

Therefore part (i) follows by Ptolemy’s Theorem (Theorem 5).
One could derive the difference formula (ii) from (i) using the fact that sine is even and cosine is odd (see Exercise 19). However, we give a geometric derivation: Consider a circle of diameter 1 containing a cyclic quadrilateral $ABCD$ in which:

- $AD$ is a diameter.
- Both $B$ and $C$ lie on one of the semi-circular arcs determined by $A$ and $D$.
- $\text{measure}(\angle BAD) = \alpha$ and $\text{measure}(\angle CAD) = \beta$.

The situation is summarized in Figure 30. From Your Turn 17 and Proposition 2 we see that

\begin{align*}
|AB| = \cos \alpha, \quad |BC| = \sin(\alpha - \beta), \quad |CD| = \sin \beta, \quad |AC| = \cos \beta, \quad |BD| = \sin \alpha.
\end{align*}

Part (ii) follows immediately from (5) by Ptolemy’s Theorem. $\Box$

From Theorem 6 one may deduce a smattering of related identities, such as:

**Corollary 7.** For angle measures $\alpha$ and $\beta$ we have

(i) (Sum formula for cosine) $\cos(\alpha + \beta) = \cos \alpha \cos \beta - \sin \alpha \sin \beta$.

(ii) (Difference formula for cosine) $\cos(\alpha - \beta) = \cos \alpha \cos \beta + \sin \alpha \sin \beta$.

(iii) (Double angle identities) $\sin(2\alpha) = 2\sin \alpha \cos \alpha$ and $\cos(2\alpha) = \cos^2 \alpha - \sin^2 \alpha$.

(iv) (Half angle identities) $\sin \left( \frac{\alpha}{2} \right) = \pm \sqrt{\frac{1 - \cos \alpha}{2}}$ and $\cos \left( \frac{\alpha}{2} \right) = \pm \sqrt{\frac{1 + \cos \alpha}{2}}$.

Proofs of these identities are addressed in Exercise 20.
Your Turn 20. One of the most common errors in trigonometry is the assumption that \( \sin(\alpha + \beta) = \sin \alpha + \sin \beta \).

(a) Produce numerical and graphical evidence showing that this assumption is false.

(b) What are some possible sources of this misconception? Discuss.

4.3. Trigonometric tables and the problem of tedious multiplication. One reason why trigonometric identities have been so important is that they enable us to compute exact values of the trigonometric functions for many angles. In the past these identities were used to construct trigonometric tables, from which one could read the values of the trigonometric functions at a large number of angles. For example, in the Almagest Ptolemy produces a trigonometric table for angles between 0° and 90°, in \( \frac{1}{2} \)° intervals. Following Ptolemy, we create our own trigonometric tables. Our starting point is the fact that

\[
\sin 72° = \frac{1}{2} \sqrt{\frac{5 + \sqrt{5}}{2}}.
\]

Ptolemy computed \( \sin 72° \) using geometric techniques that we shall not explore. So, for now, we will take this result for granted. (An outline of a proof is given in Exercise 18.) Let’s see how this helps us create a trigonometric table:

Your Turn 21. Here we create our own trigonometric table, with 3° increments. You may use a calculator if you wish, but please restrict yourself to using \(+, -, \times, \div, \sqrt{\cdot}\).

(i) Find the exact values of \( \sin 12° \) and \( \cos 12° \) without using your calculator. (Hint: Use your knowledge of the trigonometric functions at special angles together with trigonometric identities and the values of sine and cosine at 72°.)

(ii) Using trigonometric identities and your calculator, find the sine and cosine of 6° and 3°.

(iii) Starting with values \( \sin 3° \) and \( \cos 3° \), use the sum identities to find values of sine and cosine for all angles between 3° and 30°, in increments of 3°. Put your results in tabular form.

(iv) Use the result of part (ii) to improve the estimate of the distance from the earth to the sun given in part (c) of Exercise 22.
Even with the aid of the basic functions on your calculator, the calculations required to build your table (especially part (iv) above) were probably quite tedious\textsuperscript{15}. Now, imagine having to do all of those computations \textit{by hand}, especially the multiplications that occur in the sum identities. These computations proved to be a barrier in constructing accurate, extensive trigonometric tables until the advent of the logarithm in the late 16th century. As we shall see in Chapters 9 and 10, logarithms have the ability to transfer multiplication problems (which are often tedious) to simpler addition problems.

4.4. Exercises.

1. Check that Ptolemy’s Theorem is correct for $A = (2, 0), B = (1, \sqrt{3}), C = (-1, \sqrt{3}), D = (0, -2)$. Give an equation for the circle on which these four points lie.

2. Suppose that you pick a cyclic quadrilateral and after careful computation, you discover that the formula in Ptolemy’s Theorem does not hold. What can you conclude about the quadrilateral?

3. In teaching a trigonometry class, you want to prepare an example illustrating the sum formula for sine: Find reasonable angles $\alpha$ and $\beta$ such that you know exact values for $\sin(\alpha + \beta)$, $\sin \alpha$, $\sin \beta$, $\cos \alpha$, and $\cos \beta$.

4. One can illustrate the sum formula for sine using more treacherous angles, if one is willing to use a calculator to compute values. Check that the sum formula is correct (possibly up to rounding error) for $\alpha = 2$ radians and $\beta = 3$ radians.

5. Use the trigonometric table you produced in Your Turn 21 to estimate $\sin 4.8^\circ$. (Hint: Find the equation $y = mx + b$ of the line passing through the points $(3, \sin 3^\circ)$ and $(6, \sin 6^\circ)$. What is the $y$-coordinate of the point on the line whose $x$-coordinate is 4.8? This method is known as \textit{linear interpolation}.)

6. In this exercise we verify that $\sin(\frac{\pi}{2} - \alpha) = \cos(\alpha)$ and $\cos(\frac{\pi}{2} - \alpha) = \sin(\alpha)$.

   (i) Using the right triangle interpretations of sine and cosine, draw two convincing figures that illustrate these identities in the case $0 \leq \alpha \leq \frac{\pi}{2}$.

\textsuperscript{15}In the modern world we approximate values of the trigonometric functions using Taylor series, thus decreasing our direct dependence on trigonometric identities for the purpose of computation.
(ii) Verify that \( \sin\left(\frac{\pi}{2} - \alpha\right) = \cos(\alpha) \) using part (ii) of Theorem 6. Then, use this identity to verify \( \cos\left(\frac{\pi}{2} - \alpha\right) = \sin(\alpha) \).

7. Find \( \cos 15^\circ \) and \( \sin 15^\circ \) using the half-angle identity.

8. Find an expression for \( \cos(2\theta) \) that only involves \( \cos \theta \). Then, find an expression for \( \cos(2\theta) \) that only involves \( \sin \theta \).

9. Show that \( \sin^2 \theta = \frac{1}{2}(1 - \cos 2\theta) \).

10. Verify that \( \tan(\theta + \phi) = \frac{\tan \theta + \tan \phi}{1 - \tan \theta \tan \phi} \).

11. Verify that \( \tan(2\theta) = \frac{2 \tan \theta}{1 - \tan^2 \theta} \).

12. Verify \(^{16}\) that \( (\sin \theta)(\sin \phi) = \frac{1}{2} \cos(\theta - \phi) - \frac{1}{2} \cos(\theta + \phi) \).

13. Prove\(^{17}\) that \( \sin \theta + \sin \phi = 2 \sin \left(\frac{\theta + \phi}{2}\right) \cos \left(\frac{\theta - \phi}{2}\right) \).

14. Suppose that you know (via tables or otherwise) that \( \sin(0.902) \approx \frac{1.57}{2} \), \( \sin(1.097) \approx 0.89 \), \( \cos(1.999) \approx -0.415 \), and \( \cos(1.195) \approx 0.981 \). Use the identity given in Exercise 12 to estimate \( 1.57 \times 0.89 \) without performing any multiplications. \( \text{This exercise may seem silly to those of us who have modern calculators, but historically, this approach really was used for computations.} \)

15. Using trigonometric identities, express \( \cos(3\theta) \) in terms of \( \cos \theta \) and/or \( \sin \theta \). Do the same for \( \cos(4\theta) \).

16. Use geometry software or a compass to construct a large circle. Mark off an angle of \( 72^\circ \) and make measurements leading to an estimate of \( \sin(72^\circ) \). Compare your answer to (a decimal approximation of) the true value \( \sin(72^\circ) = \frac{\sqrt{5 - \sqrt{5}}}{4} \).

17. Suppose points \( A, B, C \) lie on a circle of radius 5 units and that \( \angle ABC \) is \( 20^\circ \). Find the length of the arc \( AC \) on the circle.

\(^{16}\)There are similar identities for \( (\sin \theta)(\cos \phi) \) and \( (\cos \theta)(\cos \phi) \).

\(^{17}\)There are similar identities for \( \sin \theta + \cos \phi \) and \( \cos \theta + \cos \phi \).
18. In this exercise we compute \( \cos(72^\circ) \) and \( \sin(72^\circ) \).

(i) Let \( \theta = 72^\circ \). Show that \( 16 \cos^4 \theta - 12 \cos^2 \theta + 1 = 0 \). (Hint: Since \( 5\theta = 360^\circ \), we know \( \sin(5\theta) = 0 \).

By writing \( 5\theta = 3\theta + 2\theta \), one may use the identities from Theorem 6 and Corollary 7 to begin breaking \( \sin(5\theta) \) down into sums and products of \( \sin \theta \) and \( \cos \theta \). Use the Pythagorean identity \( \sin^2 \theta + \cos^2 \theta = 1 \) to convert everything to cosines.

(ii) Substitute \( y = \cos^2 \theta \) in the equation given in part (i). Solve the equation for \( y \). Be sure to eliminate any extraneous solutions.

(iii) From your results in part (ii), verify that \( \cos \theta = \frac{1}{2} \sqrt{\frac{3-\sqrt{5}}{2}} \) and \( \sin \theta = \frac{1}{2} \sqrt{\frac{3+\sqrt{5}}{2}} \).

19. Deduce part (ii) of Theorem 6 from part (i) by using the fact that sine is odd and cosine is even.

20. Here we prove various parts of Corollary 7

(a) Prove parts (i) and (ii) of Corollary 7. (Hint: Use Exercise 6.)

(b) Prove part (iii) of Corollary 7. (Hint: use part (a) of this exercise.)

(c) Prove part (iv) of Corollary 7.

5. Trigonometry, Coordinate Geometry, and Linear Algebra

The earliest origins of trigonometry lie in Euclidean geometry, which has been the bedrock for most of the important trigonometric results we have seen thus far (e.g., the sum identities). In this section however, our purpose is to discuss how relatively modern ideas from coordinate geometry and linear algebra also contribute to our understanding of trigonometry. Along the way we discuss several important applications, including a return to the sum identities for sine and cosine, as well as a geometric interpretation of complex multiplication.

5.1. The sum formulas revisited. By means of coordinate-geometric construction, we first describe the end result of applying a rotation to a point in the plane. We should not be surprised to see that angles, which have previously been defined as the dynamic process of rotation, play a significant role. The construction, shown in Figure 31, is accomplished as follows:

---

18 These ideas date from the 18-th and 19-th centuries.
We draw two circles centered at the origin, of radius 1 and radius $r$.

The circle of radius 1 meets the positive $x$-axis at the point $e_1$ and meets the positive $y$-axis at the point $e_2$. We have

$$e_1 = (1, 0) \quad \text{and} \quad e_2 = (0, 1).$$

These points are rotated by an angle with measure $\theta$, producing new points $e'_1$ and $e'_2$. By the definitions of the sine and cosine functions, these points have coordinates:

$$e'_1 = (\cos \theta, \sin \theta) \quad \text{and} \quad e'_2 = (-\sin \theta, \cos \theta).$$

Let $P$ be the point on the circle of radius $r$, with angle measure $\theta$. In coordinates, by the definition of the sine and cosine functions,

$$P = (r \cos \theta, r \sin \theta) = (r \cos \theta)e_1 + (r \sin \theta)e_2.$$

Let $R_\phi(P)$ be the point obtained by rotating $P$ by the angle with measure $\phi$. Since the distance from $R_\phi(P)$ to the origin is $r$, and since $R_\phi(P)$ makes an angle with measure $\phi + \theta$ with the positive $x$-axis, we know that in coordinates,

$$R_\phi(P) = (r \cos(\phi + \theta), r \sin(\phi + \theta)) = r (\cos(\phi + \theta), \sin(\phi + \theta)).$$
This construction becomes useful when we notice that there is an alternate description of \( R_\phi(P) \). We imagine new “rotated” axes extending out through \( e'_1 \) and \( e'_2 \). Relative to these axes, \( R_\phi(P) \) makes an angle with measure \( \phi \), and hence

\[
R_\phi(P) = r \cos \phi e'_1 + r \sin \phi e'_2
\]

\[
= (r \cos \phi)(\cos \theta, \sin \theta) + (r \sin \phi)(-\sin \theta, \cos \theta)
\]

\[
= r(\cos \phi \cos \theta, \cos \phi \sin \theta) + r(-\sin \phi \sin \theta, \sin \phi \cos \theta)
\]

\[
= r(\cos \phi \cos \theta - \sin \phi \sin \theta, \cos \phi \sin \theta + \sin \phi \cos \theta).
\]

**Your Turn 22.** Here we revisit the sum formulas for sine and cosine.

(a) Use the descriptions of \( R_\phi(P) \) given above to show that

\[
\cos(\phi + \theta) = \cos \phi \cos \theta - \sin \phi \sin \theta
\]

\[
\sin(\phi + \theta) = \cos \phi \sin \theta + \sin \phi \cos \theta
\]

for any angles \( \phi \) and \( \theta \).

(b) Compare the current version of the sum formulas to that given in Section 4. Is the current version stronger, weaker, or equivalent to that of the previous section?

5.2. The rotation matrix and rotated coordinate systems. Given the vector\(^{19}\) \( P = \begin{pmatrix} r \cos \theta \\ r \sin \theta \end{pmatrix} \), we may rewrite (6) as

\[
R_\phi(P) = r \begin{pmatrix} \cos \phi \cos \theta - \sin \phi \sin \theta \\ \cos \phi \sin \theta + \sin \phi \cos \theta \end{pmatrix} = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} \begin{pmatrix} r \cos \theta \\ r \sin \theta \end{pmatrix} = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} P.
\]

This shows:

**Proposition 8.** The function \( R_\phi : \mathbb{R}^2 \to \mathbb{R}^2 \), given by rotation around the origin by the angle with measure \( \phi \), is a linear transformation with matrix \( \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} \).

These ideas are useful for rotating coordinate systems. From Figure 31, we see that any point \( P \) has two sets of coordinates. The first, denoted \((x, y)\), come from using the axes in the directions of \( e_1 \) and \( e_2 \). The second, denoted \((x', y')\), come from using the axes in the directions of \( e'_1 \) and \( e'_2 \).

\(^{19}\)Here, in accordance with the usual practice in linear algebra, we write vectors as column vectors.
How are the coordinates \((x, y)\) and \((x', y')\) related to each other? We can figure this out by looking at the point \(R_\phi(P)\) in Figure 31. We’ll call it \((x, y)\) in the usual coordinates; in terms of the angles that we have drawn, we have that \(x = r \cos(\phi + \theta)\) and \(y = r \sin(\phi + \theta)\). Now what are the \((x', y')\)-coordinates of \(R_\phi(P)\)? We already computed
\[
R_\phi(P) = r \cos \phi e'_1 + r \sin \phi e'_2.
\]
This tells us that
\[
x' = r \cos \phi \quad \text{and} \quad y' = r \sin \phi.
\]

So we have \(x = r \cos(\phi + \theta), y = r \sin(\phi + \theta), x' = r \cos \phi,\) and \(y' = r \sin \phi\). We’re ready to exhibit the relationship between \((x, y)\) and \((x', y')\):
\[
\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \cos(\phi + \theta) \\ r \sin(\phi + \theta) \end{pmatrix} = \begin{pmatrix} r(\cos \phi \cos \theta - \sin \phi \sin \theta) \\ r(\sin \phi \cos \theta + \cos \phi \sin \theta) \end{pmatrix} = \begin{pmatrix} x' \cos \theta - y' \sin \theta \\ y' \cos \theta + x' \sin \theta \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}.
\]

Summarizing our calculations:

**Proposition 9.** Suppose that new coordinate axes are drawn by rotating the standard axes by an angle with measure \(\theta\). Let \((x, y)\) denote coordinates with respect to the standard axes and let \((x', y')\) denote coordinates with respect to the rotated axes. Then for any point, its \((x, y)\)- and \((x', y')\)-coordinates are related by the rule
\[
x = x' \cos \theta - y' \sin \theta \\
y = x' \sin \theta + y' \cos \theta
\]
or in matrix notation,
\[
\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}.
\]

Also,
\[
\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.
\]

This proposition is useful. For example, suppose we are interested in finding the equation of an ellipse similar to the ellipse with equation \(\frac{x^2}{4} + \frac{y^2}{9} = 1\), except rotated by \(\pi/6 = 30^\circ\) (see Figure 32). In the rotated coordinate system, the equation is precisely \(\frac{(x')^2}{4} + \frac{(y')^2}{9} = 1\). To write this equation in \((x, y)\)-coordinates,
we merely must make the substitution

\[ x' = x \cos \theta + y \sin \theta = \frac{\sqrt{3}}{2} x + \frac{1}{2} y \]
\[ y' = y \cos \theta - x \sin \theta = -\frac{1}{2} x + \frac{\sqrt{3}}{2} y. \]

Thus in \((x, y)\)-coordinates, the rotated ellipse has equation

\[ \left( \frac{\sqrt{3}}{2} x + \frac{1}{2} y \right)^2 + \left( -\frac{1}{2} x + \frac{\sqrt{3}}{2} y \right)^2 = 1, \]

which can be written as

\[ 31x^2 + 10\sqrt{3}xy + 21y^2 = 144. \]

**Figure 32**

**5.3. Complex multiplication.** Multiplication of complex numbers \( z = a + bi \) and \( w = c + di \) \((a, b, c, d \in \mathbb{R})\) is given by the rule

\[ zw = (ac - bd) + (ad + bc)i. \]

We can use this definition to compute the products, but it does not give any geometric information about the meaning of complex multiplication. For that we use trigonometry.

More precisely, we can identify a complex number \( z = x + yi \) with the point \((x, y)\) in the \(xy\)-plane, which (see Figure 31) we can write as

\[ (x, y) = (r \cos \theta, r \sin \theta). \]

Therefore

\[ z = (r \cos \theta) + (r \sin \theta)i = r(\cos \theta + i \sin \theta), \]
as shown in Figure 33 below. The expression for \( z \) given in (8) is referred to as the polar representation of \( z \), where \( r \) is the *modulus* (or *absolute value*) of \( z \), and \( \theta \) is an argument for \( z \).

**Figure 33**

Now, using the sum formulas for sine and cosine, we can discover the geometric interpretation of complex multiplication:

**Theorem 10.** Given complex numbers \( z_1 = (r_1 \cos \theta_1) + (r_1 \sin \theta_1)i \) and \( z_2 = (r_2 \cos \theta_2) + (r_2 \sin \theta_2)i \), we have

\[
z_1z_2 = r_1r_2 (\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)).
\]

**Your Turn 23.**

(a) Use the sum identities for sine and cosine to verify Theorem 10.

(b) Theorem 10 says that when complex numbers are multiplied, one must _________ their lengths and _________ their angle measures. Use this principle to estimate the value of \((2 - i)(3 + \frac{1}{2}i)\) graphically.

In addition to providing geometric insight into complex multiplication, Theorem 10 has the virtue of being useful, especially when it comes to computing repeated products of complex numbers: Suppose \( z = r(\cos \theta + i \sin \theta) \) is a complex number and \( n \) is a whole number. Then

\[
z^n = r^n (\cos(n\theta) + i \sin(n\theta)).
\]

This equation is known as *DeMoivre’s formula*.\(^{20}\)

\(^{20}\)DeMoivre’s formula, which more or less marks the beginnings of trigonometry in analysis, is named after French mathematician Abraham DeMoivre (1667-1754).
To see how DeMoivre’s formula might be useful, suppose we wish to compute $(-1 + i)^7$. We could compute this by painfully repeating ordinary complex multiplication as given in Equation (7). On the other hand, observing that $-1 + i = \sqrt{2}(\cos 3\pi/4 + i \sin 3\pi/4)$, DeMoivre’s formula gives the result with relative ease:

$$(-1 + i)^7 = [\sqrt{2}(\cos(3\pi/4) + i \sin(3\pi/4))]^7 = 8\sqrt{2}[\cos(21\pi/4) + i \sin(21\pi/4)] = -8 - 8i.$$ 

5.4. Exercises.

1. Write down the matrix for a 60° rotation, and then use this matrix to rotate the point (5, 9) by 60°.

2. If the line $y = 7x$ is rotated by 60°, what is the equation of the new line?

3. Consider the ellipse with equation $\frac{x^2}{25} + \frac{y^2}{16} = 1$.
   
   (a) What are the coordinates of the four vertices?

   (b) Suppose that the ellipse is rotated by $\pi/3$. What are the coordinates of the four vertices?

4. If a particle begins at the point (4, 11) and is rotated by $\pi/4$ around the origin, what is its new position?

5. The parabola with equation $y = x^2$ is rotated 45°. What is the equation of the rotated parabola?

6. Let $z = r(\cos \theta + i \sin \theta)$ be a nonzero complex number. Prove that $z^{-1} = \frac{1}{r}(\cos \theta - i \sin \theta)$.

7. In Figure 34, two complex numbers $z$ and $w$ are marked. Plot and label the following complex numbers: $zw, z^2, w^2, z^{-1}, w^{-1}, \overline{z}, \overline{w}$, and $z + w$.

8. Let $z$ and $w$ be the complex numbers on the unit circle with angles with measures 30° and 150°, respectively (see Figure 35). Plot $z^0, z^1, z^2, \ldots, z^{11}, z^{12}, z^{13}$ and $w^0, w^1, w^2, \ldots, w^{11}, w^{12}, w^{13}$. Finally, plot $z^{100}$ and explain a quick way of finding this point.

9. (a) Find the complex number $z$ of length 1 that makes an angle of 60° with the positive $x$-axis.
(b) Use $z$ to rotate the complex number $5 + 9i$ by 60°.

10. Prove that $\sin \theta + \sin 2\theta + \cdots + \sin(n-1)\theta + \sin n\theta = \frac{\sin \left(\frac{(n+1)\theta}{2}\right) \sin \left(\frac{n\theta}{2}\right)}{\sin \left(\frac{\theta}{2}\right)}$.

11. A standard ellipse is rotated by an angle with measure $\theta$ (with $0 < \theta < \pi/2$) to produce a rotated ellipse, and the rotated ellipse has equation $98x^2 - 72xy + 77y^2 = 5000$. Find $\cos \theta$ and $\sin \theta$, and the equation

\[ \text{This is the sort of exercise that will become much easier after we learn about the connection between } e^x \text{ and } \sin x. \]
of the rotated ellipse in \((x', y')\)-coordinates. (Hint: by substitution, rewrite the equation in terms of \(x', y',\) and \(\theta\); what choice of \(\cos \theta\) and \(\sin \theta\) eliminates the \(x'y'\) term?)

12. Verify DeMoivre’s formula given in Equation (9).

13. (a) Let \(\theta\) be a real number and let \(r\) be a nonnegative real number. Use the quadratic formula to find the roots of \(x^2 - (2r \cos \theta)x + r^2\).

(b) The roots are not real numbers. How are they related to each other?

(c) What is the geometric significance of \(\theta\) in this problem?

(d) What is the angle measure \(\theta\) for the polynomial \(x^2 - \frac{28}{3}x + 49\)?

6. Application: Why We Use Radians in Calculus

We remember from calculus that the functions \(\cos x\) and \(\sin x\) are differentiable and that their derivatives are \(-\sin x\) and \(\cos x\), respectively. We might also remember that we identify the domains of these functions (i.e., the allowable input values \(x\)) with \textit{radian} angle measures, \textbf{not} degree measures. In this section we investigate our preference for radians.

\textbf{Figure 36}

The story begins, as does everything in calculus, with \textit{limits}. Consider Figure 36, featuring a small positive angle \(\theta\). Observe from the figure that the area of the circular sector \(OAC\) is bounded above and
below by the areas of triangles $OCD$ and $OAB$, respectively. Regarding $\theta$ as a radian angle measure, and assuming that the area of a circle is $\pi$ times the square of its radius, this implies that

$$\frac{1}{2} \sin \theta \cos \theta \leq \frac{\theta}{2} \leq \frac{1}{2} \tan \theta,$$

so that upon dividing by $\frac{1}{2} \sin \theta \cos \theta$ we obtain

$$1 \leq \frac{\theta}{\sin \theta \cos \theta} \leq \sec^2 \theta.$$

Since $\sec^2 \theta$ approaches 1 as $\theta$ approaches 0 through positive values, by the ‘Squeeze’ theorem we observe that $\frac{\theta}{\sin \theta \cos \theta} \to 1$ as $x \to 0^+$. Putting this together with the fact that $\cos x \to 1$ as $x \to 0^+$ gives

$$\lim_{\theta \to 0^+} \frac{\sin \theta}{\theta} = 1. \quad (10)$$

Equation (10) together with a similar argument for small negative angles $\theta$ (again, in radian measure) allows us to conclude that

$$\lim_{\theta \to 0} \frac{\sin \theta}{\theta} = 1, \quad \text{which in turn implies} \quad \lim_{\theta \to 0} \frac{\cos \theta - 1}{\theta} = 0. \quad (11)$$

Using (11) together with the sum identities (Theorem 6) we are ready to compute a derivative:

$$\frac{d}{dx} \sin x = \lim_{\theta \to 0} \frac{\sin(x + \theta) - \sin x}{\theta}$$

$$= \lim_{\theta \to 0} \frac{\sin x \cos \theta + \cos x \sin \theta - \sin x}{\theta}$$

$$= \sin x \left[ \lim_{\theta \to 0} \frac{\cos \theta - 1}{\theta} \right] + \cos x \left[ \lim_{\theta \to 0} \frac{\sin \theta}{\theta} \right]$$

$$= \sin x \cdot 0 + \cos x \cdot 1 = \cos x,$$

as expected. A similar argument produces the derivative of the cosine function, and then the quotient rule can be used to find the derivatives of the other four trigonometric functions.

Now we suppose the domains of the sine and cosine functions correspond to degree angle measures instead of radians. We shall denote the outputs of these functions by $S(x)$ and $C(x)$, respectively, where $x$ is the degree measure of an angle. Referring once again to Figure 36, if $\theta$ is a degree angle measure, then, upon converting to radians, the area of the circular sector $OAC$ is $\frac{\pi}{180} \cdot \frac{\theta}{2}$ instead of $\frac{\theta}{2}$. Making this replacement in the arguments above, we find that

$$\lim_{\theta \to 0} \frac{S(\theta)}{\theta} = \frac{\pi}{180}, \quad \text{and} \quad \lim_{\theta \to 0} \frac{C(\theta) - 1}{\theta} = 0. \quad (12)$$
Using (12) together with the sum identities (Theorem 6) we compute $S'(x)$ using the definition of the derivative:

$$
\frac{d}{dx}S(x) = \lim_{\theta \to 0} \frac{S(x + \theta) - S(x)}{\theta} = \lim_{\theta \to 0} \frac{S(x)C(\theta) + C(x)S(\theta) - S(x)}{\theta} = S(x) \left[ \lim_{\theta \to 0} \frac{C(\theta) - 1}{\theta} \right] + C(x) \left[ \lim_{\theta \to 0} \frac{S(\theta)}{\theta} \right] = S(x) \cdot 0 + C(x) \cdot \frac{\pi}{180} = \frac{\pi}{180}C(x).
$$

A similar argument produces the derivative of the $C(x)$. It is evident that the derivatives of the trigonometric functions take on their simplest form when we identify their domains with radian angle measures.

**Your Turn** 24. We investigate derivatives of the cosine function.

(a) Use the definition of the derivative to compute $\frac{d}{dx} \cos x$.

(b) Use the definition of the derivative to compute $C'(x)$.

**Your Turn** 25. The definition of $S$ and $C$ imply that $S(x) = \sin \left( \frac{x}{180} \right)$ and $C(x) = \cos \left( \frac{x}{180} \right)$. Use the chain rule to show that $S'(x) = \frac{\pi}{180}C(x)$ and $C'(x) = -\frac{\pi}{180}S(x)$.

6.1. Exercises.

1. Argue that $\lim_{\theta \to 0} \frac{\sin \theta}{\theta} = 1$. (Hint: The fact that sine is an odd function may be helpful.)

2. Show that $\lim_{\theta \to 0} \frac{\sin \theta}{\theta} = 1$ implies $\lim_{\theta \to 0} \frac{\cos \theta - 1}{\theta} = 0$.

3. Find a formula for $S^{(n)}(x)$, the $n$-th derivative of $S(x)$.

4. Suppose that $T(x) = S(x)/C(x)$. Find the derivative of $T(x)$.

7. Application: Periodic Data and Trigonometric Polynomials

We are surrounded by repeating processes. Examples include our own body temperature and heartbeat, certain predator-prey population dynamics, and the motion of planets. In order to model these and other periodic processes effectively, it seems reasonable that we use basic functions which are themselves periodic, such as the trigonometric functions.
The idea of using trigonometric functions to model periodic processes is not new. In the late 1700’s Daniel Bernoulli used infinite sums of sine and cosine functions to model the motion of a vibrating string, while in the early 1800’s Jean Fourier used similar sums of functions to determine heat flow in a rod. The process of expressing (or approximating) functions in terms of sines and cosines bears Fourier’s name, and is known today as Fourier analysis.

In this section we will explore elements of Fourier analysis that pertain to fitting periodic curves to data. In the process, we will explore a method for encoding and decoding messages.

7.1. Trigonometric polynomials and data. Our goal is to fit sums of sines and cosines to certain data sets \( \{(t_j, y_j)\} \) such as the one shown in Figure 37\(^{22}\). For the sake of simplicity, we begin by assuming that the data is more or less periodic in \( t \) of period \( 2\pi \), and that our analysis will occur over the interval \(-\pi \leq t \leq \pi\). We remember that \( \sin t \) and \( \cos t \) are periodic of period \( 2\pi \), so our first attempt to model the data might be a simple linear combination of sines and cosines, say \( p(t) = 2\cos t - \sin t \). Unfortunately, when we graph \( p \) against the data (see Figure 38), we see that our model falls well short of closely approximating the data.

**Your Turn** 26. Consider Figure 38. List at least two specific characteristics of \( p \) that cause it to fall short of accurately modeling the data.

\(^{22}\)This is a hypothetical data set. It may help to think of \( t \) as representing time and \( y \) as a roughly periodic function of \( t \), such as temperature or population.
We can make the average value of $p$ equal to the average value of the data by adding an appropriate constant to $p$, and we can increase the oscillatory behavior of $p$ by adding on ‘higher frequency’ sines and cosines, such as $\sin 2t$, $\cos 2t$, $\sin 3t$, $\cos 3t$, et cetera (see Exercises 1 and 2). These adjustments allow for significant improvement in approximation. For example, Figure 39 shows that $q(t) = 2 - \cos t + 3.5 \sin 2t + 5 \cos 2t - 5 \sin 5t$ approximates the data set in Figure 37 quite nicely. (In fact, if we add on enough higher frequency terms we can produce a function $p(t)$ whose graph passes through each of the data points.) All of this leads to the following definition, which describes a class of functions that should be very useful for modeling periodic data:
Definition 11. Let \( n \) be a non-negative integer. A trigonometric polynomial of degree \( n \) is a function of the form
\[
p(t) = \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos(kt) + b_k \sin(kt)],
\]
where \( a_k, b_k \in \mathbb{R} \) for all \( k \), and at least one of \( a_n, b_n \) is nonzero.

Your Turn 27. Without using a graphing utility, match the following trigonometric polynomials to their graphs in Figure 40. (Hint: First check whether each function is even, odd, or neither.)

(a) \(-5 \cos t + 7 \cos 2t - 5 \cos 3t\)
(b) \(-5 \sin t + 7 \sin 2t - 5 \sin 3t\)
(c) \(1 - 2 \sin t + \cos t\)
(d) \(3 - 5 \sin t + 7 \cos 2t - 5 \sin 3t\)

Your Turn 28. How is the notion of ‘degree’ for a trigonometric polynomial different from that for an ordinary polynomial?

7.2. A least squares minimization problem and orthogonality. Throughout the last section you may have wondered how we go about finding a trigonometric polynomial of degree less than or equal to a fixed number \( n \) that ‘best fits’ a set of \( m \) data points. To begin, suppose we are given a data set \( \{(t_j, y_j)\}_{j=1}^{m} \) of \( m \) data points where the \( t_j \)'s are evenly distributed\(^{23}\) (see Figure 41) throughout the interval \([-\pi, \pi]\), say,

\[
t_j = -\pi + \frac{2j}{m} \pi.
\]

\(^{23}\) The need for an even distribution of the \( t_j \)'s will be addressed shortly.
Also, we let

\[ p_n(t) = \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos kt + b_k \sin kt] \]

be a trigonometric polynomial of degree at most \( n \). In order for \( p_n(t) \) to fit the data set closely, we want the totality of differences \( |p_n(t_j) - y_j| \) to be as small. From a computational standpoint, the most feasible way to accomplish this is to minimize the sum of squared differences\(^{24}\)

\[
(14) \quad \text{err}(a_0, \ldots, a_n, b_1, \ldots, b_n) = \sum_{j=1}^{m} [p_n(t_j) - y_j]^2.
\]

Minimizing the \( \text{err} \) function is naturally a calculus problem. We find the critical points by setting the partial derivatives of \( \text{err} \) with respect to \( a_i \) and \( b_i \) equal to zero and then solving the resulting system of equations. A typical partial derivative for \( \text{err} \) is

\[
(15) \quad \frac{\partial \text{err}}{\partial b_l} = 2 \sum_{j=1}^{m} [p_n(t_j) - y_j] \sin lt_j,
\]

so, after expanding \( p_n(t_j) \) in Equation (15) and setting the partial derivative equal to zero, we find that a typical equation in the system (with unknowns \( a_k \) and \( b_k \)) determining the critical points for \( \text{err} \) is

\[
(16) \quad 0 = \sum_{j=1}^{m} \frac{a_0}{2} \sin lt_j + \sum_{k=1}^{n} \sum_{j=1}^{m} a_k \cos kt_j \sin lt_j + \sum_{k=1}^{n} \sum_{j=1}^{m} b_k \sin kt_j \sin lt_j - \sum_{j=1}^{m} y_j \sin lt_j.
\]

While not insurmountable, Equation (16) and its brethren certainly appear ugly, and our first impulse is to use \textit{Mathematica} (or some other mathematical software) to solve the system of equations. However, it turns out that by using properties of the trigonometric functions we can solve systems of equations like (16) elegantly. The key result is:

---

\(^{24}\)The method of minimizing a sum of squared differences was discussed in Chapter 2, Section 6. There are reasons for working with squared differences: In addition to making our computations easier, they heighten the influence of data points which lie ‘far away’ from the rest of the data without allowing them to dominate the approximation. Simply summing the (nonsquared) differences does not give enough weight to data points ‘out of line’ with the others. Another reason for using squared differences stems from the statistical distribution of error. For details on these issues one can consult a numerical analysis text (e.g., Burdon and Faires) and/or a text on statistical inference (e.g., Larson).
Theorem 12. (Orthogonality Relations) Suppose that $0 \leq k, l \leq n$, that $t_1, \ldots, t_m$ are as in Equation (13), and that $m \geq 2n + 1$. Let $\delta_{kl}$ be such that $\delta_{kl} = 1$ whenever $k = l$ and $\delta_{kl} = 0$ otherwise. Then

\[ \sum_{j=1}^{m} \sin kt_j \cos lt_j = 0, \quad \sum_{j=1}^{m} \sin kt_j \sin lt_j = \begin{cases} m^2 \delta_{kl} & \text{if } k + l > 0 \\ 0 & \text{if } k = l = 0 \end{cases}, \quad \text{and} \quad \sum_{j=1}^{m} \cos kt_j \cos lt_j = m^2 \delta_{kl}. \]

We delay the proof of Theorem 12 until Section 7.4, but we apply it now: Using the orthogonality relations in Theorem 12, we see that the Equation (16) boils down to

\[ 0 = b_l \frac{m}{2} - \sum_{j=1}^{m} y_j \sin lt_j, \quad \text{so that} \quad b_l = \frac{2}{m} \sum_{j=1}^{m} y_j \sin lt_j. \]

More generally, the coefficients of a best-fit trigonometric polynomial $a_0/2 + \sum_{k=1}^{n} [a_k \cos kt + b_k \sin kt]$ are found as follows:

\[ a_k = \frac{2}{m} \sum_{j=1}^{m} y_j \cos kt_j \quad \text{and} \quad b_k = \frac{2}{m} \sum_{j=1}^{m} y_j \sin kt_j. \]

The orthogonality relations will be proved in the exercises with the assistance of Lemma 13.

Your Turn 29. Consider the err function defined in Equation 14.

(a) Why do you think the function is named ‘err’?

(b) What is the domain of the err function? Why?

(c) What is the range of the err function? Why?

Your Turn 30. Where have you seen the word ‘orthogonal’ before, and what does it mean? Why do you think the word ‘orthogonal’ is used in Theorem 12?

7.3. A coding scheme: put a wave in the air. Best-fit trigonometric polynomials can be used in coding schemes, that is, the process of sending messages. For example, suppose Captain Kirk wants to send messages to Alice (his Earth girlfriend) while he’s away saving the universe aboard the starship Enterprise. During their whirlwind romance on Earth, Kirk and Alice agree that Kirk will send messages of no more than fifteen characters in length, that the alphabet shall correspond to the numbers 1 through 26 in the usual way ($a = 1$, $b = 2$, etc., and 0 represents a space), and that the text of these messages will be encoded in the coefficients of a cosine polynomial (a trigonometric polynomial with $b_1 = b_2 = \cdots = b_n = 0$, i.e., a trig polynomial with no ‘sine’ terms).
After months of roaming space, Kirk has decided in favor of a local girlfriend, so he wants to send Alice the message “I love Uhura.” To do this, Kirk first replaces each letter by the appropriate number, so that the message becomes “9 0 12 15 22 5 0 21 8 21 18 1.” Kirk uses these numbers to form coefficients for a standard cosine polynomial,

\[ p(t) = \frac{9}{2} + 12 \cos(2t) + 15 \cos(3t) + 22 \cos(4t) + 5 \cos(5t) + 21 \cos(7t) + 8 \cos(8t) + 21 \cos(9t) + 18 \cos(10t) + \cos(11t), \]

and then sends the message\(^{25}\) as a wave (see Figure 42).

![Figure 42. “I love Uhura” in wave form.](image)

Alice has equipment that will receive and sample this wave. Let’s suppose she chooses samples at 100 times \(t_1, t_2, \ldots, t_{100}\) that are evenly interspersed in the time interval \([-\pi, \pi]\), say

\[ t_j = -\pi + \frac{2j}{100} \pi \quad j = 1, \ldots, 100. \]

The samples \(\{(t_j, y_j)\}_{j=1}^{100}\) given by Alice’s machinery and the associated plot are shown in Figure 43.

In order to decode the message from the sample data \(\{(t_j, y_j)\}_{j=1}^{100}\), Alice needs to find a standard trigonometric polynomial \(d(t)\) that best fits this data (in the sense of least squares). Since the message is no more than 15 characters in length and comes in the form of a cosine polynomial, the desired trigonometric polynomial has the form

\[ d(t) = \frac{a_0}{2} + \sum_{k=1}^{14} a_k \cos(kt). \]

\(^{25}\)Actually Spock sends the message. Kirk is far too busy with administrative duties aboard the ship to send such messages himself.
To read the message, Alice must determine the numbers $a_0, \ldots, a_{14}$. As we saw previously (see Equation (17)), to do this she computes

$$a_k = \frac{2}{100} \sum_{j=1}^{100} y_j \cos(kt_j),$$

for $k = 0, 1, \ldots, 14$, from which she obtains

$a_0 = 9, a_1 = 8.73079 \times 10^{-15}, a_2 = 12, a_3 = 15, a_4 = 22, a_5 = 5, a_6 = 6.52811 \times 10^{-15}, a_7 = 21, a_8 = 8, a_9 = 21, a_{10} = 18, a_{11} = 1, a_{12} = 5.59108 \times 10^{-15}, a_{13} = 1.87850 \times 10^{-15},$ and $a_{14} = -7.04325 \times 10^{-15}$.

Upon rounding, the message becomes 9 0 12 15 22 5 0 21 8 21 18 1 0 0 0, or “I love Uhura.” Alice is heartbroken.

**Your Turn 31.** In decoding the message from the samples $\{(t_j, y_j)\}_{j=0}^{99}$, we never considered the sums

$$b_k = \frac{2}{100} \sum_{j=1}^{100} y_j \sin(kt_j).$$

Why not? Ideally, what should be the value of these sums?

**Your Turn 32.** This scheme of sending messages works well in the absence of unwanted background signals, or noise. Beyond receiving a message which seems to make no sense, explain two ways that you could test the sample data for noise.
7.4. Verifying the orthogonality relations: The triumph of DeMoivre’s formula. In this section we verify the orthogonality relations in Theorem 12. Along the way, we see how trigonometric identities and polar representations of complex numbers are crucial for our verification. We begin with:

**Lemma 13.** Let \( t_1, \ldots, t_m \) be as in Equation 13, and suppose that \( r \) is an integer with \( r \) not a multiple of \( m \). Then

\[
\sum_{j=1}^{m} \cos rt_j = \sum_{j=1}^{m} \sin rt_j = 0.
\]

**Proof.** For a real number \( \gamma \), we define the complex number

\[
e^{i\gamma} := \cos \gamma + i \sin \gamma,
\]

and we observe that to prove the first chain of equalities, it suffices to show that

\[
(18) \quad \sum_{j=1}^{m} e^{irt_j} = 0.
\]

By applying Theorem 10 and DeMoivre’s Formula (9), and computing a partial sum of a geometric series, we find

\[
\begin{align*}
\sum_{j=1}^{m} e^{irt_j} &= \sum_{j=1}^{m} e^{i(r \pi + 2j \pi)} \\
&= e^{-ir \pi} \sum_{j=1}^{m} e^{i \pi \frac{2j}{m}} \\
&= e^{-ir \pi} \sum_{j=1}^{m} \left(e^{i \pi \frac{2j}{m}}\right)^j \\
&= e^{-ir \pi} \frac{e^{i \frac{2\pi r}{m}} \left(1 - e^{i \frac{2\pi r}{m}}\right)^m}{1 - e^{i \frac{2\pi r}{m}}} \\
&= e^{-ir \pi} \frac{e^{i \frac{2\pi r}{m}} \left(1 - e^{i 2\pi r}\right)}{1 - e^{i \frac{2\pi r}{m}}} \\
&= 0.
\end{align*}
\]

We can now indicate the proof of the orthogonality relations (Theorem 12):

**Proof.** We only prove the first of the orthogonality relations. The others are proved similarly (see Exercise 7).
To begin, observe (via the sum formula\(^\text{26}\)) that
\[
\sin \theta \cos \phi = \frac{1}{2} [\sin(\theta - \phi) + \sin(\theta + \phi)].
\]

Therefore, Lemma 13 tells us that
\[
\sum_{j=1}^{m} \sin k t_j \cos l t_j = \frac{1}{2} \left[ \sum_{j=1}^{m} \sin(l - k) t_j + \sum_{j=1}^{m} \sin(l + k) t_j \right] = \frac{1}{2} [0 + 0] = 0.
\]

\[\square\]

**Your Turn 33.** *Why it is reasonable to use\( e^{i\gamma} \) as notation for\( \cos \gamma + i \sin \gamma \)? (Hint: Use Theorem 10 to compute\( e^{i\gamma_1} e^{i\gamma_2} \). What do you observe?)*

### 7.5. Exercises.

1. Verify that the average value of \( p(t) = a \cos t + b \sin t \) is zero on \([-\pi, \pi]\). For \( c \in \mathbb{R} \), how could you adjust \( p \) so that its average value is \( c \) on \([-\pi, \pi]\)?

2. Let \( n \) be a positive integer. What is the frequency (number of complete cycles over \([-\pi, \pi]\)) of \( \sin(nt) \)? Of \( \cos(nt) \)?

3. Use Equation (17) to express \( a_0 \) as a sum. Why do you suppose that the constant term of the best fit polynomial is \( a_0/2 \) instead of just \( a_0 \)? (Hint: Your Turn 26 and surrounding remarks may be helpful.)

4. Your instructor will provide you with a Geometer’s Sketchpad graph of a trigonometric polynomial of degree one. Use Sketchpad to find parameters \( t_1, t_2, t_3 \) so that the graph of \( t_1 + t_2 \cos x + t_3 \sin x \) matches the graph of the original polynomial as closely as possible. (Repeat for a trigonometric polynomial of degree 2.)

5. Consider Figure 42. Where do the tallest spikes in the graph of \( p \) occur? Do you suspect that tall spikes will occur at this location for most messages sent in this way, or is this coincidental? Discuss.

6. Using their mutually agreed upon coding scheme, Alice has been sending Kirk a barrage of insulting three-letter words. Spock, who usually helps Kirk decode these messages, is in bed with the flu. Can you help Kirk decode three messages based on the samples (at 10 evenly spaced times in \([-\pi, \pi]\)) given below?

---

\(^{26}\)See Exercise 12.
7. Prove the remaining two orthogonality relations. That is, given the hypotheses of Theorem 12, prove that

\[ \sum_{j=1}^{m} \sin(kt_j) \sin(lt_j) = \begin{cases} \frac{m}{2} \delta_{kl} & \text{if } k + l > 0 \\ 0 & \text{if } k = l = 0 \end{cases}, \quad \text{and} \quad \sum_{j=1}^{m} \cos(kt_j) \cos(lt_j) = \frac{m}{2} \delta_{kl}. \]

8. Consider the data \{(t_j, y_j)\} given in Figure 43. Assuming that this data models an even function on \([-\pi, \pi]\), deduce a new formula for recovering the coefficient \(a_k\) that requires approximately half as many calculations as the original formula

\[ a_k = \frac{2}{100} \sum_{j=1}^{100} y_j \cos(kt_j). \]

9. Consider the proof of Lemma 13

(a) Where is Theorem 10 applied?

(b) Where is DeMoivre’s formula (9) applied?

(c) Why is it important that \(2\pi\) not be a multiple of \(m\)?

(d) Why is it important that the \(t_j\)’s be evenly distributed in the interval \([-\pi, \pi]\)?

8. Exercises Involving Student Work

This section contains student solutions to various problems involving lines in the plane. We recommend that these solutions be analyzed prior to certain sections in the chapter (indicated below). For each problem and corresponding set of solutions, complete the following tasks:

1. Take 5 minutes in an attempt to solve the problem (without appealing to resources such as your text, etc.).
2. Discuss each student’s solution strategy. Indicate ways in which the student was “on-track”. Also, identify any incorrect assumptions, misconceptions, and errors revealed in each student’s work.

3. Try to determine whether the student’s numerical answer is correct.

4. Rate each solution on a scale from 1 to 5 with 5 as the highest rating and 1 as the lowest.

5. Compare your ratings in part 4 with another student (or group), and discuss your criteria for giving these ratings. Attempt to reach consensus with your partner (or group).

6. Carefully construct at least two solutions to the problem which employ different strategies. (You may use your text as a resource if you wish.)

10. (To be considered prior to Section 3) Consider a triangle with side lengths 3, 10, and \( C \). Suppose the angle opposite the side of length \( C \) is 60°. Find \( C \).
CHAPTER 5

Hyperbolic Trigonometry

Most of us first encounter the *hyperbolic trigonometric functions* in a calculus course. We see the definitions

\[
\cosh t = \frac{e^t + e^{-t}}{2} \\
\sinh t = \frac{e^t - e^{-t}}{2}
\]

from which it follows that \( \cosh' t = \sinh t \) and \( \sinh' t = \cosh t \). This enlarges the universe of functions of which we can compute derivatives, giving more opportunities to practice the sum, product, quotient, and chain rules.

Such a brief and narrow glimpse does not do justice to hyperbolic trigonometry. If we look more deeply into our calculus books, we may find an interesting application: we learn that the graph of the hyperbolic cosine function is a typical example of a *catenary*, the “hanging cable” shape that results when a cord is suspended between two posts.\(^1\) The fact that the shape of a catenary is related to \( \cosh t \) goes back to 1691, when Leibniz, Huygens, and Johann Bernoulli independently discovered this, responding to a challenge posed by Jakob Bernoulli.\(^2\) Hyperbolic trigonometric functions remain important in science and engineering; in fact, they are a basic and essential tool in relating *space* and *time* in the theory of relativity.

Defining the hyperbolic trig functions in terms of exponential functions is efficient, but we take a different approach in this chapter: we begin (plausibly enough) with *hyperbolas*. Indeed, the goal of this chapter is to play out the analogy between hyperbolic trigonometry and (the much more familiar) circular trigonometry. We will see many familiar ideas (angles, angle measure, rotations, trigonometric functions, identities, limits and derivatives) developed in the hyperbolic context. This approach reveals remarkable parallels between

\(^1\) Also, upside-down catenaries give the shape of stable arches, as discovered empirically in ancient times, and proved by Robert Hooke in 1671.

\(^2\) Galileo believed (incorrectly) that the shape of a hanging cable was given by a parabola. This was shown to be false by Jungius in 1669.
the two theories, and gives us another opportunity to reinforce the key ideas of circular trigonometry, since we will need to grapple with the same ideas in the less familiar hyperbolic setting.\footnote{Even more remarkable connections will become apparent in Chapter 11.}

1. Hyperbolas and Ellipses

It would be odd to begin a chapter on hyperbolic trigonometry without illuminating the term *hyperbolic*. In this section, we will recall some basic notions about the plane curves known as *hyperbolas*. We also discuss their cousins, the curves called *ellipses*. Both are defined in geometric terms.\footnote{Parabolas, like ellipses and hyperbolas, are defined geometrically. See Chapter 3.}

1.1. Ellipses. Suppose that \((x_0, y_0)\) and \((x_1, y_1)\) are points in the plane.\footnote{These two points are called the foci (plural of focus) of the ellipse. We leave open the possibility that they are actually the same point.} We let \(r\) be a real number such that \(2r\) is greater than the distance between \((x_0, y_0)\) and \((x_1, y_1)\). The ellipse defined by the data \\(\{(x_0, y_0), (x_1, y_1), r\}\) consists of the collection of points \((x, y)\) in the plane such that the sum of the distance from \((x, y)\) to \((x_0, y_0)\) and the distance from \((x, y)\) to \((x_1, y_1)\) equals \(2r\).

To imagine the shape of the ellipse, we picture stakes driven into the ground at the two points \((x_0, y_0)\) and \((x_1, y_1)\), with the ends of a string of length \(2r\) tied to the two stakes. (See Figure 1.) When the string is pulled taut, one obtains a point \((x, y)\) on the ellipse (why?). If we imagine pulling the string taut with a finger, and then moving our finger, the “oval” shape of an ellipse emerges.

**Your Turn 1.**

(a) Why did we assume that \(2r\) is greater than the distance between \((x_0, y_0)\) and \((x_1, y_1)\)?

(b) Describe the “ellipse” that would result if \(2r\) were equal to the distance between \((x_0, y_0)\) and \((x_1, y_1)\).

**Your Turn 2.** Give a qualitative description of the shape of the ellipse if \(2r\) is only slightly more than the distance between \((x_0, y_0)\) and \((x_1, y_1)\). Conversely, what happens to the ellipse if \(2r\) is very large relative to the distance between \((x_0, y_0)\) and \((x_1, y_1)\)?

In Exercise 8, you are asked to find the equation of the ellipse defined by the data \\(\{(x_0, y_0), (x_1, y_1), r\}\). At present, we will treat a much easier case, where the foci are the points \((\pm f, 0)\) (here \(f\) is a real number
with $f \geq 0$). Our condition on $r$ becomes $2r > 2f$ (why?), or simply $r > f$. Let’s consider the ellipse defined by this data. (See Figure 2.) We see that a point $(x, y)$ lies on the ellipse defined by this data if and only if

$$
\sqrt{(x + f)^2 + y^2} + \sqrt{(x - f)^2 + y^2} = 2r
$$

(1)
or equally well
\[
\sqrt{(x + f)^2 + y^2} = 2r - \sqrt{(x - f)^2 + y^2}.
\]

If we square Equation 2, we obtain
\[
(x + f)^2 + y^2 = 4r^2 - 4r\sqrt{(x - f)^2 + y^2} + (x - f)^2 + y^2.
\]

Simplification of Equation 3 yields
\[
r\sqrt{(x - f)^2 + y^2} = r^2 - fx.
\]

Squaring Equation 4, followed by simplification, produces the equation
\[
\frac{x^2}{r^2} + \frac{y^2}{r^2 - f^2} = 1.
\]

**Your Turn 3.** We are taught to be cautious in squaring an equation, since the new equation can have solutions that were not solutions to the original equation. Carefully explain why the transition from Equation 2 to Equation 3 does not introduce extraneous solutions. (Show that an extraneous solution would be a point \((x, y)\) such that \(\sqrt{(x + f)^2 + y^2} = -2r + \sqrt{(x - f)^2 + y^2}\). Then, give a geometric argument to show that this equation has no solutions; use the triangle inequality.)

**Your Turn 4.** We needed to square Equation 4. Again, carefully justify why squaring Equation 4 will not introduce extraneous solutions.

**Your Turn 5.**
(a) In Equation 5, why are the coefficients of \(x^2\) and \(y^2\) positive?
(b) How are the \(x\)- and \(y\)-intercepts of the ellipse related to \(f\) and \(r\)?
(c) Equations 1 and 5 both are “equations of the ellipse with data \(\{(-f, 0), (f, 0), r\}\).” For what reasons might one prefer either equation?
(d) What is the equation of the ellipse with foci at \((0, \pm f)\), and the same \(r\) as above?

1.2. Hyperbolas. The definition of a hyperbola closely resembles the definition of an ellipse, with a subtle change of sign.\(^6\)

\(^6\)This change of sign lies at the heart of the differences between circular and hyperbolic trigonometry.
Suppose that \((x_0, y_0)\) and \((x_1, y_1)\) are two different points in the plane. Let \(r\) be a positive real number such that \(2r\) is less than the distance between \((x_0, y_0)\) and \((x_1, y_1)\). The hyperbola defined by the data \(\{(x_0, y_0), (x_1, y_1), r\}\) consists of the collection of points \((x, y)\) in the plane such that the difference of the distance from \((x, y)\) to \((x_0, y_0)\) and the distance from \((x, y)\) to \((x_1, y_1)\) equals \(\pm 2r\).

What do hyperbolas look like? In Figure 3, we illustrate a hyperbola where the distance between the foci is 6 and where \(r = 2\). It consists of two branches that are captured between two asymptotes, which are lines that intersect at the midpoint of the segment joining the foci.

To get an idea of why Figure 3 is correct, try Your Turn 6. You can use either Geometer’s Sketchpad or ruler and compass.

**Your Turn 6.** In the coordinate plane, mark the foci at \((-3, 0)\) and \((3, 0)\), and set \(r = 2\). We obtain points on one branch of the hyperbola by marking points \((x, y)\) where the distance from \((x, y)\) to \((-3, 0)\) exceeds the distance from \((x, y)\) to \((3, 0)\) by four. Hence, we can find points on the hyperbola satisfying the pair of conditions in any row of the following table:

<table>
<thead>
<tr>
<th>distance from ((x, y)) to ((-3, 0))</th>
<th>distance from ((x, y)) to ((3, 0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>5.5</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>4.5</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Geometrically, this means we can produce points on the hyperbola by taking the intersection of two circles.

(a) Draw circles around \((-3, 0)\) and \((3, 0)\) of the radii indicated by the table, and mark the points that lie on the hyperbola.

(b) From your work, what are the \(x\)-intercepts of the hyperbola?

\(^7\)Again, these are the foci.
The definition of a hyperbola includes requirements about the foci and the magnitude of $r$. Your Turns 7 and 8 will clarify these choices.

**Your Turn 7.**

(a) Suppose that we allowed $(x_0, y_0)$ and $(x_1, y_1)$ to be the same point. What would the “hyperbola” consist of? (Consider the case where $r$ is positive and the case where $r = 0$.)

(b) Suppose that $(x_0, y_0) \neq (x_1, y_1)$ but $r = 0$. In this case, what does the “hyperbola” consist of?

**Your Turn 8.**

(a) Show that if $2r$ were greater than the distance between $(x_0, y_0)$ and $(x_1, y_1)$, then the “hyperbola” would be empty.

(b) If $2r$ equals the distance between $(x_0, y_0)$ and $(x_1, y_1)$, then what does the “hyperbola” consist of?

Finally, we consider the equation of a hyperbola in the special case that the foci are the points $(\pm f, 0)$ for some $f > 0$. The derivation is very similar to the case of an ellipse: from the definition, we see that the hyperbola defined by the data $\{(-f, 0), (f, 0), r\}$ has equation

$$\sqrt{(x + f)^2 + y^2} - \sqrt{(x - f)^2 + y^2} = \pm 2r,$$
which is algebraically equivalent to

\[(7) \quad \frac{x^2}{r^2} - \frac{y^2}{f^2 - r^2} = 1.\]

Note that from the equation of the hyperbola, it’s easy to find the equations of the asymptotes: we can transform Equation 7 to

\[(8) \quad \frac{y^2}{x^2} = \frac{f^2 - r^2}{r^2} - \frac{f^2 - r^2}{x^2},\]

and in the limit as \(|x| \to \infty\), we see that \(\frac{y}{x}\) approaches \(\pm \frac{\sqrt{f^2 - r^2}}{r}\). Therefore the asymptotes are the lines through \((0, 0)\) with slope \(\pm \frac{\sqrt{f^2 - r^2}}{r}\).

**Your Turn 9.**

(a) In Equation 7, is the coefficient of \(x^2\) positive or negative? Is the coefficient of \(y^2\) positive or negative? Explain.

(b) Does the hyperbola described by Equation 7 have \(x\)-intercepts? Does it have \(y\)-intercepts? How are these related to \(r\) and \(f\)?

(c) What is the equation of the hyperbola with foci at \((0, \pm f)\), and the same \(r\) as above?

### 1.3. Special ellipses and hyperbolas.

When we discuss (circular) trigonometry, we don’t use arbitrary ellipses—we use circles. In the same way, in our study of hyperbolic trigonometry, we won’t use arbitrary hyperbolas. Our attention will be focused on right hyperbolas, which we define to be hyperbolas whose asymptotes meet at right angles.

When does this happen? In the specific case of hyperbolas defined by the data \(\{(−f, 0), (f, 0), r\}\), we saw that the slopes of the asymptotes were \(\pm \frac{\sqrt{f^2 - r^2}}{r}\). The asymptotes meet at right angles when these slopes are \(\pm 1\), that is, when \(f^2 - r^2 = r^2\), which is to say that \(f = r\sqrt{2}\). A particular important case—an analogue of the “unit circle centered at the origin” that plays such a useful role in trigonometry—occurs when \(r = 1\) and \(f = \sqrt{2}\); by Equation 7, the equation of this hyperbola is \(x^2 - y^2 = 1\). (See Figure 4.)

---

Exercise 1 confirms that every circle is, an ellipse, albeit a very special one.
1.4. Exercises.

1.

(a) Using the definition of an ellipse, explain why the ellipse defined by the data \(\{(x_0, y_0), (x_0, y_0), r\}\) is a circle.

(b) What is its center? What is its radius? What simple equation has this circle as its solution set?

2. Consider the equation \(\frac{x^2}{9} + \frac{y^2}{12} = 1\).

(a) Why does the solution set form an ellipse?

(b) Find the \(x\)- and \(y\)-intercepts of the ellipse.

(c) Find the foci of the ellipse. (Do they lie on the \(x\)-axis or the \(y\)-axis?)

(d) If \((x, y)\) is a point on the ellipse, what number is the sum of its distances to the foci? (How many strategies can you find to answer this question?)

(e) For this example, what is the number \(r\)?

(f) Draw a rough graph of the ellipse.

(g) How is the ellipse given by \(\frac{x^2}{9} + \frac{y^2}{12} = 1\) related to the ellipse given by \(\frac{x^2}{12} + \frac{y^2}{9} = 1\)? To the ellipse given by \(\frac{(x-2)^2}{9} + \frac{(y+7)^2}{12} = 1\)?
(h) Find the equation of another ellipse that has the same foci as the ellipse with equation \( \frac{x^2}{9} + \frac{y^2}{12} = 1 \).

(i) Find the equation of another ellipse that has the same value of \( r \) as the ellipse with equation \( \frac{x^2}{9} + \frac{y^2}{12} = 1 \) (but has different foci).

3. Consider the equation \( \frac{x^2}{9} - \frac{y^2}{12} = 1 \).

(a) Why does the solution set form a hyperbola?

(b) Find any \( x \)- or \( y \)-intercepts of the hyperbola.

(c) Find the foci of the hyperbola.

(d) If \((x, y)\) is a point on the hyperbola, what number is the absolute value of the difference of its distances to the foci?

(e) For this example, what is the number \( r \)?

(f) Draw a rough graph of the hyperbola.

(g) Find the equations of all hyperbolas that have the same asymptotes as this one.

(h) Find the equation of another hyperbola that has the same foci as the given one.

(i) Find the equation of another hyperbola that has the same value of \( r \) as the given one (but has different foci).

4. Here we study the hyperbolas given by Equation 7.

(a) What happens to the asymptotes if \( r \) approaches 0? Sketch a typical hyperbola for which \( r \) is very small.

(b) What happens to the asymptotes if \( r \) approaches \( f \)? Sketch a typical hyperbola for which \( f \) is almost as large as \( f \).

5. Ellipses and hyperbolas as level sets.

(a) Consider the function \( F : \mathbb{R}^2 \to \mathbb{R} \), given by the rule \( F(x, y) = \frac{x^2}{9} + \frac{y^2}{12} \).

(i) What is the range of \( F \)?

(ii) Plot the level sets \( F^{-1}([c]) \) for \( c = 3, 2, 1, 0, -1, -2, -3 \).

(iii) For a number \( c \), describe the inverse image \( F^{-1}([c]) \).
(b) Consider the function \( G : \mathbb{R}^2 \to \mathbb{R} \), given by the rule \( G(x, y) = x^2 - y^2 \).

(i) What is the range of \( G \)?

(ii) Plot the level sets \( G^{-1}([c]) \) for \( c = 3, 2, 1, 0, -1, -2, -3 \).

(iii) For a number \( c \), describe the inverse image \( G^{-1}([c]) \).

6. Use graphing technology to graph the hyperbolas \( \frac{x^2}{4} - y^2 = c \) for various choices of \( c \). What happens when \( c \) is positive, negative, or zero, and as \( c \) goes to \( \pm \infty \)? What happens to the asymptotes and intercepts for various choices of \( c \)?

7. Use graphing technology to graph the hyperbolas \( \frac{x^2}{c^2} - y^2 = 1 \) for various (positive) choices of \( c \). What happens as \( c \) approaches \( \infty \) and as \( c \) approaches 0? What happens to the asymptotes and intercepts for various choices of \( c \)?

8. Find the equation of the ellipse defined by the data \( \{(x_0, y_0), (x_1, y_1), r\} \). Express your answer in the form \( p(x, y) = 1 \), where \( p(x, y) \) is a degree-two polynomial in \( x \) and \( y \).

9. Likewise, find the equation of the hyperbola defined by the data \( \{(x_0, y_0), (x_1, y_1), r\} \).

10. Explain why the curve with the equation \( xy = 1 \) is a hyperbola. Be certain to identify the foci and \( r \).

11. Consider a hyperbola defined by the data \( \{(x_0, y_0), (x_1, y_1), r\} \). Under what conditions on these data will the hyperbola be a right hyperbola? Justify your answer.

12. Give an example of an equation in \( x \) and \( y \) whose solution set is oval-shaped, but not an ellipse.

2. Hyperbolic Angles and Their Measures

In Chapter 4, we explored various ways in which the word angle is used. Sometimes ‘angle’ is the word to describe the meeting of two rays, as in the first part of Figure 5. This picture is ambiguous, however, since when it comes to measuring such an ‘angle’ one has to specify whether one is interested in the larger or smaller angle that is determined by the rays. It is somewhat more successful to describe an angle by a circular arc, as in the second part of Figure 5, since the ambiguity about “which angle” is removed.
Additionally, one often wants to consider angles that are oriented, meaning that one endpoint of the circular arc is identified as the initial point and the other endpoint is identified as the terminal point. Orientation can be indicated in a diagram by arrows (see the third part of Figure 5). The idea of orientation is linked to angle measure, since angles with counterclockwise orientation are decreed to have positive angle measure, with clockwise orientation corresponding to negative angle measure.

Indeed, one can regard an angle to be a process of rotation, as we did in Chapter 4. This has the advantage that angles with measure more than a “full rotation” of $2\pi$ can be discussed; they are inconvenient to represent via circular arcs.

In this section, our discussion of a hyperbolic angle parrots the representation of a (circular) angle in terms of a circular arc. We explore the notion of the hyperbolic measure of a hyperbolic angle, seeing the similarities and differences with radian measure of a circular angle. The notion of a hyperbolic angle as a hyperbolic rotation will be delayed until Section 4.

2.1. Hyperbolic angles. Just as we could declare a “(non-oriented) circular angle” to be an arc along a circle, we can define a (non-oriented) hyperbolic angle to be an arc along a right\footnote{Recall that a right hyperbola is a hyperbola whose asymptotes meet at right angles.} hyperbola. An angle
becomes *oriented* once we specify one of the endpoints as the initial point and the other as the terminal point. (See Figure 6.) Of course, different arcs may represent the same angle.\(^\text{10}\) (See Figure 7.)

---

\(^{10}\)Thus, a hyperbolic angle should be regarded as an *equivalence class* of hyperbolic arcs, with the equivalence relation suggested by Figure 7.
We note the first difference between the circular and hyperbolic trigonometries. On a circle, one can return to the same point by repeatedly moving around the circle; hence there exist infinitely many circular angles that possess the same pair of initial and terminal points on the unit circle. From the shape of a hyperbola, we see that no analogue of this phenomenon exists for hyperbolic angles. An initial point and a terminal point completely determine an arc on a hyperbola. There is no ambiguity about “whether to move clockwise or counterclockwise” or “how many times to loop around,” as in circular trigonometry.

2.2. Measure of hyperbolic angles. Any circle can be used to define a circular angle, and any right hyperbola can be used to define a hyperbolic angle. However, to make things more concrete, we focus our attention on the circle with equation $x^2 + y^2 = 1$, and the right hyperbola with equation $x^2 - y^2 = 1$. (See Figure 8.) We will learn how to measure angles defined by these objects.

![Figure 8](image_url)

Let’s review the circular case. A non-oriented circular angle may be regarded as an arc on the unit circle.\(^{11}\) Its radian measure is simply the length of the arc, since we are considering a circle of radius 1. But one further point needs be made about circular trigonometry, namely, that there is an equally good way to define radian measure, via area. Recall that the radian measure of the entire circle is $2\pi$, and that the radian

\(^{11}\)This definition of an angle—and the notion of its measure—must be augmented to encompass “arcs” that include complete rotations of the circle.
measure of any arc is therefore \(2\pi\) times the fraction of the circle represented by the arc. This allows us to define the radian measure \(\theta\) of an angle in terms of the area \(A\) of the circular sector associated to the arc:

\[
\theta = 2\pi \frac{\text{area of sector}}{\text{area of circle}} = 2\pi \frac{A}{\pi} = 2A.
\]

(See Figure 9.)

**Your Turn 10.** Consider an arc equal to one quarter of the unit circle.

(a) What is the radian measure \(\theta\) of the arc?

(b) What is the length of the arc?

(c) What is the area \(A\) of the circular sector determined by the arc?

(d) Verify that \(\theta = 2A\).

By analogy, a nonoriented hyperbolic angle\(^{12}\) is an arc on the right-hand branch of the hyperbola \(x^2 - y^2 = 1\). We now have a choice to make in defining the hyperbolic measure of this hyperbolic angle. One reasonable choice would be, “the length of the intercepted arc,” and a second reasonable choice would be, “twice the area of the sector determined by the arc.” (See Figure 9.)

---

\(^{12}\)Note that we are restricting attention to \(x^2 - y^2 = 1\). It’s not difficult to define the measure of angles on other right hyperbolas, but we won’t pursue it.
In the hyperbolic setting, these two notions do not give the same definition of angle measure! In fact, we can compute arc length and sector area for a variety of hyperbolic arcs with endpoints \((1, 0)\) and \((x, y)\). From our table, we see that the arc length is not twice the sector area. Indeed, the ratio of arc length to sector area is not constant:

<table>
<thead>
<tr>
<th>((x, y))</th>
<th>arc length</th>
<th>sector area</th>
</tr>
</thead>
<tbody>
<tr>
<td>((2, \sqrt{3}))</td>
<td>2.03762</td>
<td>0.658479</td>
</tr>
<tr>
<td>((4, \sqrt{15}))</td>
<td>4.96773</td>
<td>1.03172</td>
</tr>
<tr>
<td>((6, \sqrt{35}))</td>
<td>7.8268</td>
<td>1.23894</td>
</tr>
<tr>
<td>((8, \sqrt{63}))</td>
<td>10.6702</td>
<td>1.38433</td>
</tr>
</tbody>
</table>

Which measurement should we choose to define hyperbolic angle measure? From the circular case, there is no reason to prefer the ‘arc length’ or the ‘sector area’ definition of angle measure—after all, arc length is always just twice the sector area! And of course, both are preserved by rotation: if we take an arc of the circle \(x^2 + y^2 = 1\) and rotate around the center by any amount, the new arc has the same length, and the same sector area, as the original. In the hyperbolic case, however, there is an important mathematical difference. As we will see in Section 4, Exercise 6, arc length is not preserved by the hyperbolic analogue of a rotation, but sector area is preserved (Lemma 5 of the same section).\(^{14}\) This justifies our decision: the hyperbolic measure of an angle defined by an arc on the hyperbola with equation \(x^2 - y^2 = 1\), is defined to be twice the area of the hyperbolic sector determined by the arc.

2.3. Producing a standard angle from a number. Given a real number \(\theta\), we can produce an oriented angle whose measure is \(\theta\). This idea works in both circular and hyperbolic trigonometry.

Let’s begin with the circular case. Given a real number \(\theta\), we can produce an oriented circular angle as follows, using the circle with equation \(x^2 + y^2 = 1\). If \(\theta\) is positive, we begin at the point \((1, 0)\), and produce

\(\text{Recall from calculus that both arc length and area can be computed from integrals. We have used a computer algebra system to give numerical approximations of these integrals.}
\(\text{Moreover, in Exercise 6 in Section 5, we will show that a modified notion of arc length is equal to twice the sector area.}\)
a circular arc by traveling counterclockwise until an arc of length $\theta$ is produced (or equally well, a sector of area of $\theta/2$ is swept out). If $\theta$ is negative, we begin at the point $(1,0)$, but produce an arc by traveling clockwise, until a sector of area $|\theta/2|$ is swept out. If $\theta = 0$, the corresponding angle consists of the single point $(1,0)$.

We say that the angle produced by this process is the \textit{standard angle coming from} $\theta$, and that such an angle is in \textit{standard position}, meaning that the initial point is $(1,0)$.

\textit{Standard hyperbolic angles} are produced in an analogous way, using the right-hand branch of the hyperbola with equation $x^2 - y^2 = 1$. Given a real number $\theta$, we produce an oriented arc on the hyperbola, with initial point $(1,0)$. The terminal point is in the quadrant with $x > 0$ and $y > 0$ if $\theta$ is positive, and in the quadrant with $x > 0$ and $y < 0$ if $\theta$ is negative. The terminal point is chosen so that the sector that is swept out has area $|\theta/2|$. (See Figure 10.)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure10.png}
\caption{Two standard hyperbolic angles}
\end{figure}

There is one subtlety that must be addressed. Using improper integrals, one can show that the region above the positive $x$-axis, to the left of the hyperbola, and below the asymptote $y = x$ has infinite area. (For evidence, see Exercise 3.) This means that given any real number $\theta$, no matter how large, we can indeed produce an angle whose hyperbolic measure is $\theta$. 
We again observe the non-periodic behavior of the hyperbola. Recall that in circular trigonometry, if \( \theta_1 \) and \( \theta_2 \) are two real numbers whose difference is a multiple of \( 2\pi \), their standard angles have the same terminal point on the unit circle. In hyperbolic trigonometry, this phenomenon does not occur. The quantity \( 2\pi \) plays no role. As the hyperbolic angle measure \( \theta \) goes to \( \infty \), the terminal point of the standard angle moves up and to the right along the hyperbola, farther and farther away from the origin. As \( \theta \to -\infty \), the point \( Q \) moves down and to the right along the hyperbola, farther and farther from the origin.

2.4. Exercises.

1. Why have we not defined the degree measure of a hyperbolic angle? (Why don’t we simply divide the hyperbola into 360 equal pieces, as we did the circle?)

2. Suppose that beyond the hyperbola with equation \( x^2 - y^2 = 1 \), we consider a more general hyperbola with equation \( x^2 - y^2 = r^2 \). How must we modify our notions of hyperbolic arc and measure?

3. Using improper integrals, show that the region above the \( x \)-axis, below the hyperbola \( y = 1/x \), to the right of the \( y \)-axis, and to the left of \( x = 1 \) has infinite area.

3. The Hyperbolic Trigonometric Functions

3.1. Circular and hyperbolic trigonometric functions. Using the idea of standard angles from the last section, it’s easy to define both the circular and hyperbolic trigonometric functions. We will use Figure 11, which shows the essential constructions of circular trigonometry on the left, and hyperbolic trigonometry on the right.

The circular trigonometric functions can be defined using the circle with equation \( x^2 + y^2 = 1 \). Given any real number \( \theta \), we produce a standard angle coming from \( \theta \), as described in the last section (perhaps involving multiple trips around the circle). Regarding the angle as an arc of the circle, we let \((x, y)\) denote the terminal point of the arc. The six circular trigonometric functions of \( \theta \) are defined by the rules \( \sin \theta = y \), \( \cos \theta = x \), \( \tan \theta = y/x \), \( \csc \theta = 1/y \), \( \sec \theta = 1/x \), and \( \cot \theta = x/y \).

Hyperbolic trigonometry uses exactly the same ideas, except that the unit circle is replaced by the right-hand branch of the hyperbola with equation \( x^2 - y^2 = 1 \). Again, let \((x, y)\) denote the terminal point on the
Hyperbolic Trigonometry

The six hyperbolic trigonometric functions of \( \theta \) are defined using the rules:

\[
\sinh(\theta) = y, \quad \cosh(\theta) = x, \quad \tanh(\theta) = \frac{y}{x}, \quad \text{csch}(\theta) = \frac{1}{y}, \quad \text{sech}(\theta) = \frac{1}{x}, \quad \text{coth}(\theta) = \frac{x}{y}.
\]

Your Turn 11. On the basis of the definitions just made, give a reasonable, “natural” domain (inside the real numbers) for each of the six circular trigonometric functions. Do the same for the six hyperbolic trigonometric functions.

We emphasize that the circular trigonometric functions and the hyperbolic trigonometric functions are different. For example:

- To compute \( \cos(.75) \) and \( \sin(.75) \), one starts on the unit circle at \((1,0)\), and travels counterclockwise until the circular sector has area 0.375 (remember \( .75 = \theta = 2A \)). Let’s denote the point on the unit circle where we have stopped as \((x,y)\). Then \( \cos(.75) = x \) and \( \sin(.75) = y \). According to a scientific calculator, \( \cos(.75) \approx .7317 \) and \( \sin(.75) \approx .6816 \). Note that \( .7317^2 + .6816^2 = .99996345 \approx 1 \), reflecting the fact that \((x,y)\) is on the unit circle. (See the first half of Figure 12.)

- On the other hand, to compute \( \cosh(.75) \) and \( \sinh(.75) \), one starts at the point \((1,0)\) on the hyperbola \( x^2 - y^2 = 1 \), and travels up and to the right along the hyperbola, until the area of the sector is 0.375. If
the stopping point is \((x', y')\), then \(\cosh(0.75) = x'\) and \(\sinh(0.75) = y'\). According to a scientific calculator, \(\cosh(0.75) \approx 1.2947\) and \(\sinh(0.75) \approx 0.8223\). Note that \(1.2947^2 - 0.8223^2 = 1.0000708 \approx 1\), which is no surprise since \((x', y')\) is on the hyperbola \(x^2 - y^2 = 1\). (See the second half of Figure 12.)

**Figure 12.** \(\theta = 0.75\) in circular and hyperbolic trigonometries

This illustrates the basic identity for hyperbolic trig functions:

\[
(9) \quad \cosh^2 \phi - \sinh^2 \phi = 1 \text{ for every real number } \phi.
\]

Equation 9 follows from the fact that \((\cosh \phi, \sinh \phi)\) lies on the hyperbola \(x^2 - y^2 = 1\).

**Your Turn 12.**

(a) *To which identity from circular trigonometry is the identity* \(\cosh^2 \phi - \sinh^2 \phi = 1\) *analogous? Why?*

(b) *Find the identity in hyperbolic trigonometry that is analogous to the identity* \(1 + \tan^2 \phi = \sec^2 \phi\).

**3.2. Exercises.**

1. Is \(\cosh \phi\) an even function, an odd function, or neither? Is \(\sinh \phi\) an even function, an odd function, or neither? Explain. Compare to the analogous circular trigonometric functions.

2. Without a calculator, find \(\cosh 0\) and \(\sinh 0\). Explain briefly. Compare to the analogous circular trigonometric functions.
3. Using a calculator, make a table of values for \( \cosh \phi \) and \( \sinh \phi \) for \( \phi = 0, \pm 0.5, \pm 1, \pm 1.5, \pm 2, \pm 2.5, \) and \( \pm 3 \). Use these to give rough graphs of \( \cosh \phi \) and \( \sinh \phi \). Then, plot the ordered pairs \(( \cosh \phi, \sinh \phi )\) along the hyperbola \( x^2 - y^2 = 1 \).

4. What is the range of the function \( \cosh \phi \)? What is the range of the function \( \sinh \phi \)? Compare to the analogous circular trigonometric functions. *(Examine the right-hand branch of the hyperbola with equation \( x^2 - y^2 = 1 \), and the circle with equation \( x^2 + y^2 = 1 \). It is not necessary to have the graphs of the trigonometric functions.)*

5. Show that the range of the hyperbolic tangent function is the open interval \((-1, 1)\). *(Consider a point \(( \cosh \phi, \sinh \phi )\) on the hyperbola \( x^2 - y^2 = 1 \).) In comparison, what is the range of the (circular) tangent function?

6. From your answers to Exercises 4 and 5, deduce the range of the hyperbolic secant, cosecant, and cotangent functions. Compare to the analogous circular trigonometric functions.

7.

(a) If \( \cosh \phi = 2 \), what can \( \sinh \phi \) be?

(b) Is there any \( \phi \) with \( \cosh \phi = \frac{1}{2} \)?

(c) If \( \sinh \phi = 3 \), what can \( \cosh \phi \) be?

(d) If \( \sinh \phi = -\frac{1}{3} \), what can \( \cosh \phi \) be?

Explain briefly. *(You will need to use the identity \( \cosh^2 \phi - \sinh^2 \phi = 1 \), but make certain that your answers are compatible with your answer to Exercise 4 and with the picture of the right-hand branch of the hyperbola \( x^2 - y^2 = 1 \).)*

8. How do the results of the Exercise 7 compare to results for the analogous circular trigonometric functions?

9. In (circular) trigonometry, we learn that the sine and cosine functions are closely related to each other. In particular, their graphs are related to each other by a graphing transformation, thanks to the identity
sin(t) = cos(π/2 − t). Are the hyperbolic sine and cosine functions related to each other by a graphing transformation? Discuss.

10. Which points in the (x, y)-plane can be written in the form r(cosh φ, sinh φ) for some r > 0? Is the answer different if you allow r ≤ 0 also? How does this differ from the notion of polar coordinates (using the circular trigonometric functions)?

4. Hyperbolic Rotations

In Chapter 4, we argued that an angle could be thought of as a “dynamic process of rotation.” In the usual (circular) sense, this notion of rotation is understood viscerally by anyone who has every turned the steering wheel of a car or operated a faucet. In this section, we explore the hyperbolic analogue, which has a quite different geometric nature.

4.1. What should a “rotation” do? Circular rotations (with center at the origin) are easy to describe (see the first part of Figure 13). If we were to rotate a point $P$ by $φ$, we preserve $P$’s distance to the origin, but we increase by $φ$ the measure of the angle that $P$ makes with the positive real axis.

It’s helpful to write this in polar coordinates: if we write $P$ as $(r \cos θ, r \sin θ)$ and write $R_φ : \mathbb{R}^2 → \mathbb{R}^2$ for the transformation that rotates each point in the plane by $φ$, then we have

\[
R_φ(r \cos θ, r \sin θ) = (r \cos(φ + θ), r \sin(φ + θ)).
\]

Let’s try the same idea in the context of hyperbolic trigonometry. Here, we can’t immediately consider all points in the plane; we focus our attention instead on a region where the hyperbolic analogue of polar coordinates makes sense.\textsuperscript{17} This region is the wedge $W$ that lies above the line $y = −x$ and below the line $y = x$ (see the second part of Figure 13). In $W$, each point can be written uniquely as a positive scalar multiple of a point on the right-hand branch of the hyperbola with equation $x^2 − y^2 = 1$; in other words, we can write any $P ∈ W$ as $(r \cosh θ, r \sinh θ)$ for a unique $θ ∈ \mathbb{R}$ and unique $r > 0$. By analogy to the circular case, we should require a hyperbolic rotation $H_φ$ to preserve each hyperbola $x^2 − y^2 = r^2$, and to increase

\textsuperscript{15}Here we are assuming that $P$ is not the origin. If $P$ were the origin, it would not be moved by any rotation.
\textsuperscript{16}In other words, rotation preserves circles centered at the origin.
\textsuperscript{17}You may have explored this in Exercise 10.
the hyperbolic angle of $P$ by $\phi$. Symbolically, this requirement is

$$H_\phi(r \cosh \theta, r \sinh \theta) = (r \cosh(\phi + \theta), r \sinh(\phi + \theta))$$

The problem that remains for us is to find a useful, concrete way to describe $H_\phi$. Naturally, we will use the circular trigonometry for inspiration.

4.2. Rotations as linear transformations. In Chapter 4, we found that $R_\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ (rotation by $\phi$) could be described by a matrix. Using column vectors to represent points in $\mathbb{R}^2$, we found:

$$R_\phi \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$  

Two facts need to be underlined. First, since $R_\phi$ is given by a matrix, it’s automatic that $R_\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a linear transformation, as studied in a linear algebra course. Second, any proof of Equation 12 is tantamount to a proof of the sum formulas for the cosine and sine functions, as you can verify in Your Turn 13.

**Your Turn 13.** Write the point $\begin{pmatrix} x \\ y \end{pmatrix}$ in polar form as $\begin{pmatrix} r \cos \theta \\ r \sin \theta \end{pmatrix}$. Then, perform the matrix multiplication in Equation 12 to obtain the sum formulas for cosine and sine.
Entirely analogous ideas govern the notion of a hyperbolic transformation. Our main result is the following:

**Theorem 1.** A hyperbolic rotation \( H_\phi : W \to W \) extends to a linear transformation \( H_\phi : \mathbb{R}^2 \to \mathbb{R}^2 \) that is given by the matrix \[
\begin{pmatrix}
cosh \phi & \sinh \phi \\
\sinh \phi & \cosh \phi
\end{pmatrix}.
\]

Theorem 1 is beautiful, but not self-evident, and it will take some work to prove it. As a start, let’s note that Theorem 1 can be restated as the following concrete matrix equation:

**Theorem 2.** For any real numbers \( \phi \) and \( \theta \), we have the following matrix equation:

\[
\begin{pmatrix}
cosh \phi & \sinh \phi \\
\sinh \phi & \cosh \phi
\end{pmatrix}
\begin{pmatrix}
cosh \theta \\
\sinh \theta
\end{pmatrix} = \begin{pmatrix}
cosh(\phi + \theta) \\
\sinh(\phi + \theta)
\end{pmatrix}.
\]

**Your Turn 14.** Explain why Theorem 1 is a consequence of Theorem 2.

By expanding the matrix product in Theorem 2—exactly as you did in Your Turn 13—we can see that Theorem 2 implies the **sum formulas for hyperbolic sine and cosine**:

**Corollary 3.** For any real numbers \( \phi \) and \( \theta \),

\[
cosh(\phi + \theta) = \cosh \phi \cosh \theta + \sinh \phi \sinh \theta
\]

\[
\sinh(\phi + \theta) = \sinh \phi \cosh \theta + \cosh \phi \sinh \theta.
\]

**Your Turn 15.** How are these formulas similar/different from the sum formulas for sine and for cosine?

The task that remains is verification of Theorem 2.

**4.3. A proof of Theorem 2.** We begin by establishing a few lemmas.

**Lemma 4.** If \[
\begin{pmatrix}
x \\
y
\end{pmatrix}
\]
is on the right-hand branch of the hyperbola with equation \( x^2 - y^2 = 1 \), then so is \[
\begin{pmatrix}
cosh \phi & \sinh \phi \\
\sinh \phi & \cosh \phi
\end{pmatrix}
\begin{pmatrix}
x \\
y
\end{pmatrix}.
\]

**Proof.** First we compute the product:

\[
\begin{pmatrix}
cosh \phi & \sinh \phi \\
\sinh \phi & \cosh \phi
\end{pmatrix}
\begin{pmatrix}
x \\
y
\end{pmatrix} = \begin{pmatrix}
x \cosh \phi + y \sinh \phi \\
x \sinh \phi + y \cosh \phi
\end{pmatrix}.
\]
To check that this point is on the hyperbola, we compute

\[
(x \cosh \phi + y \sinh \phi)^2 - (x \sinh \phi + y \cosh \phi)^2
= (x^2 \cosh^2 \phi + 2xy \cosh \phi \sinh \phi + y^2 \sinh^2 \phi)
- (x^2 \sinh^2 \phi + 2xy \cosh \phi \sinh \phi + y^2 \cosh^2 \phi)
= x^2(\cosh^2 \phi - \sinh^2 \phi) + y^2(\sinh^2 \phi - \cosh^2 \phi)
= x^2 - y^2 = 1.
\]

Finally, to check that this vector is on the right-hand branch of the hyperbola, we look at its \(x\)-coordinate, which is \(x \cosh \phi + y \sinh \phi\). We must show that this is positive. Since \(x \cosh \phi\) is positive (why?), this would follow from the statement that \(|x \cosh \phi| > |y \sinh \phi|\), so we prove this statement:

\[
|x \cosh \phi| > |y \sinh \phi| \iff x^2 \cosh^2 \phi > y^2 \sinh^2 \phi
\]

\[
\iff x^2 \cosh^2 \phi > (x^2 - 1)(\cosh^2 \phi - 1)
\]

\[
\iff x^2 \cosh^2 \phi > x^2 \cosh^2 \phi - \cosh^2 \phi - x^2 + 1
\]

\[
\iff \cosh^2 \phi + x^2 > 1
\]

which is true since \(x \geq 1\) and \(\cosh \phi \geq 1\). □

**Lemma 5.** If a region in the plane has area \(A\), then its image under the matrix \(\begin{pmatrix} \cosh \phi & \sinh \phi \\ \sinh \phi & \cosh \phi \end{pmatrix}\) also has area \(A\).

**Proof.** It is a general fact that under a \(2 \times 2\)-matrix \(M\), the area of any region expands by a factor of \(|\det M|\), and here we have \(\det \begin{pmatrix} \cosh \phi & \sinh \phi \\ \sinh \phi & \cosh \phi \end{pmatrix} = \cosh^2 \phi - \sinh^2 \phi = 1\). □

**Your Turn 16.** To what statements in circular trigonometry are Lemmas 4 and 5 analogous?

We can now give a proof of Theorem 2:

**Proof.** We use Figure 14. Let \(M\) denote the matrix \(\begin{pmatrix} \cosh \phi & \sinh \phi \\ \sinh \phi & \cosh \phi \end{pmatrix}\). The matrix \(M\) takes \(O\) to itself, and \(M\) takes \(S = (1,0)\) to \(Q = (\cosh \phi, \sinh \phi)\). By Lemma 4, \(M\) carries \(P = (\cosh \theta, \sinh \theta)\) to some other point \(R\) on the right-hand branch of the hyperbola, which we denote as \((\cosh \alpha, \sinh \alpha)\) for some real
number $\alpha$. It is easy to see that $M$ takes the sector $OPS$ to the sector $ORQ$. By Lemma 5, these two sectors must have the same area.

Note that proving the theorem means showing that $\alpha = \phi + \theta$ (why?). We do this by comparing areas, remembering that the hyperbolic measure of an angle is twice the area of the corresponding sector:

$$\frac{\phi}{2} + \frac{\theta}{2} = \text{area(sector } OQS) + \text{area(sector } OPS)$$
$$= \text{area(sector } OQS) + \text{area(sector } ORQ)$$
$$= \text{area(sector } ORS)$$
$$= \frac{\alpha}{2}.$$ 

Hence $\phi + \theta = \alpha$. $\square$

4.4. Exercises.

*Identities that follow from the sum formulas*

1. Find expressions for $\sinh(2\alpha)$ and $\cosh(2\alpha)$ in terms of $\sinh \alpha$ and $\cosh \alpha$.

2. Using Exercise 1, find an expression for $\tanh(2\alpha)$ in terms of $\tanh \alpha$. 
3. Below is a list of identities involving circular trigonometric functions. Find and prove analogous identities involving hyperbolic trigonometric functions. Briefly explain how you found the correct hyperbolic versions.

(a) \( \tan(\theta + \phi) = \frac{\tan \theta + \tan \phi}{1 - \tan \theta \tan \phi} \)

(b) \( \sin \theta \sin \phi = \frac{1}{2} \cos(\theta - \phi) - \frac{1}{2} \cos(\theta + \phi) \)

(c) \( \sin \theta + \sin \phi = 2 \sin \left( \frac{\theta + \phi}{2} \right) \cos \left( \frac{\theta - \phi}{2} \right) \)

Properties of hyperbolic rotations

4. A hyperbolic rotation is very different from a circular rotation, but it is a simple type of linear transformation that usually is discussed in a linear algebra class.

Consider the matrix \( H_\phi = \begin{pmatrix} \cosh \phi & \sinh \phi \\ \sinh \phi & \cosh \phi \end{pmatrix} \).

(a) Show that \( \begin{pmatrix} 1 \\ 1 \end{pmatrix} \) is an eigenvector of \( H_\phi \). What is the corresponding eigenvalue?

(b) Show that \( \begin{pmatrix} -1 \\ 1 \end{pmatrix} \) is an eigenvector of \( H_\phi \). What is the corresponding eigenvalue?

(c) In contrast, what are the eigenvalues of the circular rotation \( R_\phi = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} \)?

5. Explain why hyperbolic rotations are given exactly by the matrices of the form \( \begin{pmatrix} a & b \\ b & a \end{pmatrix} \), where \( a^2 - b^2 = 1 \) and \( a \) is positive.

6. Let \( S \) denote the arc on the hyperbola with equation \( x^2 - y^2 = 1 \) whose endpoints are \((1, 0)\) and \((2, \sqrt{3})\). Let \( H \) be the hyperbolic rotation with matrix \( \begin{pmatrix} 2 & \sqrt{3} \\ \sqrt{3} & 2 \end{pmatrix} \). Let \( S' \) denote the hyperbolic arc obtained by applying \( H \) to \( S \).

(a) What are the endpoints of \( S' \)?

(b) Via a sketch, show that the length of \( S' \) does not equal the length of \( S \).

(This exercise shows that arc length is not preserved by hyperbolic rotations, in general. In contrast, sector area is preserved by hyperbolic rotations, as shown in Lemma 5. This helps explain why we used sector area, and not arc length, to define the hyperbolic measure of a hyperbolic angle.)

7. Let \( H \) be the hyperbolic rotation with matrix \( \begin{pmatrix} 2 & \sqrt{3} \\ \sqrt{3} & 2 \end{pmatrix} \).
(a) Find the image of each of the following points under $H$ and under its iterates $H^2$ and $H^3$.

(i) $(1,0)$
(ii) $(-1,0)$
(iii) $(0,1)$
(iv) $(0,-1)$
(v) $(1,1)$
(vi) $(-1,-1)$
(vii) $(1,-1)$
(viii) $(-1,1)$

(b) In words, or with a diagram, explain “where a point $(x,y)$ in the plane moves” under higher and higher powers of $H$.

Additional exercises

8. Explain why any oriented circular angle (an oriented arc on the unit circle) can be transformed to a standard circular angle by means of a circular rotation. Similarly, explain why any oriented hyperbolic angle (an oriented arc on the right-hand branch of the hyperbola with equation $x^2 - y^2 = 1$) can be transformed to a standard hyperbolic angle by means of a hyperbolic rotation.

9. Investigate the geometric properties of the linear transformations given by the matrices
   \[
   \begin{pmatrix}
   \cos \phi & \sin \phi \\
   \sin \phi & \cos \phi
   \end{pmatrix}
   \]
   and
   \[
   \begin{pmatrix}
   \cosh \phi & -\sinh \phi \\
   \sinh \phi & \cosh \phi
   \end{pmatrix}
   \]
   which did not arise as circular or hyperbolic rotations. (To begin, determine whether they have real eigenvalues.)

5. Differential Calculus and the Hyperbolic Trigonometric Functions

We have defined and explored hyperbolic trigonometry in geometric terms, but a deeper appreciation of the features of these functions will result when we bring the ideas of calculus into our study. A first step is the computation of the derivatives of the hyperbolic trigonometric functions.

5.1. The derivative of the hyperbolic sine and cosine functions. By definition,

\[
\sinh'(t) = \lim_{h \to 0} \frac{\sinh(t+h) - \sinh(t)}{h}.
\]
Using the sum formula for the hyperbolic sine function, we can write
\[
\frac{\sinh(t + h) - \sinh(t)}{h} = \frac{\sinh t \cosh h + \cosh t \sinh h - \sinh t}{h} = \sinh t \left( \frac{\cosh h - 1}{h} \right) + \cosh t \left( \frac{\sinh h}{h} \right).
\]
Thus we will be able to compute \( \sinh'(t) \) as soon as we can compute two limits: \( \lim_{h \to 0} \frac{\sinh h}{h} \) and \( \lim_{h \to 0} \frac{\cosh h - 1}{h} \).

**Your Turn 17.**

(a) Why can’t we find these limits simply by substituting \( h = 0 \)?

(b) Why can’t we use L’Hopital’s Rule to find these limits?

Since the hyperbolic sine and cosine functions are defined geometrically using standard hyperbolic angles, there is no alternative but to use these notions in our investigation of these limits. The results are remarkably similar to the circular case:

**Theorem 6.** \( \lim_{h \to 0} \frac{\sinh h}{h} = 1 \) and \( \lim_{h \to 0} \frac{\cosh h - 1}{h} = 0 \).

**Figure 15.** Proving \( \lim_{h \to 0^+} \frac{\sinh h}{h} = 1 \)

**Proof.** The limit calculation ultimately relies on a comparison of areas. We will use Figure 15, and take \( h > 0 \). Observe that \( P = (\cosh h, \sinh h) \), \( R = (1, 0) \), \( S = (\cosh h, 0) \), and \( Q = (1, \frac{\sinh h}{\cosh h}) \) (the \( x \)-coordinate
is 1, and the slope of $OQ$ equals the slope of $OP$, which is $\frac{\sinh h}{\cosh h}$). Then we can compute the areas of certain regions: $\text{area}(\triangle OQR) = \frac{1}{2} \sinh h \cosh h$, area(sector $OPR$) = $\frac{1}{2} h$ (recall that $h$ is the hyperbolic measure of the angle), and area($\triangle OPS$) = $\frac{1}{2} \sinh h \cosh h$. It is clear that $\text{area}(\triangle OQR) \leq \text{area(sectorOPR)} \leq \text{area($\triangle OPS$)}$, hence $\cosh h \sinh h \geq h \geq \frac{\sinh h}{\cosh h}$, so $\cosh h \geq \frac{h}{\sinh h} \geq \frac{1}{\cosh h}$. Since $\lim_{h \to 0^+} \cosh h = 1 = \lim_{h \to 0^+} \frac{1}{\cosh h}$, we conclude that $\lim_{h \to 0^+} \frac{h}{\sinh h} = 1$, and then taking reciprocals, we have $\lim_{h \to 0^+} \frac{\sinh h}{h} = 1$.

Your Turn 18 handles the limit through negative values of $h$. The proof that $\lim_{h \to 0} \cosh h - \frac{1}{h} = 0$ is left to Exercise 1.

**Your Turn 18.** Show that $\lim_{h \to 0} -\frac{\sinh h}{h} = 1$. (Use the fact that the hyperbolic sine function is an odd function.)

With the limits as $h$ approaches zero of $\frac{1-\cosh h}{h}$ and $\frac{\sinh h}{h}$ available to us, we can finish the computation from the beginning of this section:

**Theorem 7.** $\sinh'(t) = \cosh(t)$

**Your Turn 19.** Prove that $\cosh'(t) = \sinh(t)$.

### 5.2. Exercises.

1. Prove that $\lim_{h \to 0} \frac{\cosh h - 1}{h} = 0$. (Begin with $\lim_{h \to 0} \frac{\sinh h}{h} = 1$, then multiply numerator and denominator by $\cosh h + 1$.)

2. Draw graphs of $\cosh t$ and $\sinh t$ and explain why it is plausible that $\cosh' t = \sinh t$ and $\sinh' t = \cosh t$ (using the interpretation of derivative as slope).

3. Using the derivatives of $\sinh \theta$ and $\cosh \theta$, compute the derivatives of $\tanh \theta$, $\text{sech} \theta$, $\coth \theta$, and $\text{csch} \theta$. Compare to the derivatives of the analogous circular trigonometric functions.

4. Compute the higher derivatives of $\cosh t$ and $\sinh t$ at 0, and deduce the Maclaurin series of $\cosh t$ and $\sinh t$. How do these series compare with the Maclaurin series of $\cos t$ and $\sin t$?

5.
(a) Show that \( \sinh t \) and \( \cosh t \) are both solutions of the differential equation \( y''(t) = y(t) \).

(b) Find a similar differential equation of which \( \sin t \) and \( \cos t \) are solutions.

6. In this exercise, we note the modification to the usual notion of arc length that is necessary if we want to use a sort of “arc length” to measure hyperbolic angles.

In calculus, one shows that it is reasonable to define the length of a parameterized curve \((x(t), y(t))\) (with \(a \leq t \leq b\)) by the formula

\[
\text{length} = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} \, dt.
\]

(a) Use this to show that for the parameterized curve \((\cos t, \sin t)\) for \(0 \leq t \leq \theta\), the arc length is \(\theta\).

(This reinforces the fact that along the unit circle, the circular (radian) measure of an angle equals the length of the intercepted arc.)

In contrast, the hyperbolic measure of a hyperbolic angle is not equal to the length of the intercepted arc.

However, something similar is true; one must simply make a sign change:

(b) Let \(x(t) = \cosh t\) and \(y(t) = \sinh t\) for \(0 \leq t \leq \theta\). Prove that

\[
\theta = \int_0^\phi \sqrt{-\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} \, dt.
\]

6. Hyperbolic Trigonometric Functions and the Exponential Function

6.1. Calculus to the rescue. We return to the astonishing fact with which we began this chapter:

**Theorem 8.** For any \( \phi \), \( \cosh \phi = \frac{e^\phi + e^{-\phi}}{2} \) and \( \sinh \phi = \frac{e^\phi - e^{-\phi}}{2} \).

**Lemma 9.** If \( g(t) \) is a function that is equal to its own derivative, then \( g(t) = Ce^t \) for some constant \( C \).

**Proof.** By the quotient rule, \( \left(\frac{g(t)}{e^t}\right)' = \frac{e^t g'(t) - g(t)e^t}{e^{2t}} = 0 \), so \( g(t) = Ce^t \) is a constant. \( \square \)

Now we can prove Theorem 8:

**Proof.** We consider the function \( g(\phi) = \cosh \phi + \sinh \phi \). We have that \( g'(\phi) = \sinh \phi + \cosh \phi = g(\phi) \). By Lemma 9, the only function equal to its own derivative is \( Ce^\phi \), where \( C \) is a constant. Thus
cosh φ + sinh φ = Ceφ for some constant C. To find C, we evaluate at φ = 0, and discover that C = Ce0 =
cosh 0 + sinh 0 = 1 + 0 = 1, so C = 1. Hence cosh φ + sinh φ = eφ.

Substituting −φ for φ, we have that e−φ = cosh(−φ) + sinh(−φ) = cosh φ − sinh φ, using that cosh is an
even function and sinh is an odd function.

Finally, from eφ = cosh φ + sinh φ and e−φ = cosh φ − sinh φ, one can solve for cosh φ (or sinh φ) by
adding (or subtracting) the two equations. This yields the equations in the statement of the theorem. □

6.2. Exercises.

Using exponential functions to study hyperbolic trigonometric functions

1. Many calculus books simply define the hyperbolic sine and cosine functions by the rules

\[
\begin{align*}
cosh t &= \frac{e^t + e^{-t}}{2} \\
\sinh t &= \frac{e^t - e^{-t}}{2}
\end{align*}
\]

without any reference to hyperbolas, area, or hyperbolic measure of angles. (The other hyperbolic trigono-
metric functions are defined, as usual, in terms of cosh and sinh.)

Using only these definitions and properties of the exponential function, give proofs of the following:

(a) The point \((x, y) = (\cosh t, \sinh t)\) lies on the hyperbola \(x^2 - y^2 = 1\).

(b) The derivative of \(\cosh t\) is \(\sinh t\) and the derivative of \(\sinh t\) is \(\cosh t\).

(c) The derivative of \(\tanh t\) is \(\text{sech}^2 t\).

(d) \(\cosh(\phi + \theta) = \cosh \phi \cosh \theta + \sinh \phi \sinh \theta\) and \(\sinh(\phi + \theta) = \sinh \phi \cosh \theta + \cosh \phi \sinh \theta\).

(e) The hyperbolic cosine function is even and has range \([0, \infty)\).

(f) The hyperbolic sine function is odd, is injective, and has range \(\mathbb{R}\).

(g) The hyperbolic tangent function is odd, injective, and has range \((-1, 1)\).

(h) For \(\theta_0 > 0\), the area of the region bounded by the line through \((0, 0)\) and \((\cosh \theta_0, \sinh \theta_0)\), the
hyperbola \(x^2 - y^2 = 1\), and the positive \(x\)-axis is \(\theta_0/2\). (This is rather difficult to prove using
one-variable calculus, but is quite easy using the change-of-variables formula for double integrals.

Try the transformation \((r, \theta) \mapsto (r \cosh \theta, r \sinh \theta)\).

2. Solve the following equations for \(\theta\):
(a) \( \cosh \theta = 7 \)

(b) \( \sinh \theta = -4 \)

(c) \( \tanh \theta = \frac{3}{10} \)

(Hint: write in terms of \( e^\theta \) and \( e^{-\theta} \). In parts (a) and (b), multiply by \( 2e^\theta \), and use the quadratic formula. Part (c) does not require the quadratic formula.)

3.

(a) In Exercise 2, you should have found two solutions in part (a), and one solution in parts (b) and (c). Using the right-hand branch of the hyperbola with equation \( x^2 - y^2 = 1 \), explain why this was to be expected.

(b) Verify that the two solutions you obtained to the equation \( \cosh \theta = 7 \) are opposites of each other.

(Why is this expected?)

4. Express the hyperbolic cosine, sine, and tangent functions in terms of exponential functions. Make small tables of values for these functions, and use the tables to give rough graphs of the functions. Compare your graphs to those produced by a graphing calculator or computer algebra system.

Inverse hyperbolic trigonometric functions

5. Give reasonable natural domains and ranges of the inverse functions of the hyperbolic cosine, sine, secant, and tangent functions. (Begin with your answers to Your Turn 11 and Exercises 4, 5, and 6 from Section 3. Do you need to restrict the domain of any hyperbolic trigonometric function to make it one-to-one?)

6. On the basis of your answer to Exercise 5, find simple formulas for each of the following, along with the values of \( t \) for which each expression is defined:

(a) \( \sinh(\text{arcsinh } t) \)

(b) \( \text{arcsinh}(\sinh t) \)

(c) \( \cosh(\text{arccosh } t) \)
(d) \( \text{arccosh}(\cosh t) \)

(e) \( \sinh(\text{arccosh} t) \)

(f) \( \cosh(\text{arcsinh} t) \)

7. Compute the derivatives of the inverse functions of the hyperbolic sine, cosine, secant, and tangent functions. Show your work. (Use the equation that relates the derivatives of a pair of inverse functions.)

(b) Check your answers using a computer algebra system.

(c) Compare your answers to the derivatives of the inverse functions of the (circular) cosine, sine, secant, and tangent functions.

8. Prove the following rules for the inverse hyperbolic trigonometric functions:

(a) \( \text{arctanh}(x) = \ln \sqrt{1 + x} + \frac{1}{2} \ln \frac{1 + x}{1 - x} = \frac{1}{4} (\ln |1 + x| - \ln |1 - x|) \)

(b) \( \text{arcsinh}(x) = \ln(x + \sqrt{1 + x^2}) \)

(c) \( \text{arccosh}(x) = \ln(x + \sqrt{x^2 - 1}) \)

(Use the strategies from Exercise 2, taking due care with domains and ranges.)

9. Using your answers to Exercise 8, obtain expressions for the derivatives of the inverse hyperbolic tangent, sine, and cosine functions. Simplify your answers. (This gives an alternate strategy to Exercise 7.)

Additional exercises

10. (This is a continuation of Exercise 6.)

Find formulas for each of the following (in terms of exponential and logarithmic functions), along with the values of \( t \) for which each expression is defined:

(g) \( \text{arcsinh}(\cosh t) \)

(h) \( \text{arccosh}(\sinh t) \)
11. Use a computer algebra system to plot the functions in (g) and (h) of Exercise 10. Compare these to graphs of the analogous circular trigonometric functions. Why are the graphs simpler in the circular case?

12. Find an exact value for the area of the sector of the hyperbola (with equation $x^2 - y^2 = 1$) determined by the arc with endpoints $(2, \sqrt{3})$ and $(4, \sqrt{15})$.

13. Express the eigenvalues of the hyperbolic rotation $H_\phi = \begin{pmatrix} \cosh \phi & \sinh \phi \\ \sinh \phi & \cosh \phi \end{pmatrix}$ in terms of $e^\phi$ and $e^{-\phi}$.

7. Linking the Two Trigonometries

In Chapters 4 and 5, we have developed circular and hyperbolic trigonometry, and revealed parallels between the two theories. In sections 7.1 and 7.2, we discuss two explicit ways of connecting the circular and hyperbolic trigonometric functions.\[18]

7.1. Circles and hyperbolas together. We begin by relating hyperbolas and circles (see Figure 16).

We have marked an arbitrary point $H$ on the right-hand branch of the hyperbola with equation $x^2 - y^2 = 1$, and a point $C$ on the circle with equation $x^2 + y^2 = 1$; given $H$, $C$ is uniquely determined by the requirement that $C$ and $H$ lie on a ray with vertex at the origin. There is a bijective function $\Phi$ from the right-hand branch of the hyperbola to the quarter-circle with $x^2 + y^2 = 1$ and $|y| < x$, defined by $\Phi(H) = C$. Equally well, $\Phi^{-1}$ is a bijection from the quarter-circle to the right-hand branch of the hyperbola.

We can reinterpret $\Phi$ in terms of hyperbolic and circular angles and their measures. We know that there is a real number $\theta_h$ such that $H = (\cosh \theta_h, \sinh \theta_h)$, and there is a real number $\theta_c$ (between $-\pi/4$ and $\pi/4$) such that $C = (\cos \theta_c, \sin \theta_c)$. This defines a bijection function $\phi : \mathbb{R} \to (-\pi/4, \pi/4)$, assigning the output $\theta_c$ to the input $\theta_h$. The function $\phi$ conveys the same information as the function $\Phi$.

It is easy to describe $\phi$ in terms of trigonometric functions. From Figure 11, by considering the slope of the ray from the origin through $C$ and $H$, we have that $\tan \theta_c = \tanh \theta_h$, from which we conclude that $\phi(\theta_h) = \arctan(\tanh \theta_h)$. The inverse function $\phi^{-1}$ is given by the rule $\arctanh \circ \tan$.

Finally, we recall the relationship between angle measures and area. Let $A_h$ denote the area of the hyperbolic sector determined by $(0,0)$, $(1,0)$, and $H$, and let $A_c$ denote the area of the circular sector determined by $(0,0)$, $(1,0)$, and $C$. Recall that $\theta_h = 2A_h$ and $\theta_c = 2A_c$. We conclude that $A_c = \frac{1}{2} \phi(2A_h)$.

\[18\] Further relationships will be discussed in Chapter 11.
7.2. Parameterizations of the hyperbola, and the Gudermannian function. We have seen that the right-hand branch of the hyperbola with equation $x^2 - y^2 = 1$ is conveniently parameterized by the map $t \mapsto H(t) := (\cosh t, \sinh t)$. But equally well, the branch can be parameterized by the map $s \mapsto Q(s) := (\sec s, \tan s)$, with $\pi/2 < s < \pi/2$.

Your Turn 20.

(a) Explain why for all $t$, $Q(s)$ lies on the hyperbola.

(b) Explain why the range of $Q$ is the entire right-hand branch of the hyperbola.

Therefore, there is a function $f$ defined by $Q(f(t)) = H(t)$; otherwise said, $s = f(t)$ is the value of the parameter that makes $Q(s) = H(t)$. Explicitly, we have $(\sec(f(t)), \tan(f(t)) = Q(f(t)) = H(t) = (\cosh t, \sinh t)$, from which it follows that $f(t) = \arctan(\sinh(t))$. The bijective function $f : \mathbb{R} \to (-\pi/2, \pi/2)$ is called the Gudermannian function$^{19}$ and naturally, its inverse function $g$ is the inverse Gudermannian function.

$^{19}$The German mathematician Christoph Gudermann (1798–1852) began his career as a mathematics teacher in a secondary school in Kleve, and later became a professor in Münster. He was a student of Gauss, and as a teacher, had a strong influence on Weierstrass, who worked for his secondary school teacher’s certificate in Münster.
7.3. Exercises. These exercises make use of facts about hyperbolic trigonometric functions, their inverse functions and derivatives, and trigonometric identities, as explored in the exercises from previous sections.

1. Relating $\phi$ and $f$.
   
   (a) Prove that $f(t) = 2\arctan(\tanh(t/2))$. (The double angle identity for $\tanh$ is helpful.)
   
   (b) Deduce that $\phi(t) = \frac{1}{2} f(2t)$.
   
   (c) Find a similar expression for $f(t)$ in terms of $\phi$.

2. 
   
   (a) Use graphing technology to plot $f$. Is your answer consistent with what you expect the range of $f$ to be?
   
   (b) According to Exercise 1(b), how is the graph of $\phi$ obtained from the graph of $f$?

3. From the definition of $f$, we have that $\tan \circ f = \sinh$. Compute and simplify $\sin \circ f$ and $\cos \circ f$. In both cases, you will obtain one of the six basic hyperbolic trigonometric functions.

4. Prove that $f$ is an antiderivative of the hyperbolic secant function.

5. Prove that $g$ is an antiderivative of the (circular) secant function.

6. Prove that $g(t) = \ln(\sec t + \tan t)$.

7. Prove that $f(t) = 2\arctan(e^t) - \pi/2$. 
CHAPTER 6

Numbers

*Number systems* are the collections of numbers that we learn about from elementary school through university:

- \(N\) natural numbers
- \(W\) whole numbers
- \(Z\) integers
- \(Q\) rational numbers
- \(R\) real numbers
- \(C\) complex numbers

In this chapter, we will concern ourselves with the sets \(N, W, Z, Q,\) and \(R\). The complex numbers, as well as the operations on all of these number systems, will be studied in Chapter 7. To prepare the way, consider the following exercise:

**Let’s Go 1.**

(a) What *intuitive pictures* or *physical models* do you have in mind to represent that some number \(x\) is *less than* some other number \(y\)?

(b) Do these pictures/models work equally well for whole numbers, for fractions, and for decimals? Do they work for negative numbers?

(c) What *computational techniques* do you have to show that \(x < y\)?

(d) Consider the computational techniques that you have identified. Do they apply to whole numbers (in base ten notation)? Do they apply to fractions? Do they apply to decimals? Do they apply to negative numbers?
During the long course of your mathematical education, you probably will examine these number systems from two distinctly different points of view. The first glimpse at number systems, starting from toddler years, is intuitive. The natural numbers are the numbers we use for counting objects; rational numbers (written as fractions) are the symbols we use to describe parts of a whole; real numbers (probably written as decimals) mark our position along a line. The operations of addition, subtraction, multiplication, and division arise naturally as we encounter daily life problems, like dividing a box of doughnuts among five people, or figuring out how many cans of paint we'll need to paint a room.

On the other hand, mathematics majors also may see a rigorous development of these number systems. This development relies on a great deal of mathematical background that we typically acquire in the first two years of college: sets, functions, sequences, limits, and (of overriding importance) the notion of an equivalence relation. If you study number systems rigorously—perhaps in this chapter, perhaps in still greater depth in advanced mathematics courses—you will gain an appreciation for the value of a careful discussion of the number systems. For example, you will learn that progress in calculus would have ground to a halt without a common understanding among mathematicians of the properties of real numbers. You will learn about the crisis in mathematics that occurred when mathematicians struggled to put basic notions of mathematics (like “sets” and “integers”) on a rigorous footing.

As you work to integrate your dual roles as mathematician and as educator, it will be important for you to appreciate, and to connect, the intuitive and rigorous views of number systems. An appreciation of the rigorous definitions will enrich the classroom experiences you design for your students. It will protect you against the misconceptions than can arise from superficial understanding, and equip you with new tools for answering students’ questions. Having said that, let’s recognize that none of us would ever begin a middle-school lesson with a phrase like “We define a rational number to be an equivalence class of symbols \((a, b)\), where…” That would be ridiculous, since the rigorous definitions only make sense to us after years of schooling in which we develop our intuitive pictures. The importance of these intuitive models should never be underestimated. They point to the critical applications of basic mathematics that are essential to your students’ success in daily life and virtually every profession, and they plant the seeds of understanding that will lead to success in higher mathematics for your students. This is why a good math teacher can give
several explanations of what multiplication “means,” and can give quick arguments to explain why \( ab = ba \) or why the distributive law holds. Reaching this level of familiarity with the intuitive models is an important goal for you to set for yourself as you read this chapter, and the next.

1. The Natural Numbers and Whole Numbers: An Intuitive View

1.1. Numbers and counting. The “natural” place to begin a study of number systems is with the natural numbers

\[ \mathbb{N} = \{1, 2, 3, 4, 5 \ldots \} \]

or perhaps a little more generally with the whole numbers

\[ \mathbb{W} = \{0, 1, 2, 3, 4, 5 \ldots \} \].

If we asked ourselves where these numbers “come from,” or where we first encountered them, most of us would say that they come from counting. This act of counting is one of our first mathematical activities, and remains indispensable throughout our mathematical experience.

We learn to count by considering a wide variety of sets of objects: these sets might be a serving of peas on a plate, a collection of coins on a table, a herd of cattle in a field, a string of lights on a Christmas tree, or any set of objects of which we wish to keep track. At the end of the counting process, we assign a number to each set. The conceptual milestone we all made early in life was that even though sets of cattle have very little in common with sets of coins, the procedure for counting cattle is the same as that for counting coins.

**Your Turn 1.** Give several examples that illustrate the connection between counting and measurement. *(You might consider ways of measuring length or volume that can also be interpreted as “counting.”)*

It might be difficult to visualize “all possible sizes of herds of cows” or “all possible serving sizes of peas,” but there is a convenient way to visualize “all whole numbers,” namely, along a line (Figure 1). This

---

1Among texts, there is not universal agreement in these definitions. Some books will call 0 a natural number, and some books don’t deem 0 to be a whole number.
line should also remind us of the connection between counting and measurement: the line might as well be a yardstick (a tool for counting inches) or the scale on a big measuring cup.

Let’s recall the notion of comparison. Our years of schooling have made us very comfortable with inequalities like \(10 < 12\) and \(128 > 100\). Where do these abstractions come from? We might see a herd of cattle on one hillside and another herd across the valley; we might take the time to count, and learn that there are 10 on the hillside and 12 across the valley, and conclude that there are “more cattle across the valley than on the hill.” That would be an eminently practical way of answering the question, “Which location has more cattle?” However, probably the notion of “more” and “fewer” comes to us even earlier than counting. We learn very early, that if we have a pile of peas on our plate, and Mama unloads another spoonful, then we have more than we had originally—and this does not require us knowing the precise number that were there originally, the number added, or the final number on the plate. Likewise, one can compare the length of two sticks by holding them up against each other.

**Your Turn 2.** If Tommy has a small number of peas on his plate and so does Catherine, how can they conclusively learn who has more peas, without actually counting them? (Your solution will show that one can compare two finite sets without counting their elements.)

In terms of our “number line model,” the statement \(10 < 12\) reflects that 10 occurs to the left of 12. If we interpret 10 as the entire segment on the number line from 0 to 10, then the fact that 10 is less than 12 is connected tangibly to the statement that a ten-foot stick is shorter than a twelve-foot stick. (See Figure 2.)

**Your Turn 3.** Is the number zero important for counting or for measuring? In what ways is zero important, even for discussing natural numbers?
1. We know that in the collection of whole numbers \{3, 8, 30, 2, 10\}, the smallest element is 2 and the largest is 30.

   (a) Give an example of a collection of whole numbers that does not have a largest element.
   
   (b) Does every collection of whole numbers have a smallest element?
   
   (c) Give an example of a collection of whole numbers where the smallest and largest elements are equal.

2. (Trichotomy Property) We all know that if \(a\) and \(b\) are whole numbers, then exactly one of the following is true: \(a = b\), \(a < b\), or \(b < a\). Interpret this statement in terms of a number line and give an intuitive explanation of the truth of the statement.

3. (Transitive Property) We all know that if \(a\), \(b\), \(c\) are whole numbers and \(a < b\) and \(b < c\), then \(a < c\). Interpret this statement in terms of a number line and give an intuitive explanation of the truth of the statement.

4. In the set \{3, 8, 30, 2, 10\}, we have failed to list the elements in increasing order. Describe an algorithm that allows one to take the list 3, 8, 30, 2, 10 (or any finite list of natural numbers) and arrange them in
increasing order. (Make certain that the steps in your algorithm are “simple.” Avoid complicated steps such as “pick the smallest element in the list,” since this already requires a multi-step algorithm.)

5. Take a blank piece of paper and (randomly) draw approximately thirty dots on it.

Now imagine that you are seeing the piece of paper for the first time and need to count the number of dots.

(a) Count them.

(b) Describe the strategy you used for counting them. (How did you decide on the order in which to count them? How did you keep track of which dots were already counted?)

(c) Describe several alternate strategies for counting the dots. Which strategies require marking the page? Which strategies would be equally effective if there were 200 dots on the page? Would the distribution (pattern) of the dots affect your choice of strategy?

6. Exercise 5 is interesting, insofar as few of us have the ability to glance at the page and immediately recognize that there were precisely thirty dots on it. By experimentation, determine how many dots on a page you and your classmates can determine immediately (without counting “one at a time” and without dividing the dots into groups).²

7. If you have experience with small children, describe the counting errors you have observed. Through what “stages” do children progress as they become more effective in counting objects?

2. The Integers: An Intuitive View

2.1. The need for negative numbers. From elementary school, we are familiar with the integers

\[ \mathbb{Z} = \{ \ldots -4, -3, -2, -1, 0, 1, 2, 3, 4 \ldots \} \]

What motivates the introduction of “negative numbers”? 

There are simple real-world situations for which the use of whole numbers is inadequate, even misleading. For example, in your relationship with a bank, the number 1,750 (as a number of dollars) is ambiguous. It might represent your deposit of $1,750, held by the bank, which must be returned to you on demand; it

²This topic has been researched extensively. Try an internet search with key work *subitization*. 
might instead represent a loan of $1,750 that you must eventually repay. These are “opposite” situations (in
the sense that owing $1,750 is the opposite of being owed $1,750), and of course a familiar way to represent
this is through the placement of a dash in front of the number in the case of a debt in your relationship with
a bank. Thus $−1,750 would represent a debt to the bank, whereas $1,750 represents a credit. We regard
1,750 as a whole number (just as in the last section), which we call “positive,” while regarding $−1,750 as a
“new number,” called “negative.”

Your Turn 4. In what other real-world settings do negative numbers arise?

To represent the whole numbers and the negative integers at the same time, we again use positions along
a line, typically with positive numbers to the right and negative numbers to the left (see Figure 3).

![Figure 3. Integers on the number line](image)

As in the case of whole numbers, for integers we also interpret $a < b$ to mean that $a$ is to the left of $b
on the number line.

Your Turn 5.

(a) We have used integers to describe our relationship with a bank. In this interpretation, what does
$a < b$ mean? (Think of the integers $a$ and $b$ as describing Alice and Bob’s relationship with the
bank, respectively.)

(b) In contrast, what does the statement $|a| < |b|$ mean (still in the context of banking)?

2.2. Exercises.

1.

(a) Give an example of a collection of integers that has both a smallest element and a largest element.
(b) Give an example of a collection of integers that does not have a largest element, but has a smallest element.

(c) Does every collection of integers have a smallest element?

2. Find a closed interval on the number line that is 0.9 units long, and contains no integers.

3. Suppose $a$ and $b$ are integers. State in words what $a > b$ means in terms of the following models for the integers:
   
   (a) Elevation above/below sea level
   
   (b) Temperature
   
   (c) Annual excess/deficit of rainfall compared to normal
   
   (d) Time before/after liftoff of a rocketship

4. In each item in Exercise 3,
   
   (i) What units would be reasonable to use?
   
   (ii) Does every integer (no matter how large or small) represent a meaningful physical situation? (Discuss each example separately.)
   
   (iii) Are there physically meaningful quantities in the examples that cannot be represented by integers?

5. A student attempts to define absolute value in the following way: “If the number is $x$, then the absolute value is $x$, and if the number is $-x$, then the absolute value is $x$.” Is this definition correct? Can you identify any assumptions or misconceptions on the student’s part? Can you give a clearer definition of absolute value?

6. Is there any good reason that we use positive numbers to represent credits and negative numbers to represent debts, as opposed to the reverse?

3. The Rational Numbers: An Intuitive View

**Let’s Go** 2. First create several convincing explanations for the fact that $\frac{2}{10} = \frac{9}{15}$. Then, for each of your explanations, determine whether:
(a) you used a picture representing an intuitive model (e.g., a number line)

(b) you used a computational technique representing an algorithm.

(c) your explanation is algebraic or geometric.

(d) your explanation uses technology.

3.1. Parts of a whole. If you want a broom handle that is longer than 4 feet, you aren’t limited to a broom handle of 5 or more feet. In the same way, a bank account can reflect a balance or a debt that is a portion of a dollar, and by adding a few pebbles to the top of a mountain, you can guarantee that its elevation is not a whole number of feet above sea level. These physical realities, all of which involve dividing some entity into parts, motivate our understanding of rational (and real) numbers.

Specifically, fractions (and the rational numbers they represent) are conceptually related to parts of a whole and the process of subdivision. In a typical model for a fraction, we might imagine cutting a pie into four pieces. The fraction $\frac{1}{4}$ or $\frac{1}{4}$ is used to signify the amount of pie contained in one of the four pieces, the fraction $\frac{2}{4}$ represents two of the four pieces, etc. The number $\frac{5}{4}$ represents . . . five of the four pieces? Well, that is probably not the best language to use for improper fractions. It might be better to speak of “five copies of a fourth of a pie” or “five pieces of pie, each equal to a fourth of a pie.”

Any object that can be subdivided can be used in modeling fractions, and conversely, fractions are the appropriate abstraction to represent real-world examples of subdivision. The hope and expectation of our elementary school teachers was that by understanding two-thirds of a pie and two-thirds of an inch, we students would come to understand some less tangible parts of a whole that are important for functioning adults, like two-thirds of a year, two-thirds of the world’s people, or two-thirds of the calories in a frankfurter.

What ties these disparate examples together? To make a representation or model of a fraction, we need a unit. The unit is some object which we consider to be “whole,” but which we will ultimately subdivide into pieces. The possible models for units are endless. Besides “years” and “caloric content of a hot dog,” examples include a single cookie, a box of 12 cookies, a line segment, or some region in the plane like a rectangle (see Figure 4). Regardless of our choice of model, we let the number 1 represent the unit.

3 Most of us would picture a whole pie plus a quarter of another pie, which steers us toward the mixed number $1\frac{1}{4}$. 
Once we have a unit, we are ready to make the following informal definition. Let \( a, b \in \mathbb{W} \) be whole numbers, with \( b \neq 0 \). The fraction \( a/b \) represents ‘\( a \)’ parts of a unit which has been split into ‘\( b \)’ equal parts.

Of course, models for a given fraction will depend on our choice of unit. For example, Figure 5 shows models for \( 1/6 \) and Figure 6 shows models for \( 5/3 \), using the units from Figure 4.

**Your Turn 6.** Explain how Figure 7 can be used to model both \( 3/4 \) and \( 4/3 \). (What is the unit in each case?)
A crucial concept we all had to master was the notion of equivalent fractions. Let’s remind ourselves how equivalent fractions are related to “parts of a whole.” Consider the fractions 3/5, 6/10, 9/15, and 18/30, depicted in linear and area models in Figures 8 and 9. Observe that the models for 6/10 and 9/15 are obtained from the model for 3/5 by simply making further subdivisions of the unit (each piece of the unit is divided in two for 6/10, and is divided in three for 9/15), while the overall portion of the unit represented by each fraction remains the same. Informally, we say that $a/b$ and $c/d$ are equivalent if they represent the
same overall portion of the unit. From the preceding discussion, we may see that given a fraction \( a/b \) and a nonzero whole number \( c \), the fractions \( a/b \) and \( (ac)/(bc) \) are equivalent.

\[ \text{Figure 8. Equivalent fractions: linear model} \]

In this example, we conclude that \( 6/10 \) and \( 9/15 \) are equivalent to each other, since both are equivalent to \( 3/5 \), which we should recognize as the “simplest form” (or “reduced form”) of both \( 6/10 \) and \( 9/15 \), or as the outcome of putting both \( 6/10 \) and \( 9/15 \) into “lowest terms.” An alternative strategy for showing the equivalence of \( 6/10 \) and \( 9/15 \) is to subdivide the unit further, by a factor of three for \( 6/10 \) and a factor of two for \( 9/15 \) (also illustrated in the Figures). This reveals that \( 6/10 \) is equivalent to \( 18/30 \) and so is \( 9/15 \). All told, this gives two distinctly different “subdivision strategies” for showing that \( 6/10 \) and \( 9/15 \) are equivalent.\(^4\)

\textbf{Your Turn 7.} We have verified the equivalence of \( 6/10 \) and \( 9/15 \), first by showing that both are equivalent to \( 3/5 \), and second by showing that both are equivalent to \( 18/30 \).

(a) Why does 5 arise as the denominator of a fraction equivalent to both \( 6/10 \) and \( 9/15 \) in our verification?

\(^4\)In the next section, we’ll revisit the “cross multiplication rule” that is commonly used to check whether fractions are equivalent.
Figure 9. Equivalent fractions: area model

(b) Why does 30 arise as the denominator of a fraction equivalent to both 6/10 and 9/15 in our verification?

(c) What other denominators might students use in showing the equivalence of 6/10 and 9/15? Why?

Your Turn 8. Show how to use a subdivision strategy to conclude that 4/7 > 2/7.

The number line model provides an extremely important conceptual tool in understanding rational numbers (see Figure 10). This figure indicates several important concepts. First, it collects all the rational numbers together, allowing us to compare numbers (as with the integers, being “smaller” coincides with “lying to the left”). Second, observe that fractions lying in the same vertical column are equivalent to one another. Third, it becomes clear that rational numbers can be chosen arbitrarily close to each other, and that between any two rational numbers lie many (even infinitely many!) other rational numbers. Thus, for rational numbers, unlike integers, there is no sensible way of defining “the next rational number after a given rational number.” Likewise there are no “gaps” in the rational numbers—that is, there are no intervals of positive length that are devoid of rational numbers.
Your Turn 9. In Figure 10, we see that each rational number occurs many times as a fraction (always along the same vertical line). What is special about the top fraction that occurs along each vertical line? (Hint: consider how the numerator and denominator of this fraction are related to each other.)

3.2. Exercises.

1. Model each fraction using each of the following units: a solid circle (pie), a line segment, and a rectangle.

   (a) 4/7.
   
   (b) 7/4.
   
   (c) 5/8.
2. Use a linear model, and also a rectangular area model, to model \( \frac{4}{7} \), together with two fractions that are equivalent to it.

3. Use the two subdivision strategies from the text to show that \( \frac{30}{36} \) and \( \frac{35}{42} \) are equivalent fractions. (You need not draw the linear and area models, but you should identify the new fractions that are equivalent to both \( \frac{30}{36} \) and \( \frac{35}{42} \), and the “subdivision factors.”)

4. Consider the following pairs of rational numbers. Use the idea of subdivision (or common denominators) to determine whether the first number is larger than, equal to, or smaller than the second number. Briefly explain how your method uses the idea of subdivision.
   
   (a) \( \frac{17}{31} \) and \( \frac{19}{37} \).
   
   (b) \( \frac{3}{4} \) and \( \frac{57}{86} \).
   
   (c) \( \frac{2}{1111} \) and \( \frac{3}{1888} \).

5. Give a plausible explanation for the term rational number. (Hint: rational is not used in the sense of logical, and it’s important that students in school not think that irrational numbers are “illogical.”)

6. In what real-world settings does one need to use negative rational numbers?

7. Mrs. D. has attempted to teach her students an algorithm for comparing two positive fractions. (See Figure 11.) Suzie concludes that \( \frac{6}{15} < \frac{4}{9} \), while Jimmie finds that \( \frac{6}{15} > \frac{4}{9} \).
   
   (a) Which student is correct?
   
   (b) Give a precise statement of Mrs. D.’s algorithm.
   
   (c) Explain Jimmie’s mistake in implementing Mrs. D.’s algorithm.
   
   (d) Give a mathematical justification of Mrs. D.’s algorithm.

8. Suzie has been attempting to generalize Mrs. D.’s algorithm (from Exercise 7) to rational numbers that need not be positive. She has been experimenting with fractions such as \( \frac{1}{2} \), \( -\frac{1}{2} \), \( \frac{1}{2} \), and \( \frac{-3}{3} \) and she suspects that Mrs. D.’s algorithm is not correct.
Conduct a similar investigation to find an efficient algorithm for comparing the fractions \( \frac{a}{b} \) and \( \frac{c}{d} \), where \( a, b, c, d \) are integers with \( b \neq 0 \) and \( d \neq 0 \). Give a mathematically convincing argument to explain why your algorithm is correct.

9. On the subject of comparing fractions, one of your students announces, “The bigger the denominator, the smaller the fraction.” In what sense is your student correct? In what sense might he be incorrect?

10. A novice teacher uses the diagram in Figure 12 to model \( 1/4 \). What criticisms can you make of this choice?
11. Find three rational numbers lying strictly between $2/3$ and $4/5$. Explain how you found your answers.

12. Suppose that our “unit” is a solid cube. Indicate with a sketch how this unit can give a volume model for the fraction $1/8$.

13. Use technology to find a simplest-form fraction that is equivalent to $\frac{123,889}{9,085,117}$. Do the same for $\frac{244,944}{684,288}$. Are these fractions easy to put in simplest form without technology? Discuss.

14. It’s “Big Numbers Day” in Mrs. D.’s 8th grade math class. She asks the class to consider the numbers

\[ x = \frac{10^{50} + 3}{7 \cdot 10^{20} + 2} \quad \text{and} \quad y = \frac{10^{50} + 6}{7 \cdot 10^{20} + 4} \]

and to try to figure out which one is bigger.

- Jimmie responds that they are equal, since his calculator says they are both equal to $1.42857142 \times 10^{29}$.
- Suzie says they are equal, since $\frac{3}{7} = \frac{6}{4}$, and aside from that, you’re adding the same thing to the numerators and you’re adding the same thing to the denominators.

Imagine that you are Mrs. D.

(a) Respond to Jimmie. In your response, indicate whether his reasoning is correct, and carefully explain why.

(b) Do the same for Suzie.

(c) Which is bigger ($x$ or $y$), or are they equal? You might want to begin with some scratchwork, but be certain that you end up giving a clear and convincing explanation. Please don’t give an answer that relies on technology (calculator or computer).

4. The Real Numbers: An Intuitive View

Let’s Go 3. We’ve seen earlier that rational numbers can be used to quantify parts of a whole. What are irrational numbers good for? Discuss.
4.1. The number line. Intuitively, the “real numbers” are intended to account for all points along the number line. We will consider a real number \( a \) to be less than another real number \( b \) if \( a \) lies to the left of \( b \) on the number line. This gives a satisfying completion of our pictures of the previous number systems (whole numbers, integers, and rational numbers), since these number systems are represented by some, but not all of the points on the line.

Computationally, we are all experienced in using decimals to represent points on the line:

Your Turn 10.

(a) Explain how a decimal like 3.582 produces a point on the number line. (Precisely how does one use the digits of 3.582 to mark a location on the number line?)

(b) Explain how a nonterminating decimal, like the decimal expansion of \( \pi \), also produces a point on the number line. How does this differ from the case of a terminating decimal?

The simple intuitive idea—that the real numbers “form a line”—is actually very complex, as we shall see in Section 8. At present, let’s pass the following milestone, by considering the basic question, “Why isn’t every point along the line a rational number?” After all, if every point on the line were rational, the real numbers would be the same as the rational numbers, and there would be no need for this section at all.

Indeed, there are real numbers that are not rational, but this is neither trivial nor intuitive. It would be easy to be misled by Figure 10, which shows that there are lots of rational numbers. By considering all fractions with denominator 2, 3, 4, 5, 6..., or even merely the denominators 2, 4, 8, 16..., we see that any point on the line can be approximated by a rational number, up to an error that we can make as small as we like.

Nevertheless, if we believe the Fundamental Theorem of Arithmetic\(^5\) (that is, the fact that all integers larger than one can be written uniquely as a product of primes), we can produce irrational numbers.\(^6\) Specifically, we’ll show that \( \sqrt{2} \) is a real number that is not rational. First, let’s remind ourselves why there should be a real number called “the square root of two.” Observe that by the Pythagorean Theorem, we may

\(^5\)The Fundamental Theorem of Arithmetic will be proved in Chapter 8.

\(^6\)The existence of irrational numbers was well-known to the ancient Greeks. Plato (4-th century BCE) asserts that the irrationality of certain square roots had been demonstrated by a certain Theodorus of Cyrene.
easily construct a segment of length $\sqrt{2}$ as the hypotenuse of a right triangle whose two legs are of length 1 (see Figure 13). Thus $\sqrt{2}$ is a real number. Why is $\sqrt{2}$ irrational? Suppose to the contrary that $\sqrt{2}$ were rational. As a consequence, there would exist nonzero natural numbers $a$ and $b$ such that \((\frac{a}{b})^2 = 2\), that is, $a^2 = 2b^2$. Let’s see how this leads to a contradiction. For any natural number $k$ we let $n_k$ represent the number of 2’s in the prime factorization of $k$. (For example, since $120 = 2^3 \cdot 3^1 \cdot 5^1$, we have $n_{120} = 3$.) From the prime factorizations for $a$ and $b$ guaranteed by the Fundamental Theorem of Arithmetic, one may obtain factorizations for $a^2$ and $2b^2$. We then find that $n_{a^2} = 2n_a$ (why?), which is even, while $n_{2b^2} = 2n_b + 1$ which is odd. Thus $n_{a^2} \neq n_{2b^2}$, from which we conclude $a^2 \neq 2b^2$ by the uniqueness part of the Fundamental Theorem of Arithmetic.

The points on the number line that do not correspond to rational numbers are dubbed *irrational numbers*. We’ve seen, without too much difficulty, that $\sqrt{2}$ is not rational. Similar arguments show that $\sqrt{n}$ is irrational, for every whole number $n$ that is not the square of a whole number. For similar reasons, $\log_3 7$ is irrational, as you will show in the Exercise 4. More difficult arguments have been found that show that $\pi$ and $e$ are irrational. If you have studied “cardinality” in another math course, you will know that irrational numbers are not a rarity: there are “many more” irrational numbers than rational numbers. The exercises will provide an opportunity to ‘write down’ some of these numbers.
At this point, we clarify some mental pictures we might have of both the rational and the real numbers. Here is a fact about the real numbers:

- **Between any two different real numbers, there is another real number (for example, the midpoint of the segment joining the two numbers).**

However, before we think that such a statement somehow “characterizes” the real numbers, we should remember that this statement is already true about the rational numbers! So, what property can we identify that truly distinguishes the rational numbers from the real numbers? A rough version is:

- **There are no holes in the real numbers, whereas there are holes in the rational numbers.**

A “hole” should mean “a number that is missing.” We have already seen that $\sqrt{2}$ is missing from the rational numbers; on the other hand, our intuitive picture of a line suggests an absence of such holes. This is a very non-rigorous statement, and mathematicians’ ability to prove many mathematical statements (the sorts that occur at key moments in calculus) would be hampered without a precise articulation of the “no holes” property. We do so in Section 8, where we give a more rigorous definition of the real numbers.

### 4.2. Absolute value.

Let’s record an official definition of the absolute value of a real number. The definition should be familiar, but we’ll state it since it will be used so often in the text. It relies on the fact that every real number has an opposite.

**Definition 1.** If $x$ is a real number, then we define the absolute value of $x$, denoted $|x|$, by the rule $|x| = x$ if $x \geq 0$, and $|x| = -x$ if $x < 0$.

Observe that $|x|$ has a clear geometric meaning: it’s the distance between $x$ and 0, or equally well, the length of the segment on the number line with 0 and $x$ as endpoints.

### 4.3. Exercises.

1. One of your students appears to believe that $0.125 > 0.8$, and when you ask why, he responds, “because 0.125 has more numbers.” Discuss the source of the student’s confusion and a strategy for remediation.
2. In the text, we used a right triangle with side lengths of 1 and 1 to produce $\sqrt{2}$ as a length (the length of the hypotenuse). Explain why we can’t get $\sqrt{3}$ as the length of the hypotenuse of a right triangle whose legs have whole number lengths.

3. A teacher states, “there are no gaps in the rational numbers,” but also, “there are holes in the rational numbers.” Carefully explain the difference between “having gaps” and “having holes.”

4. In the text, we showed that the Fundamental Theorem of Arithmetic implies that $\sqrt{2}$ is not rational. In a similar way, show that $\log_3 7$ is not rational.

5. Give a sequence of steps that shows how to construct (using compass and straightedge) a segment of length $\sqrt{2}$, beginning with a segment of length 1.

6. In this exercise, we see that two numbers written using radicals may be equal to each other, without this equality being transparent.

   Let $p(x) = x^4 - 10x^2 + 1$.

   (a) Let $r_1 = \sqrt{2} + \sqrt{3}$, $r_2 = \sqrt{2} - \sqrt{3}$, $r_3 = -\sqrt{2} + \sqrt{3}$, and $r_4 = -\sqrt{2} - \sqrt{3}$. By substitution, show that each of these numbers is a root of $p(x)$.

   (b) Using the quadratic formula, find expressions for the four roots of $p(x)$.

   (c) Identify each of the roots in (b) as $r_1$, $r_2$, $r_3$, or $r_4$. How many strategies can you find for correctly pairing the two lists of roots?

   (d) If you had been given $p(x)$ and had be asked to find all the roots, would you have been more likely to find the list in (a) or in (b)?

   (e) Using the (known) roots of $p(x)$, find three different ways of factoring $p(x)$ into a product of two quadratic polynomials. (To find these quadratic polynomials, do you prefer using the list of roots in (a) or the list in (b)?)

7. Let $q(x) = x^2 - 14\sqrt{2}x + 87$. Find a degree-four polynomial $p(x)$ with integer coefficients whose roots include the two roots of $q(x)$. What are the other two roots of $p(x)$? Explain your strategy.
8.

(a) The real number $37 + 20\sqrt{3}$ can be written as the square of a number of the form $a + b\sqrt{3}$ (where $a$ and $b$ are integers). Find $a$ and $b$.

(b) Find a simpler expression for the number $\sqrt{37 + 20\sqrt{3}} - \sqrt{12}$. Explain your strategy.

9. In school, many of us learned a procedure called “rationalizing the denominator.” As a typical example, we computed that

$$\frac{1 + 7\sqrt{2}}{5 + 3\sqrt{2}} = \left(\frac{1 + 7\sqrt{2}}{5 + 3\sqrt{2}}\right) \left(\frac{5 - 3\sqrt{2}}{5 - 3\sqrt{2}}\right) = -\frac{37 + 32\sqrt{2}}{7} = -\frac{37}{7} + \frac{32}{7}\sqrt{2}$$

(a) What is the purpose of “rationalizing the denominator”?

(b) Perform a similar computation to simplify $\frac{1 + 7i}{5 + 3i}$. (Perhaps you are familiar with this procedure through your study of complex numbers.) What seems to be the purpose of this computation?

(c) More generally, suppose $q$ is a rational number that is not the square of another rational number.

Prove that for every choice of rational numbers $\{a, b, c, d\}$ (with $cd \neq 0$), the real number $\frac{a + b\sqrt{q}}{c + d\sqrt{q}}$ can be written as $A + B\sqrt{q}$ for some other rational numbers $A$ and $B$.

(d) Investigate: how might you “rationalize the denominator” if the expression is $\frac{1}{2 + 7\sqrt{2} + 5\sqrt{3}}$?

(e) Investigate: how might you “rationalize the denominator” if the expression is $\frac{1}{\sqrt{2} + 1}$?

5. The Natural Numbers and Whole Numbers: A More Rigorous View

5.1. A confession about set theory. We must begin our discussion with an admission: our treatment of the whole numbers will be sketchy, leaving many subtle but important points unaddressed. Why is this?

Whole numbers are concerned with counting. What does one count? We might answer, “We count the objects in some collection of objects.” Indeed, the objects we count are clumped together in collections, like the five peas we might count on our plate. In mathematics courses, we routinely call these collections of objects “sets.” However, mathematical discoveries in the late 19th century revealed that contradictions arise if one treats the notion of ‘set’ or ‘collection’ so casually. Specifically, Russell’s Paradox (see Exercise 9) shows that one cannot regard the collection of all sets as a set itself, at least if one expects to do the typical constructions we all do with sets. (For example, we expect that arbitrary unions and intersections of sets will
again be sets, and we expect that the collection of objects within a set for which some property holds will again be a set.\footnote{As an example of “forming the subset of a set consisting of the elements for which some property holds,” consider the set of all real numbers \( x \) such that the property \((x - 1)(x - 5) > 0\) holds. In mathematics we often form sets this way, so it’s clearly something we want set theory to permit.} In light of Russell’s Paradox, the mathematical notion of a “set” has to be more restrictive than our intuitive notion of a “collection of objects.” The solution to this conundrum is to present a list of axioms that are assumed to hold for those collections of objects that are deemed to be “sets.” This was accomplished in the early 20th century, yielding amongst others the Zermelo-Fränkel axioms of set theory.\footnote{An internet search will produce a list of the Zermelo-Fränkel axioms for you. It’s fair to say that most working mathematicians are not experts in this area, and that most research in mathematics goes on simply viewing sets as collections of objects.}

It would take a separate book to motivate and describe this system and to use it to give a truly complete account of the whole numbers. We will try to make our discussion of the whole numbers as approachable as possible, but the subtleties of set theory will lead us through some treacherous waters.

5.2. Finding the next biggest number: the successor function. In counting, we wish to associate a whole number to any set.\footnote{Richard Dedekind (1831-1916) was the first to propose a ‘set theoretical’ definition of the natural numbers. Subsequently, other (equivalent) definitions were offered by Gottlob Frege, Georg Cantor, and finally Guiseppe Peano. The uniqueness of natural numbers follows from Dedekind’s Recursion Theorem (1888).} We have an intuitive/symbolic representation of the whole numbers as the list 0, 1, 2, 3, 4, etc., with each numeral representing a number that is “larger by one” than its predecessor. We might first ask, “Given any whole number that we might intuitively imagine, is there really a set with that number of elements?”\footnote{For example, have you actually ever seen a set with 23, 465, 937 elements?} Intuitively, we get big sets by starting with smaller sets and adding elements “one at a time.” The mathematically rigorous way to discuss this is through the notion of a successor function \( S \) defined as follows: Given a set \( A \), we define \( S(A) \) to be \( A \cup \{A\} \). This probably looks weird at first glance, so let’s see an example: if \( A = \{a, b, c, d\} \) (a set with four elements), then \( S(A) \) is the set \( \{a, b, c, d, \{a, b, c, d\}\} \) (a set with five elements). Thus, the tool of the successor function accomplishes the intuitive goal of increasing the size of a set by one. Now as a very special case, we can do the following: we define ‘0’ to be the empty set \( \emptyset \), we define ‘1’ to be \( S(0) = \{\emptyset\} \) (the successor set of \( 0 \)), we define ‘2’ to be \( S(1) = S(S(0)) = \{\emptyset, \{\emptyset\}\} \) (the successor set of 1), etc.

Your Turn 11. Use the notion of the successor function to list the elements of each of the following sets: \( 0, 1 = S(0), 2 = S(S(0)), 3 = S(S(S(0))), \text{ and } 4 = S(S(S(S(0)))) \).
The previous discussion\textsuperscript{11} gives a rough idea how one obtains the set of whole numbers via a successor function, but in order to do anything with more with this idea, such as defining operations on the whole numbers and rigorously establishing the properties thereof, we will need to be more precise in our definition of the whole numbers.

**Definition 2.** The whole numbers are a nonempty set \( W \) together with a function \( S : W \to W \) and an equivalence relation ‘=’ satisfying the following properties, known as the Peano\textsuperscript{12} axioms:

1. \( W \) contains an element 0.
2. \( S(n) \in W \) for every \( n \in W \), and \( S(n) \neq 0 \).
3. \( S(m) = S(n) \) implies \( n = m \) for all \( n, m \in W \).
4. (Induction Principle) If \( K \subset W \) such that
   i. \( 0 \in K \), and
   ii. \( n \in K \) implies \( S(n) \in K \)

then \( K = W \).

We shall refer to these properties from time to time as we proceed.

5.3. **When do two sets have the same number of elements?** Whole numbers are supposed to be useful for counting. What, in fact, do these mysterious sets 0, \( S(0) \), \( S(S(0)) \), etc. have to do with counting? Intuitively, our intention is that a set should “have three elements” if it “has the same number of elements as \( S(S(0)) \).” We need to make clear what we mean by “having the same number of elements.” What makes a set of ten cows have “the same number of elements” as a set of ten peas? The reason is that we can match up perfectly the ten cows and the ten peas (for instance, by giving each of the cows exactly one pea to eat). In mathematical terms, “matching up” two sets means producing a bijective function from one set to the other. If we call the set of cows \( A \) and the set of peas \( B \), we have a bijection \( f : A \to B \), defined by the rule \( f(\text{cow}) = \text{the pea eaten by the cow} \). Why is \( f \) a bijective function? First, \( f \) is a function, since

\textsuperscript{11}Of course in the discussion above, we have “done some things” with sets, and we should remark that these operations are consistent with the Zermelo-Fränkel axioms of set theory. Specifically, the axioms assert the existence of the empty set, and allow the formation of unions (which we needed to form successor sets). Moreover, the axioms assert that “the collection of whole numbers” also is a set.

\textsuperscript{12}These are due to Guiseppe Peano.
each cow eats exactly one pea. The function $f$ is injective (one-to-one), since two different cows don’t eat the same pea. The function $f$ is surjective (onto) since each pea gets eaten. Since $f$ is both injective and surjective, it is a bijection, as we claimed.\(^{13}\)

This example motivates the definition of *equivalence of sets*:

**Definition 3.** Let $A$ and $B$ be sets. We say that $A$ is equivalent to $B$ if there exists a bijection (that is, a function that is both one-to-one and onto) $f : A \rightarrow B$. If $A$ is equivalent to $B$ then we write $A \sim B$.\(^{14}\)

You will verify that this property is reflexive, symmetric, and transitive in Exercise 5.

To finish the story of counting, we simply say that a set has $n$ elements if it is equivalent to the set ‘$n$’ = $\{\{\emptyset\}, \{\{\emptyset\}\}, \ldots \}$ formed via successor sets.\(^{15}\) This agrees with our grade-school procedure for counting, which we can see by formally discussing the example of counting ten cows. We can let $A$ be the collection of ten cows, and let $B$ be the list\(^{16}\) of words \{one, two, three, four, five, six, seven, eight, nine, ten\}, which is the list of “names” for our successor sets. Then we have a bijection $f : A \rightarrow B$ given by

$$f(\text{cow}) = \text{the word we say when we are pointing at that cow.}$$

Thus as we recite the words “one, two, three…” while we run through the objects we are counting, the word on our lips as we point to the last object is the word we use for “the number of objects in the set.”

**Your Turn 12.** Explain how the process of counting given immediately above could be described equally well by a function $g : B \rightarrow A$. Give an explicit definition of the function $g$. How is it related to the function $f$?

**Your Turn 13.** Let’s peer in on a three-year-old attempting to count a penny, nickel, dime, and quarter. Let $A = \{\text{“one”, “two”, “three”}\}$, $B = \{\text{“one”, “two”, “three”, “four”, “five”}\}$, and let $C = \{\text{penny, nickel, dime, quarter}\}$.

---

\(^{13}\)To review the meanings of *injective*, *surjective*, and *bijective* see Chapter 1, Section 5.

\(^{14}\)Additionally, we must declare the empty set to be equivalent to itself but to no other set. Many of the statements in this section require some additional (trivial) statement to cover the case of the empty set. To keep our discussion brief, we will omit these statements.

\(^{15}\)However, in the full story, one has to verify that the successor sets ‘0’, ‘1’, ‘2’… are mutually non-equivalent.

\(^{16}\)We are calling $B$ a set, but for the purposes of counting, the elements of $B$ must be arranged in the usual order!
(a) Suppose Tommy counts the penny, nickel, and dime correctly, but forgets to consider the quarter.

Define a function \( h : A \to C \) which could describe the correspondence (there is more than one possibility). Is \( f \) a bijection? If not, why not?

(b) Suppose Tommy counts the penny twice by pointing to it and saying “one, two”, and then counts the other coins correctly. Define a function \( j : B \to C \) which could describe the correspondence (there is more than one possibility). Is \( g \) a bijection? If not, why not?

(c) Suppose Tommy points to the penny and says “one,” then points to the nickel and says “one,” and continues to count the rest of the coins correctly. Is there a function \( k : A \to C \) describing the correspondence? If so, give an example. If not, why not?

We have accomplished our goal of defining the whole numbers, and showing why they are used for counting. The remainder of this section deals with some special issues on the topics of sets and cardinality.

5.4. Comparing the size of two sets. How do we make precise the notion of a set being “bigger” than another set? If \( A \) and \( B \) are sets, we write \( A \preceq B \) if there is a one-to-one function from \( A \) to \( B \). It is easy to see that this property is transitive. It can be shown that given any two sets \( A \) and \( B \), either \( A \preceq B \) or \( B \preceq A \). Moreover, it’s true (but not obvious) that if \( A \preceq B \) and \( B \preceq A \), then \( A \) and \( B \) can be put in one-to-one correspondence, and thus are equivalent (this result is called the Schröder-Bernstein Theorem).

When \( A \) and \( B \) are finite sets, we are merely describing the usual notion of \( \leq \) for the whole numbers. That is, if \( a \) and \( b \) are whole numbers, then the statement “\( a \leq b \)” about whole numbers is shorthand for the following statement about sets: There exists a set \( A \) with \( a \) elements and a set \( B \) with \( b \) elements, such that \( A \preceq B \).

5.5. Finite and infinite sets. The notion of “infinity” already arises in elementary school. Namely, what would happen if we attempted to count the natural numbers themselves? We would say “one” as we point to “one,” we would say “two” as we point to “two,” et. cetera... but since there would be no “last”

\[17\] Another astounding discovery from the late 19th century was the fact that there exist infinite sets that are not equivalent to each other. This will be discussed in the Exercises.
object we would point to, there would be no “word on our lips as we point to the last object,” and hence, there would be no “number of elements in the set.” This is how we grew accustomed to thinking of \( \mathbb{N} \) as an “infinite set.”

Intuitively, we picture that finite sets are the sets that are “small enough” to be counted using the whole numbers, and infinite sets are “too big” to be counted using the whole numbers. In rigorous terms, one can declare a set to be finite if it’s equivalent to one of the sets \( S_n \), and infinite if it’s not.\(^{18}\)

However, we prefer a different definition of finiteness:

**Definition 4.** A set is said to be *finite* if and only if it is not equivalent to any of its proper subsets.\(^{19}\)

This is a very clean definition, but it requires some motivation. If we start with a set with ten elements, it is plausible that if we throw out one or more of the elements, then the resulting subset cannot be put in one-to-one correspondence with the original set. Indeed, our intuitive notion of finite sets is very closely tied to this very picture, that any proper subset of a finite set has fewer elements. On the other hand, if we start with our mental picture of the whole numbers and throw out one element (say, the number zero), then the resulting set (the natural numbers) still is in one-to-one correspondence with the original set, namely, by the bijective function \( f : \mathbb{W} \rightarrow \mathbb{N}, f(x) = x + 1 \). Even more, one can show that if a set \( X \) contains a subset that is in one-to-one correspondence with \( \mathbb{W} \), then one can do the same trick (find a one-to-one correspondence between \( X \) and some proper subset). This connects our two definitions of finiteness. (You will be invited to explore this in detail in Exercise 7.)

### 5.6. Equivalence classes of finite sets?

Finally, we would like to sketch an appealing, but ultimately unsatisfactory way to define the whole numbers. As above, one can define what it means for two sets to be *equivalent*. Next, one can define what it means for a set to be *finite* (it is not equivalent to any of its proper subsets). Since the notion of equivalence is reflexive, symmetric, and transitive, it’s tempting to define the whole numbers to be the equivalence classes of finite sets.

---

\(^{18}\)To complete the picture, if \( B \) is an infinite set one needs to verify \( S_n \leq B \) and \( B \) is not in one-to-one correspondence with \( S_n \) for all \( n \).

\(^{19}\)This definition is due to Richard Dedekind, appearing in *Was sind und was sollen die Zahlen?* (1888).
This looks fine, but there is a subtle logical problem: by Russell’s Paradox, there is no “set of all finite sets,” and consequently, there is no “set of equivalence classes of finite sets.” (See Exercise 10.)

5.7. Exercises.

1. Let $A = \{a, b, c, d\}$, $B = \{b, a, c, d\}$, $C = \{x, y, z\}$, $D = \{x, y, t\}$, $E = \{l, m, n, p\}$, $F = \{u, v\}$, $G = \{u, v, w\}$, and $H = \{c, a, d, b\}$.

(a) List all pairs of equivalent sets from the sets given above.

(b) List all pairs of equal sets from the sets given above.

(c) For each pair of sets in part (a), give an explicit one-to-one correspondence between the sets, thus showing the sets in each pair are equivalent. (You may draw pictures (mapping diagrams) to define the functions.)

(d) If two sets are equal, are they necessarily equivalent? Explain.

(e) How many bijective functions from $A$ to $E$ exist?

2. List five sets in the equivalence class denoted by the whole number 3, as well as one set that is not in this class.

3. Consult the textbook used in the mathematics content course for elementary education majors at your university. Read its development of the whole numbers, particularly its definitions of set, finite set, and whole number. Which of the definitions appear to be rigorous, and which definitions appear to be vague? How effective is the textbook’s strategy in discussing bijections and equivalence of sets? Does it use alternate terminology for these notions? Finally, how does the discussion of whole numbers balance and integrate the intuitive models and the rigorous definitions?

4. Suppose $f : A \to B$ and $g : B \to C$ are bijections. Show that the composition $g \circ f$ is a bijection. (Break this into two parts: show that the composition of two injective functions is injective, and the composition of two surjective functions is surjective.)

5. Prove that the property of “equivalence” for sets in Definition 3 is reflexive, symmetric, and transitive.

You may use Exercise 4.
6. Show that any subset of a finite set is finite. (Hint: Suppose $B$ is finite and $A \subset B$. Let $C$ be a proper subset of $A$. For a contradiction, assume that there is a bijection $f : A \rightarrow C$. Use $f$ to construct a bijection from $B$ to a proper subset of $B$, thus obtaining a contradiction. You will need to use the official definition of “finite” in Definition 4.)

7. Let $A$ be a set, let $B$ be a subset of $A$, and suppose there exists a bijective function $f : \mathbb{W} \rightarrow B$. Prove that there is a bijection from $A$ to a proper subset of $A$. (This exercise is meant to demonstrate that a set that contains a copy of the whole numbers is “infinite” in the sense of Definition 4.)

8. A basic but astounding fact that was only discovered in the late 19th century is the following: not all infinite sets are equivalent. This says (roughly) that there are “different sizes of infinity.” This idea is often explored in a discrete mathematics or real analysis class, where one sees that $\mathbb{W}$, $\mathbb{Z}$, and $\mathbb{Q}$ are equivalent, but these are not equivalent to $\mathbb{R}$.

In this exercise, we take a different route to showing there are inequivalent infinite sets. Recall that the power set of a set is the set of all subsets of the set. For example, if $X = \{a, b, c\}$, then the power set $\mathcal{P}(X)$ has eight elements, namely

$$\emptyset \{a\} \{b\} \{c\} \{a, b\} \{a, c\} \{b, c\} \{a, b, c\}.$$  

**Theorem 5.** No set is equivalent to its power set.

In particular, the power set of $\mathbb{W}$ is not equivalent to $\mathbb{W}$. This gives two inequivalent infinite sets.

Your job is to prove Theorem 5, using the following outline: attempt a proof by contradiction. Assume that $f : X \rightarrow \mathcal{P}(X)$ is a bijection. Call an element $x \in X$ good if $x \in f(x)$ and bad if $x \notin f(x)$. Let $B \subset X$ be the set of all bad elements. Show that $B$ is not in the range of $f$.

9. In this exercise, we show why it is nonsense to speak of “the set of all sets.”

Let $A$ be any set. Call an element $x \in A$ good if $x \notin x$ and bad if $x \in x$.\(^{20}\) Let $G$ be the set of all good elements of $A$.

\(^{20}\)The notion that a set might be an element of itself should seem odd to you, but whether there are actually any bad elements is irrelevant to this exercise.
(a) Show that it is impossible that \( G \in A \).

(b) As a consequence of (a), show that there does not exist “a set of all sets.”

10. As a follow up to the previous exercise, we will show why it is even nonsense to speak of “the set of all finite sets.” We will use a proof by contradiction.

Let \( F \) be the alleged “set” of all finite sets. Let \( O \subset F \) be the subset of \( F \) consisting of all sets with exactly one element. Let \( X \) be the union of all the elements of \( O \) (since the elements of \( O \) are sets, it makes sense to speak of the union of them). One might think of \( X \) as “the set of all objects.”

How does the result of the previous exercise lead to a contradiction? (Think of how \( X \) is related to the alleged “set of all sets,” and remember that within a set, a collection of elements determined by a property is again a set.)

11. Using a college mathematics textbook\(^{21}\) as a resource, write a report containing the following:

(a) The definition of countable\(^{22}\) set.

(b) A proof that any subset of a countable set is countable (infinite countable or finite).

(c) A proof (using decimals) that the real numbers are not countable. (This puts our assertion that “there are far more irrational numbers than rational numbers” on solid footing.)

6. The Integers: A More Rigorous View

6.1. Defining the integers. Our goal now is to define the integers, taking for granted the definition and properties of the whole numbers. We will repeat this pattern as we define the rational numbers and finally, the real numbers: when each new number system is defined, we will make free use of all the properties of the previously-defined number systems.\(^{23}\)

We explore a definition of the integers that is less intuitively obvious than the idea of the integers that we presented in Section 2, but which is more convenient for proofs.\(^{24}\) Here we think of an integer as coming

\(^{21}\)Try a textbook in foundations of mathematics, real analysis, or number theory. Search the index for the word countable.

\(^{22}\)Beware of inconsistencies among texts: some texts insist that a countable set be infinite, while others do not. The text you choose may also use the word denumerable.

\(^{23}\)According to Leopold Kronecker, who introduced the algebraic notion of integral domain in his Grundzüge einer arithmetischen Theorie der algebraischen Größen, “The Good Lord made the integers and all the rest is the work of man.”

\(^{24}\)The idea of representing integers by pairs of natural numbers originated with Dedekind.
from an ordered pair \((a, b)\), where \(a\) and \(b\) are both whole numbers. (It might help to visualize the ordered pair \((a, b)\) as an arrow along a number line, where \(a\) as the starting point of the arrow and \(b\) as the ending point. Rightward pointing arrows correspond to positive integers and leftward pointing arrows correspond to negative integers.) How does this alternative notion of an integer relate to the point of view we adopted in Section 2? We think of the positive integer 2 as represented by any of the ordered pairs \((0, 2)\), \((1, 3)\), \((2, 4)\), \((3, 5)\) \ldots \ (any arrow extending two units to the right), whereas \(-2\) is represented by the ordered pairs \((2, 0)\), \((3, 1)\), \((4, 2)\), \((5, 3)\) \ldots \ (any arrow extending two units to the left). We need an equivalence relation on the set of ordered pairs to obtain the integers, and you should convince yourself that the relation is

\[
(a, b) \sim (a', b') \quad \text{if} \quad a + b' = a' + b.
\]

We then define the integers to be the set of equivalence classes. Thus, in this alternate definition, an integer is an equivalence class \([a, b]\), where \(a\) and \(b\) are whole numbers. (Note that we are abiding by the rules: this equivalence relation is merely a statement about whole numbers.)

How do we describe the “less than” relation, using this definition of the integers? A brief investigation should convince you that the appropriate definition is

\[
[a_1, b_1] < [a_2, b_2] \quad \text{if} \quad a_1 + b_2 < a_2 + b_1;
\]

again, we are “playing by the rules” since the latter inequality is a statement about whole numbers. \(^{25}\)

6.2. Exercises.

1. Show that the relation given in Equation (1) is reflexive, symmetric, and transitive, hence an equivalence relation. (In the proof of the transitive property, feel free to use well-known properties of addition and cancellation.)

2. Explain why the order relation \(<\) given in Equation (2) appears to be appropriate.

3. Prove the transitive property of \(<\) for the integers.

\(^{25}\)It must be verified that this definition is well defined, that is, independent of the representatives of the equivalence classes. We will omit this. If you are not familiar with the notion of “well defined,” see the discussion of \(<\) in the rational numbers in the following section.
4. (a) If an integer is the equivalence class \([a,b]\), then how would one write the opposite (additive inverse) of this equivalence class?

(b) Prove that your answer in (a) is well defined. In other words, if \((a,b)\) is equivalent to \((a',b')\), you must show that the opposite of \([a,b]\) equals the opposite of \([a',b']\).

7. The Rational Numbers: A More Rigorous View

7.1. Fractions and rational numbers. Put

\[ \mathcal{F} = \{(a,b) \in \mathbb{Z} \times \mathbb{Z} \mid b \neq 0\}. \]

We call \(\mathcal{F}\) the set of fractions. In this notation, we will be writing \(\frac{4}{10}\) as \((4,10)\) and \(\frac{6}{15}\) as \((6,15)\). Strictly speaking, \((4,10)\) and \((6,15)\) are not equal fractions because they are two distinct members of the set \(\mathcal{F}\). However, we can establish an equivalence between \((4,10)\) and \((6,15)\) by defining a relation on \(\mathcal{F}\) inspired by the “cross-multiplication rule”:

**Definition 6.** Let \((a,b)\) and \((c,d)\) be fractions. We say that \((a,b)\) is equivalent to \((c,d)\) if \(ad = bc\). We write \((a,b) \sim (c,d)\) to indicate that \((a,b)\) and \((c,d)\) are equivalent.\(^{26}\)

This definition is firmly grounded in our notion of parts of a whole: We only need to recall that by the process of subdivision, we have a strategy for comparing fractions. Specifically, by a subdivision argument, we saw in Section 3 that the fractions \(\frac{4}{7}\) and \(\frac{16}{14}\) represent the same parts of a whole, so should be regarded as equivalent. We now apply this with a clever choice of the “\(k\),” which should remind you of “common denominators.” Namely, \((a,b)\) should be equivalent to \((ad,bd)\), and \((c,d)\) should be equivalent to \((bc,bd)\). Thus \((a,b)\) should be equivalent to \((c,d)\) exactly when \((ad,bd)\) is equivalent to \((bc,bd)\). However, both of these fractions come from subdividing the unit into the same number of parts \((bd\) parts); formally, we have two fractions with a “common denominator” \(bd\). These fractions will be equivalent exactly when they “count” the same number of parts, that is to say, when \(ad = bc\).

\(^{26}\)As an example, we have that \((4,10) \sim (6,15)\) since \(4 \cdot 15 = 10 \cdot 6\). We re-emphasize that our notation is chosen merely to clarify the fact that \(\frac{4}{10}\) and \(\frac{6}{15}\) are unequal fractions that are equal as rational numbers.
In Exercise 8, you will prove that the relation ∼ on \( F \) given in Definition 6 is an equivalence relation. Hence, we are able to define:

**Definition 7.** The rational numbers, denoted by \( \mathbb{Q} \), consist of all equivalence classes in \( F \) under ∼.

In our proofs about the rational numbers, we will tend to write fractions in the form \((a, b)\) and rational numbers as \([a, b]\) (since they are equivalence classes of fractions). When we are not actively involved in proving properties of rational numbers, we will use the customary fractional notation \( \frac{a}{b} \) or \( a/b \).

Next, we ponder the order on rational numbers. Here is a formal definition.

**Definition 8.** Given rational numbers \([a, b]\) and \([c, d]\), we will write that \([a, b]\) < \([c, d]\), and say that \([a, b]\) is less than \([c, d]\), if \(bd(bc - ad) > 0\) (with the usual notion of “less than” for the integers).

Of course, this definition did not drop out of the sky. Rather, it is the consequence of interpreting “less than” in physical terms (position along along the number line), along with some experimentation.

**Your Turn 14.** Give a convincing argument to explain why Definition 8 is reasonable. (Why does it give the same order relation as the number line model?)

From the rigorous point of view, a subtle but important issue arises in Definition 7. Here we are trying to define what it means for one equivalence class to be less than another equivalence class (the classes \([a, b]\) and \([c, d]\)), but we have made the definition using specific representatives of the equivalence class (namely, \((a, b)\) and \((c, d)\)). Whenever such a definition is made, one needs to verify that it is well defined. That is, we check that the definition doesn’t depend on the specific representatives. How does this play out? Let’s pick different representatives: some \((a', b')\) that’s equivalent to \((a, b)\), and some \((c', d')\) that’s equivalent to \((c, d)\). Since we could have used these representatives for the equivalence classes in Definition 8 instead of the original ones, we are required to show that

\[
0 < b'd'(bc' - ad') \quad \text{exactly when} \quad 0 < bd(bc - ad),
\]

27The rational numbers were never treated with the same suspicion borne by the negative integers or imaginary numbers. The first systematic treatment of the rationals occurs in Book VII of Euclid’s Elements. The notion of a rational number as an equivalence class of pairs of integers first appears in Wilhelm Weber’s *Lehrbuch der Algebra* (1895).

28...which you may have conducted yourself, if you solved Exercises 7 and 8 back in Section 3.

29Imagine the weirdness that would result if this were not true. Would it make any sense if \( \frac{2}{5} < \frac{8}{5} \) but \( \frac{6}{10} > \frac{24}{27} \)?
which, via multiplication, amounts to showing that $b'd'(bc - ad)(b'c' - a'd') > 0$. However,

$$\text{(3)} \quad b'd'(bc - ad)(b'c' - a'd') = b'd'(bc - ad)(bb'c'd - dd'a'b'),$$

so it suffices to show that the righthand side of (3) is positive. Finally, since $(a', b') \sim (a, b)$ and $(c', d') \sim (c, d)$,

we know that $a'b = b'a$ and $c'd = d'c$, and so

$$b'd'(bc - ad)(bb'c'd - dd'a'b') = b'd'(bc - ad)b'd'(bc - ad) = [b'd'(bc - ad)]^2,$$

which is indeed positive, so Definition 8 is well defined.

7.2. Exercises.

1. In school, we may learn phrases such as “6 is to 4 as 15 is to 10.”

   (a) Translate the phrase to a statement about fractions or rational numbers.

   (b) Give a sketch that illustrates a geometric interpretation of the phrase.

2. Consider the following pairs of rational numbers. Using the cross-multiplication rule (Definition 8), determine whether the first number is larger than, equal to, or smaller than the second number.

   (a) $[(17, 31)]$ and $[(19, 37)]$

   (b) $[(3, 4)]$ and $[(57, 86)]$

   (c) $[(2, 1111)]$ and $[(3, 1888)]$

   (d) $[(-3, 5)]$ and $[(-7, 11)]$

   (e) $[(-3, 5)]$ and $[(7, -11)]$

   (f) $[(3, -5)]$ and $[(7, -11)]$

3. Using Definition 6, verify that $(51, 102) \sim (16, 32)$, and that $(17, 51) \sim (1, 3)$.

4. List five elements of each of the following equivalence classes: $[(1, 2)], [(2, 1)]$, and $[(5, -4)]$.

5. (This exercise shows that the rational numbers are an ordered set.) Let $a, b, c$ be whole numbers with $b, c \neq 0$. Using Definition 6, verify that $(a, b)$ and $(ac, bc)$ are equivalent fractions.

6. Let $x, y, z$ be rational numbers and let $<$ be as in Definition 8.
8. The Real Numbers: A More Rigorous View

(a) (Anti-symmetry) Show that one and only one of the statements \( x < y, \ x = y, \ y < x \) is true.

(b) (Transitivity) Show that if \( x < y \) and \( y < z \) then \( x < z \).

7.

(a) What is the additive inverse of the rational number \([a, b] \)?

(b) Prove that your answer to (a) is well-defined.

8. Let \( \sim \) be as in Definition 6. Prove that \( \sim \) is reflexive, symmetric, and transitive, hence an equivalence relation. (You may use well-known properties of integer addition and multiplication.)

8. The Real Numbers: A More Rigorous View

If the points of a line are divided into two classes, in such a way that each point of the first class lies to the left of every point of the second class, then there exists one and only one point of division which produces this particular subdivision into two classes, this cutting of the line into two parts.

(R. Dedekind, Stetigkeit und irrationale Zahlen, 1872)

8.1. A formal definition. From a pedagogical standpoint, the real numbers are in a peculiar position. On the one hand, they are exceedingly familiar to all of us from our early education. We have a compelling picture of the real numbers, simply as a line. From elementary school through calculus, we use properties of the real numbers and make computations using real numbers. On the other hand, the real numbers are difficult to define rigorously, and this partially explains occasions in a calculus text where the author must write “The proof is beyond the scope of this course.”

Intuitively, we think of real numbers as all the points on the number line, with the irrational numbers serving to “fill in the holes” on the line among the rational numbers. The chief usefulness of the set of real numbers is its completeness, reflecting the fact that the “holes” in the rational numbers have been filled in. This property is crucial in mathematics, for example, in defining functions like \( \sqrt{x} \), and for virtually all of calculus. The intuitive definition of real numbers, however, is inadequate for proving statements about the real numbers. Our goal now is to give a precise definition of the real numbers that does not refer to the physical model of the line (but is inspired by it).
Two precise definitions of the real numbers, due to Cantor and Dedekind, respectively, were published in 1872. We now give a brief description of Cantor’s construction. Some information about Dedekind’s construction can be found in Exercises 13–17.

Cantor’s construction is an outgrowth of the idea that any point on the line can be approximated by rational numbers. For example, since we know that we can extract a decimal expansion for \( \sqrt{2} \) (it begins 1.41421356237...), we can think of \( \sqrt{2} \) as a “limit” of the following sequence of rational numbers:

\[
x_1 = 1, \quad x_2 = \frac{14}{10} = 1.4, \quad x_3 = \frac{141}{100} = 1.41, \quad x_4 = \frac{1414}{1000} = 1.414, \quad x_5 = \frac{14142}{10000} = 1.4142, \quad \ldots
\]

Of course, we can’t simply define \( \sqrt{2} \) to be the “limit of the sequence \((x_1, x_2, \ldots)\)” since this assumes that there is a number system in which this limit exists, when this is exactly what we are trying to establish! So a back-door strategy is needed. The trick is not merely to think of \( \sqrt{2} \) as the limit of \((x_1, x_2, \ldots)\); instead we think of \( \sqrt{2} \) as the sequence \((x_1, x_2, \ldots)\) itself. More generally, we aim to define a real number as a sequence of rational numbers—a sequence that seems to converge to a point on the number line.

There are a few technical issues. First, it is ridiculous to consider all sequences of rational numbers. For example, the sequences \((0, 1, 4, 9, 16, 25, \ldots)\) and \((0, 1, 0, 1, 0, 1, \ldots)\) do not appear to converge to any point on the line, so we need to exclude sequences like these from our construction. The “right” sequences are the ones that do seem to converge to a point on the line. These sequences are called Cauchy sequences of rational numbers:

**Definition 9.** A sequence \((x_1, x_2, x_3, \ldots)\) of rational numbers is a *Cauchy sequence of rational numbers* if given any rational number \(\epsilon > 0\), there is a positive integer \(N(\epsilon)\) such that \(|x_i - x_j| < \epsilon\) whenever \(i > N(\epsilon)\) and \(j > N(\epsilon)\).\textsuperscript{33}

\textsuperscript{30}Georg Cantor (1845-1918) and Richard Dedekind (1831-1916) were two of the premier German mathematicians of the late 19th century. In an age of acrimony in the German mathematical community, Cantor and Dedekind were friends, and kept a frequent correspondence. Cantor’s work in defining the real numbers eventually led to important breakthroughs in set theory, including the development of transfinite ordinal numbers.

\textsuperscript{31}Further details concerning the construction of the real numbers can often be found in real analysis texts. For example, Dedekind’s construction appears in Rudin’s *Principles of Mathematical Analysis* (McGraw-Hill, 1976).

\textsuperscript{32}Augustine Louis Cauchy (1789-1857) was among the first mathematicians to establish logical underpinnings for calculus by developing an acceptable theory of limits.

\textsuperscript{33}Note that subtraction has snuck in here. We will return to operations such as subtraction in the chapter *Operations in Number Systems*. 
This definition should remind the reader of the definition of the limit of a sequence. The difference is that in the definition of a limit of a sequence, the terms \( x_i \) approach some number \( L \). Here the number \( L \) is not yet available, since it might not be rational! So instead of requiring that the terms \( x_i \) approach \( L \), we require that the terms “approach each other.”

The second technical issue is that a point on the line can be represented by many different Cauchy sequences. For example, \( \sqrt{2} \) is represented by

\[
(1, 1.4, 1.41, 1.414, 1.4142, \ldots),
\]

but equally well by

\[
(2, 1.5, 1.42, 1.415, 1.4143, \ldots)
\]

(approaching \( \sqrt{2} \) from above, not below). In fact, there are myriads of Cauchy sequences that seem to approach \( \sqrt{2} \). We must represent \( \sqrt{2} \) by the whole collection of Cauchy sequences of rational numbers that seem to approach it. Mathematically, this means we want an equivalence relation on the set of Cauchy sequences. Intuitively, two sequences should be equivalent if they seem to have the same limit. A precise definition that succeeds in conveying this is:

**Definition 10.** Two Cauchy sequences \((x_1, x_2, x_3, \ldots)\) and \((y_1, y_2, y_3, \ldots)\) are equivalent if the sequence \((x_1, y_1, x_2, y_2, x_3, y_3, \ldots)\) is a Cauchy sequence.

This definition makes sense: if the sequences \((x_i)\) and \((y_i)\) approach the same point on the line, then the sequence obtained by interweaving the terms will also approach that same point. Conversely, if \((x_1, y_1, x_2, y_2, x_3, y_3, \ldots)\) approaches a point on the line, then the subsequences \((x_i)\) and \((y_i)\) clearly approach that point too.

Finally, one has the following definition of the real numbers:

**Definition 11.** A real number is an equivalence class of Cauchy sequences of rational numbers.

**Your Turn 15.** Give three Cauchy sequences of rational numbers that represent \(4/3\). (What is the easiest example?)
Cantor’s definition succeeds, but there is a tremendous amount that needs to be checked, starting with
the fact the our alleged “equivalence relation” actually is reflexive, symmetric, and transitive (Exercise 11).
Moreover, we’ll take for granted that one can add and subtract real numbers by manipulating the sequences
that define them. For example, if \((x_i)\) and \((y_i)\) are Cauchy sequences of rational numbers, giving real numbers
\(x = [(x_i)]\) and \(y = [(y_i)]\), then one can define \(x - y\) by \([(x_i - y_i)]\), which is a Cauchy sequence of rational
numbers. Likewise, we’ll take for granted that one can define the opposite of a real number as follows: if
\((x_i)\) is a Cauchy sequence representing a real number, then one can define the opposite of the real number
using the Cauchy sequence \((-x_i)\). Some of these issues will be explored in Exercises 10 and 11.

8.2. Comparing real numbers. Let’s turn to the order properties of real numbers. Viewing (intu-
itively) real numbers as points on a line, and \(a < b\) as meaning that \(a\) is to the left of \(b\), we all are ready to
accept the following proposition:

**Proposition 12.** Let \(x, y, z \in \mathbb{R}\).

(i) Exactly one of the three statements \(x < y\), \(x = y\), \(x > y\) is true.

(ii) If \(x < y\) and \(y < z\), then \(x < z\).

To treat “less than” in a rigorous sense, we need to appeal to the formal definition of real numbers.
Unfortunately, the definition of \(<\) is quite complicated in Cantor’s model. Recall that real numbers are
equivalence classes of Cauchy sequences of real numbers. Given \([(x_i)]\) and \([(y_j)]\), we will say that \([(x_i)] < [(y_j)]\)
if \([(x_i)] \neq [(y_j)]\) and if there exists a whole number \(N\) such that \(x_n < y_n\) whenever \(n > N\). This definition
may seem a bit complicated; it will be explored in Exercise 9.34

8.3. No holes in the reals. The key distinction between the rational numbers and the real numbers
is that the real numbers “have no holes.” Mathematically, this is reflected in two commonly-used properties
of the real numbers: the Least Upper Bound Property and the Completeness Property. Students who take
a course in real analysis become very familiar with these properties, since they are crucial ingredients in

\[34\] It is already work to show that this definition is well-defined, that is, that the statement about equivalence classes doesn’t
depend on the choice of representatives of the equivalence classes.
proving many theorems. We will not prove these properties, but they can be proved using either Cantor’s or Dedekind’s definition of the real numbers. Further details can be found in the Exercises.

Let’s begin our discussion of the Least Upper Bound Property:

**Definition 13.** Let \( S \) be any nonempty subset of the real numbers. An upper bound for \( S \) is any real number \( y \) such that \( x \leq y \) for all \( x \in S \).

For example, if \( S = \{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots\} \), then any number \( y \) with \( y \geq 1 \) serves as an upper bound for \( S \).

Obviously, 1 is the smallest number that is an upper bound for \( S \), and we call it the least upper bound of \( S \).

In general, if a set \( S \) has an upper bound, then a least upper bound for the set is an upper bound for the set that is less than or equal to all upper bounds of the set.

**Proposition 14.** (Least Upper Bound Property) If \( S \) is a nonempty set of real numbers, and if \( S \) has an upper bound, then \( S \) has a least upper bound.

For example, let \( S \) be the collection of all real numbers that are less than \( \sqrt{2} \). The upper bounds for \( S \) are all the numbers that are greater than or equal to \( \sqrt{2} \). The least upper bound is \( \sqrt{2} \) itself (which does not happen to lie in the set \( S \) itself, but that is not a requirement of the definition).

Note that \( \mathbb{Q} \) does not possess the least upper bound property. For example, if we let \( S \) be the collection of all rational numbers that are less than \( \sqrt{2} \), then the (rational) upper bounds for \( S \) would be the set of all the rational numbers that are greater than \( \sqrt{2} \). This latter set does not have a least element—there is no “smallest rational number greater than \( \sqrt{2} \).” This shows that the failure of the Least Upper Bound Property for the rational numbers is intimately tied with the existence of “holes” in the rational numbers.

We now move on to the Completeness Property. This property revolves around the relationships among Cauchy sequences, limits, and convergence.

**Definition 15.** A sequence of real numbers \((x_1, x_2, x_3, \ldots)\) is a Cauchy sequence if, given any real number \( \epsilon > 0 \), there exists a whole number \( N(\epsilon) \) such that \( |x_i - x_j| < \epsilon \) whenever \( i \geq N(\epsilon) \) and \( j \geq N(\epsilon) \).

The next definition is the familiar one from calculus:
Definition 16. A sequence of real numbers \((x_1, x_2, x_3, \ldots)\) converges to the real number \(L\) if, given any real number \(\epsilon > 0\), there exists a whole number \(N(\epsilon)\) such that \(|L - x_i| < \epsilon\) whenever \(i \geq N(\epsilon)\).

One can prove the following property, which quite accurately reflects the absence of holes in the real numbers:

Proposition 17. (Completeness Property) Every Cauchy sequence of real numbers converges to a real number.

8.4. Exercises.

1. Give three Cauchy sequences of rational numbers that represent \(\pi\).

2. For each set below state whether the set has an upper bound. If the set has an upper bound, state the least upper bound.
   
   (a) \(W\)
   (b) the set of negative integers
   (c) the open interval \((3, 7)\)
   (d) the closed interval \([3, 7]\)

3. If \(S\) is a nonempty set of irrational numbers, and if \(S\) has an irrational upper bound, does \(S\) necessarily have a least irrational upper bound? Discuss.

4. Prove that every point on the line has distance less than 0.01 from an irrational number. (Hint: think about the collection of rational multiples of \(\sqrt{2}\). By the way, there is nothing special about 0.01.)

5. Find two Cauchy sequences of rational numbers, \((x_i)\) and \((y_i)\), such that \(x_i < y_i\) for all \(i\), but \([-[(x_i)] = [-(y_i)]\).

6. Students often learn about limits of sequences in an intuitive way, bypassing the “\(\epsilon\) definition.” They might express the fact that a sequence of real numbers \((x_i)\) converges to \(L\) by saying that the terms of the sequence get “closer and closer” to \(L\).
(a) Unfortunately, this choice of language is imprecise. Illustrate this by finding a sequence that doesn’t converge, but has the property that the terms get “closer and closer” to a number $L$.

(b) Likewise, find a sequence that does converge to a limit $L$, but whose terms might be interpreted as not getting “closer and closer” to $L$. (In a convergent sequence, must each term be closer to $L$ than its predecessor?)

7. Prove the converse of Proposition 17: any convergent sequence of real numbers is a Cauchy sequence.

8. (For students who have studied formal logic and quantifiers.) A basic fact about the rational and real numbers is, Given any real number $x$ and any positive real number $\epsilon$, one can find a rational number $q$ such that $|x - q| < \epsilon$. One can state the fact in logical symbols as

$$\forall x \in \mathbb{R} \quad \forall \epsilon > 0 \quad \exists q \in \mathbb{Q} \quad |x - q| < \epsilon$$

(a) Explain why this fact is true.

(b) A student attempts to state this rule as, Given any real number $x$, there is a rational number that is infinitessimally close to it. How would you translate this statement into logical symbols? Is it a true statement?

9. A student studying Cantor’s model of the real numbers says, “A real number $[(x_i)]$ should be less than a real number $[(y_i)]$ exactly when $x_i < y_i$ for all $i$.” Why would this definition lead to undesirable consequences?

10. Suppose that $x = [(x_i)]$ is a real number (equivalence class of Cauchy sequences of rational numbers).

   (a) How would you define the opposite (additive inverse) of $x$? Justify your answer.

   (b) Prove that your answer to (a) is well-defined.

11. Prove that the equivalence relation on Cauchy sequences of rational numbers is reflexive, symmetric, and transitive.

12. We have indicated in the text that properties of the real numbers are crucial in the proofs of some of the most important theorems in calculus. This exercises deals with two of these theorems.
(a) The **Maximum Value Theorem** states that if $f$ is a continuous function on $[a, b]$, then there is a number $c \in [a, b]$ such that $f(c) \geq f(x)$ for all $x \in [a, b]$.

Find an example of $f$, $a$, and $b$, such that there is no rational number $c$ meeting the requirements of the theorem.

(b) The **Intermediate Value Theorem** states that if $f$ is a continuous function on $[a, b]$, and if $c$ is between $f(a)$ and $f(b)$, then there exists a $d$ in $[a, b]$ such that $f(d) = c$.

Find an example of $f$, $a$, $b$, and $c$, such that $f(x)$ is rational whenever $x$ is rational, and such that $c$ is rational, but such that no rational $d$ exists with $f(d) = c$.

Exercises 13–17 are concerned with Dedekind’s construction of the real numbers. Dedekind defined a real number to be a “cut” (defined below). Here is the motivation. A point $p$ on the line essentially gives a way of “cutting” the rational numbers on the line into either two or three pieces, namely:

(i) the rational numbers to the left of $p$

(ii) the rational numbers to the right of $p$

(iii) $p$ itself (if $p$ is a rational number).

Conversely, given a set of the form (i), we can recover the point $p$ as the “missing right endpoint.”

In Dedekind’s construction, a real number is (more or less) a set of the form (i). All that is missing is a characterization of these sets, without referring to the “line”:

**Definition 18.** A **Dedekind cut** is a collection $C$ of rational numbers satisfying the following conditions:

- $C$ is not empty and is not all of $\mathbb{Q}$.
- If $x, y \in \mathbb{Q}$, $x \in C$, and $y < x$, then $y \in C$.
- If $x \in C$, then there exists some $y \in C$ with $y > x$ (a cut has no largest element).

Finally, we have the definition

**Definition 19.** A **real number** is a Dedekind cut.
13. Prove that $\{x \in \mathbb{Q} : x < \frac{2}{3}\}$ is a Dedekind cut.

14. Give an intuitive argument to explain why $\{x \in \mathbb{Q} : x \leq 0\} \cup \{x \in \mathbb{Q} : x > 0 \text{ and } x^2 < 2\}$ is a Dedekind cut.

15. Suppose the real number $x$ corresponds to the cut $C$, and the real number $x'$ corresponds to the cut $C'$. How does one interpret the statement $x < x'$ in terms of $C$ and $C''$? Is transitivity of the order relation easy to prove?

16. Is the union of two cuts a cut? Is the intersection of two cuts a cut? Is the complement of a cut a cut?

17. If the real number $x$ corresponds to the cut $C$, how do you describe the cut that corresponds to $-x$?

9. Exercises Involving Student Work: Fractions and the Number Line

For the Problem below,

- Complete the given task yourself. Compare your responses with those of a partner or small group.
- Read through the student responses provided.
- Comment on the quality of each student response. Identify ways in which the students’ thinking is correct and ways in which the student’s thinking is incorrect or incomplete.

Two students in your class, Jim and Sue, are debating about where to place the fraction $1/5$. They have already placed the fractions $1/4$ and $1/6$ on the number line. Jim says that $1/5$ should be placed exactly in the middle of $1/4$ and $1/6$, since $1/5$ is halfway between $1/4$ and $1/6$. Sue isn't sure whether Jim’s claim is right.

(a) Is $1/5$ halfway between $1/4$ and $1/6$?

(b) Explain how you determined your answer.

(c) Place the fractions $1/4$, $1/5$, and $1/6$ on the number line, as accurately as possible.
Let’s Go 1. Describe several mathematical patterns that a student could discover by examining a 9 × 9 addition or multiplication table.

We were well served when our teachers made us “learn our math facts” and master key computational skills. But it would be tragically inadequate if a student’s expertise in number systems were limited to memorizing their addition or multiplication tables, or even to facility in computing with multi-digit numbers, fractions and decimals. If we realize the meaning of operations in number systems (that is, addition, subtraction, multiplication, and division), then we are closer to having a true understanding and appreciation of these operations. In Chapter 6 we saw that numbers themselves ‘mean something’: there are a host of intuitive models and physical representations (collections of discrete objects; lengths of sticks; parts of a whole; the number line; areas of regions…) that give concrete meaning to the abstract notions of whole number, real number, rational number, and real number. As we shall see throughout this chapter, the abstract notions of addition and multiplication (and subtraction and division too) have concrete meaning within these systems, and the familiar properties of these operations (such as the distributive property) have concrete justifications through these intuitive models. A firm grasp of the meaning of these operations is crucial for students in both applying mathematics to problems and in laying the foundation for algebra.

You can read most of this chapter profitably without mastering all of the material in Chapter 6. The following intuitive ideas are sufficient for most of this chapter:

- The natural numbers \( \mathbb{N} = \{1, 2, 3, 4 \ldots \} \) and the whole numbers \( \mathbb{W} = \{0, 1, 2, 3, 4 \ldots \} \) are familiar to us because of their role in counting.
7. OPERATIONS IN NUMBER SYSTEMS

- The integers \( \mathbb{Z} = \{ \ldots -3, -2, -1, 0, 1, 2, 3 \ldots \} \) are obtained from the whole numbers by appending to \( \mathbb{W} \) a symbol \(-x\) for each natural number \( x\).

- Fractions are used to represent parts of a whole (see Chapter 6, Section 3). Two fractions are equivalent if they represent the same portion of the whole, and by a subdivision argument, it’s easy to see that \( \frac{a}{b} \) and \( \frac{c}{d} \) are equivalent exactly when \( ad = bc \). The rational numbers are the equivalence classes of fractions.

- Real numbers can be pictured intuitively as the points along a line. This model will serve us well as we study operations on real numbers.

The more rigorous definitions from Chapter 6 are necessary only in Sections 1.2 and 5, and in some of the B exercises.

1. Addition and Subtraction in Number Systems

Let’s Go 2. How would you convince your students of the reasonableness of the following facts: \((-10) + 3 = -7\), \(10 + (-3) = 7\), \((-3) + 10 = 7\), \(3 + (-10) = -7\), and \((-3) + (-7) = -10\)? Do your explanations use formal rules, or do they use intuitive models? What properties of addition are you using?

1.1. Intuitive models. We will start with addition of whole numbers. Our most basic picture, going back early in elementary school, involves joining collections of objects. One might think of joining a group of 4 apples together with another group of 3 apples to make 7 altogether (see Figure 1). We will refer to this as a discrete model for addition. The discrete model for addition can be phrased in terms of sets: If

\[ a \text{ and } b \text{ are whole numbers, pick finite sets } A \text{ and } B \text{ with } |A| = a \text{ and } |B| = b, \text{ along with the important requirement that } A \cap B = \emptyset \text{ (why?)}. \]

We then compute \( a + b \) by determining \(|A \cup B|\). For this model to be valid we need to check that if \( A \sim A' \) and \( B \sim B' \) (and \( A' \cap B' = \emptyset \)), then \( A \cup B \sim A \cup B' \) (see Exercise 8). This mathematical statement is a rigorous version of an intuitively obvious statement, namely, that it

Figure 1. Joining apples: a discrete model for addition
doesn’t matter which sets of 3 or 4 objects one uses in showing $3 + 4 = 7$ (apples, pears, sausages...), so long as the two sets are disjoint.

In Exercise 6, you will be asked for a set-theoretic model for $a - b$, where $a$ and $b$ are whole numbers with $a \geq b$.

We have other ways of interpreting whole numbers aside from the discrete model, so we can look for other ways to model addition. Let’s think for a moment about measurement. If we interpret 3 as the length of a stick and 4 as the length of another stick, then “3 + 4 = 7” is interpreted as the fact that gluing the two sticks end-to-end produces a stick of that is seven units long (see Figure 2). This a linear model for addition. However, other “measurable attributes” lead to models of addition of whole numbers: adding the surface area of non-overlapping regions in the plane (an area model); mixing two containers of water in a single large container (a volume model), etc.

Let’s move on to addition of (positive) rational numbers. Just as we were able to think of addition of whole numbers as “joining whole objects,” we now think about addition of positive rational numbers as “joining parts of a whole.” If we have a common subdivision of the two fractions, this is purely a matter of counting the parts, as we see in Figure 3. Otherwise, we need to replace the fractions with equivalents that

![Figure 2. Joining line segments: a linear model for addition](image)

![Figure 3. $3/4 + 7/4 = 10/4$](image)
use a common subdivision (our old friend, the common denominator). We illustrate this with a linear model in Figure 4 and with an area model in Figure 5. Figure 4 is important, since when we ignore the subdivision,

\[
\frac{3}{4} + \frac{2}{3} = \frac{9}{12} + \frac{8}{12} = \frac{17}{12}
\]

in a linear model

Figure 4.

\[
\frac{3}{4} + \frac{2}{3} = \frac{9}{12} + \frac{8}{12} = \frac{17}{12}
\]
in an area model

we have the basic fact that in the linear model, addition of positive rational (or even real) numbers means adding lengths. In terms of our all-important number line model, we can say the following: to add a positive number \( a \) and a positive number \( b \), start at position \( a \) on the number line, then walk a total distance of length \( b \) to the right. The real number at the place where one stops is \( a + b \). This is a standard explanation in elementary school texts.

Your Turn 1. Suppose you have taught the rules for adding and multiplying fractions:

\[
\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \quad \text{and} \quad \frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}
\]
Your students object, “Why can’t we just add fractions by adding numerators and denominators, like with multiplication?” What responses can you give to aid your students’ understanding?

The number 0 plays an important role in addition, since \( a + 0 = a \) for all real numbers \( a \). Zero is termed the additive identity or the identity for addition, since \( a + 0 \) is “identical” to \( a \), for all \( a \).

Now let’s think about subtraction. Suppose first that \( a \) and \( b \) are whole numbers with \( a \geq b \geq 0 \); then \( a - b \) should be the number of objects left when \( b \) objects are removed from a collection of \( a \) objects. For (positive) rational numbers, one need only substitute “parts of the whole” for “objects.” For (positive) real numbers in general (including whole numbers and positive rational numbers), the length model is very compelling: if \( a \) and \( b \) are real numbers with \( a \geq b \), then \( a - b \) is the length of the stick that must be joined to a stick of length \( b \) to obtain a stick of length \( a \). Hence, we can view \( a - b \) as the solution to a problem, the Subtraction Problem: What number \( x \) must be added to \( b \) to obtain \( a \)? In symbols, this is the algebra problem of finding \( x \) so that \( b + x = a \).

This context for subtraction (only considering \( a - b \) if \( a \geq b \)) is set up to avoid the need for negative numbers as the solution to subtraction problems. Indeed, students typically are exposed to subtraction before learning about negative numbers. Now it’s time to extend our definition of addition (and subtraction) to include negative numbers. This is a topic that requires focused attention on our part: over our years of schooling, we may become so accustomed to principles like \( a - b = a + (-b) \) that we lose sight of the mathematical issues involved, the intuitive models that motivate the definitions, and the difficulty that some students have assimilating the ideas.

Let’s think about how we interpret \( a + c \) if \( c \) is negative (as you did in Let’s Go 2). Perhaps the number line model gives the most convenient picture. To obtain \( a + c \), we start at position \( a \). Then we walk left a distance of \( |c| \).\(^1\) The stopping position is precisely \( a + c \). Whether \( a \) is positive, negative or zero is irrelevant. Hence we arrive at \( 5 + -2 \) by walking two steps to the left from 5 (arriving at 3), and we arrive at \( -7 + -4 \) by walking four steps to the left from \( -7 \), arriving at \( -11 \).

\(^1\)Note that \(|c| = -c \) since \( c \) is negative.
The purpose of this intuitive definition is to make a connection with subtraction: definitions of addition and subtraction should be chosen so that for all real numbers $a$ and $b$, we have that $a - b = a + (-b)$. In words: $a$ minus $b$ should equal $a$ plus the additive inverse (or opposite) of $b$. Now there is a real issue here: we have an intuitive model for $a - b$, at least if $a \geq 0$, $b \geq 0$, and $a \geq b$ (what was the model?). We have just introduced a model for addition of any two real numbers, using movement along the number line. Your job is to verify that these models are in agreement:

**Your Turn 2.** Suppose that $a \geq b \geq 0$. Show that the intuitive models of $a - b$ and $a + (-b)$ give the same “answer.” (Hint: use the “walking along the number line” model to interpret $a + (-b)$. Then, show that this length is the length that must be added to $b$ to obtain a length of $a$.)

1.2. A more rigorous viewpoint. Having revisited our intuitive pictures of addition and subtraction, we now turn to some remarks on the rigorous treatment of these operations.

A rigorous definition for addition of whole numbers relies on the formal Peano definition of the whole numbers given in Chapter 6, Definition 2. The definition is recursive: For whole numbers $a, b$ we declare

\[ a + 0 = a \quad \text{and} \quad a + S(b) = S(a + b), \]

where $S$ is the successor function described in Chapter 6, Definition 2.

We fully expect addition to be commutative, associative, etc., but these are things that must be proved (see Section 4, Exercises 4 through 6). For now, to give an idea of what these proofs are like, we verify that $0 + n = n$ for all whole numbers $n$. We do this by the induction principle given in Chapter 6, Definition 2. Observe that the result holds when $n = 0$ because $0 + 0 = 0$ by the definition of addition. Now suppose the result holds for $n$. Then, by the definition of addition followed by the induction hypothesis, we have $0 + S(n) = S(0 + n) = S(n)$. We conclude from the induction principle that $0 + n = n$ for all whole numbers $n$.

This definition of addition can also be used to establish an order on the whole numbers (see Section 5, Exercises 11 through 13): If $m, n \in \mathbb{W}$ we declare

---

\(^2\)We definitely want this to be true; the notation $a - b$ clearly intended as an abbreviation of $a + (-b)$. 

Let’s move on to \textit{addition of integers}. Recall that the integers can be defined as equivalence classes $[(a, b)]$ of pairs of whole numbers (see Chapter 6, Section 6). Under this model of the integers we can define addition by
\begin{equation}
[(a, b)] + [(c, d)] = [(a + c, b + d)]
\end{equation}
for whole numbers $a, b, c, d$. Observe that $[(0, b)] + [(0, d)] = [(0, b + d)]$, so this definition of integer addition restricts to ordinary addition on the whole numbers.

For this definition to make sense, we need to verify that if $(a', b') \sim (a, b)$ and $(c', d') \sim (c, d)$, then $(a + b' + c' + d') \sim (a' + c', b' + d')$. In other words, we must check that the definition of integer addition is \textit{well-defined}, meaning, it doesn’t depend on the representatives of the equivalence classes that occur in the definition.\footnote{To make this completely concrete: Note $2 = [(3, 5)]$ and $-3 = [(9, 6)]$ and that $2 + (-3) = [(3 + 9, 5 + 6)] = [(12, 11)]$ according to Equation (3). If however we write $2 = [(10, 12)]$ and $-3 = [(8, 5)]$, then we obtain $2 + (-3) = [(10 + 8, 12 + 5)] = [(18, 17)]$. For the definition to make sense we demand that our two results, $[(12, 11)]$ and $[(18, 17)]$, be equivalent, and indeed they are: $12 + 17 = 11 + 18$.} Let’s do this: Since $(a, b) \sim (a', b')$ and $(c, d) \sim (c', d')$, we know that $a + b' = b + a'$ and $c + d' = d + c'$, respectively. Adding gives $(a + c) + (b' + d') = (b + d) + (a' + c')$, so we conclude $(a + c, b + d) \sim (a' + c', b' + d')$.

Having defined the sum of any two integers, we simply \textit{define} subtraction of integers by the rule $a - b = a + (-b)$, which makes sense regardless of the signs of $a$ and $b$. (Let’s agree to do this for rational numbers and real numbers too.)

We now move on to \textit{addition of rational numbers}, from a rigorous point of view. In terms of our equivalence classes of fractions, we define
\begin{align*}
[(a, b)] + [(c, d)] &= [(ad + bc, bd)].
\end{align*}
This definition is expected, since if we use the old-fashioned notation, we compute the sum as
\begin{align*}
\frac{a}{b} + \frac{c}{d} &= \frac{ad}{bd} + \frac{cd}{bd} = \frac{ad + bc}{bd}.
\end{align*}
though, of course, the formal definition conceals the intuitive motivation we discussed earlier (adding parts of a whole, or adding lengths)!

To be technically correct in our rigorous definition, we have to worry about a few things. To begin, why is \([ad + bc, bd]\) actually a rational number? We have to know that both entries are integers (no problem), but also that the second entry is nonzero. We haven’t talked about multiplication yet, but we can (and will) show that \(bd\) can’t be zero, since neither \(b\) nor \(d\) is zero (this should be a familiar principle to you).\(^4\)

Just as we did for the integers, we must check that the definition of rational addition is well-defined. You will do this in Exercise 23.

Since we regard real numbers as Cauchy sequences of rational numbers, we can use the definition of addition of rational numbers to get a definition of addition of real numbers. Specifically, if \((x_i)\) and \((y_i)\) are Cauchy sequences of real numbers, we define addition by term-by-term addition of the sequences:

\[
[(x_i)] + [(y_i)] = [(x_i + y_i)].
\]

Exactly as with rational numbers, we have to check a few things. First, why is \([(x_i + y_i)]\) a real number? (We would need to check that \((x_i + y_i)\) is a Cauchy sequence.) Second, why is the definition independent of the representives of the equivalence classes? These appear as Exercises 28 and 29.

1.3. Exercises.

Whole numbers

1. How might an elementary school teacher use a discrete model of addition to demonstrate that \(3 + 12 = 5 + 10\)? Why would the teacher be interested in conveying the fact \(3 + 12 = 5 + 10\) itself?

2. Suppose you are a cashier and need to select bills and coins totalling $47.89.

   (a) Which bills and coins would you be most likely to select?

   (b) What principle seems to govern this selection (i.e. why would you probably not select 47 singles and 89 pennies?)?

\(^4\)In terms of the traditional notation for fractions, we are just verifying that when one adds two fractions, the sum never has a denominator of zero.
(c) What subconscious algorithm did you use to produce the precise selection of bills and coins? Discuss carefully.

3.

(a) A customer pays for a $0.37 purchase with a dollar bill. Find all combinations of coins that could be used to make change for the purchase.

(b) Which of these combinations in (a) would you be most likely to choose, if you were the person giving change? What influences this choice?

(c) Describe several different strategies that one could use in making change (assuming the customer gives you a dollar and the purchase price is less than one dollar). On what mathematical ideas or algorithms do these strategies rely? Do these strategies ever result in giving a different combination of coins (for the same purchase price)?

4. On a drive through western Ohio, a textbook author stopped at a fast food restaurant and ordered two large iced teas. The purchase price was $2.64. The author gave three one-dollar bills to the cashier, who mistakenly selected two one-dollar bills, two quarters, a dime and four pennies from the cash drawer (totalling $2.64) as “change.” (The cashier mistakenly looked at the wrong line on the transcript for the Change Due.)

(a) What principle of paying-and-making-change makes the incident ridiculous (and should have alerted the cashier to her mistake)?

(b) Test your answer to (a) in the following situations. Is the transaction reasonable, or has the customer or cashier made an inappropriate payment? If the transaction is reasonable but “unusual,” what might the customer’s or cashier’s rationale be? In each example, the purchase price is $0.57.

(i) Customer pays three quarters and receives a dime, a nickel, and three pennies.
(ii) Customer pays four quarters and receives a quarter, a dime, a nickel, and three pennies.
(iii) Customer pays with a dollar bill and receives a quarter, a dime, a nickel, and three pennies.
(iv) Customer pays with a dollar bill, a nickel, and two pennies, and receives two quarters.
(v) Customer pays with a dollar bill and receives forty three pennies.
(v) Customer pays with a dollar bill, a nickel, and two pennies, and receives a quarter, two dimes, and a nickel.

5. Let the sets $C$, $D$, $G$ be as in Exercise 1 of Chapter 6, Section 5.
   
   (a) What is the whole number associated to $C$?
   
   (b) What is the whole number associated to $C \cup D$?
   
   (c) What is the whole number associated to $C \cup G$?

6. Let $a$ and $b$ be whole numbers with $a \geq b$. How would one use sets to describe $a - b$?

7. Let $a$ be a whole number. What statement about sets explains the fact that $a + 0 = a$?

8. Suppose $A$, $B$, $A'$, and $B'$ are sets, with $A \cap B = \emptyset = A' \cap B'$, $A \sim A'$, and $B \sim B'$. Prove that $(A \cup B) \sim (A' \cup B')$.

   **Integers, and issues involving negative numbers**

9. Why is the sum of two negative numbers negative?

10. Explain why it is reasonable to call $|a - b|$ the distance between the real numbers $a$ and $b$.

11. Sometimes students are taught the following rule for adding numbers of opposite sign: *If two numbers have opposite sign, subtract their absolute values, and give this the sign of the number whose absolute value is larger.*

   (a) Illustrate this rule with a few well-chosen examples.

   (b) Do you believe that this rule helps students understand the meaning of addition and subtraction? Why or why not?

12. Let $x, y \in \mathbb{Z}$. Is it always true that $|x + y| = |x| + |y|$? Explain.

13. In Figure 6, an algorithm for computing $623 - 59$ is presented.

   (a) Explain how this algorithm works.
(b) What advantage might it have over the standard algorithm?

14. Prove that $|x + y| \leq |x| + |y|$ for all real numbers $x$ and $y$.

15. In this exercise, we consider the definition of the integers that was given in Chapter 6, Section 6.

(a) What is the reasonable way to define $[(a, b)] + [(c, d)]$? Justify your answer.

(b) Using your answer to (a), prove that addition of integers is well-defined.

*Rational numbers*

16. Describe several real-world situations where one would have to consider subtraction of fractions, such as $\frac{2}{3} - \frac{1}{4}$.

17. Use a linear model to illustrate and solve the addition problem $\frac{3}{5} + \frac{7}{2}$. (Draw your diagram carefully, and in a few sentences, give an explanation of what you have done. Don’t forget to give the numerical solution.)

18. Repeat Exercise 17, using an area model.
19. Let’s consider the addition problem $3/20 + 14/15$ (note that both addends are fractions in simplest form). If we simply apply the definition, we have that

$$\frac{3}{20} + \frac{14}{15} = \frac{3 \cdot 15 + 14 \cdot 20}{20 \cdot 15} = \frac{45 + 280}{300} = \frac{325}{300}.$$ 

Here $\frac{325}{300}$ is a correct answer, but we have a long ways to go if we insist on an answer in simplest form, since the greatest common divisor of 325 and 300 turns out to be 25. To get an answer in simplest form, we would have to continue our calculation, perhaps in this way:

$$\frac{325}{300} = \frac{325}{300/5} = \frac{65}{60} = \frac{65/5}{60/5} = \frac{13}{12}.$$ 

Most of us learned a way to add fractions that gives an answer that is “closer” to being in simplest form. Namely, we use the least common multiple of the denominators, instead of their product. In the problem we have discussed, it’s easy to see that the least common multiple of 20 and 15 is 60. Writing out all the details, we might compute

$$\frac{3}{20} + \frac{14}{15} = \frac{3 \cdot 3 + 14 \cdot 4}{60} = \frac{9 + 56}{60} = \frac{65}{60}.$$ 

This time, though our answer is not in simplest form, it is “closer” to simplest form than $\frac{325}{300}$ was, and we only have to divide numerator and denominator by 5 to obtain $\frac{65}{60} = \frac{13}{12}$.

(a) Illustrate both techniques for adding fractions, with the addition problem $\frac{13}{35} + \frac{33}{70}$. Be certain to carry through your computations until you can give an answer in simplest form.

To review our methods for addition fractions: our first method for adding the fractions $\frac{a}{b}$ and $\frac{c}{d}$ was simply the addition formula

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}.$$ 

Our second method really amounted to the following:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad}{\gcd(b,d)} + \frac{bc}{\gcd(b,d)} = \frac{ad}{\gcd(b,d)} + \frac{bc}{\gcd(b,d)} \cdot \frac{\gcd(b,d)}{\text{lcm}(b,d)}.$$ 

(b) Substitute $a = 3$, $b = 20$, $c = 14$, $d = 15$ into the last formula.

(c) Find the easiest example you can to show that the fraction $\frac{ad + bc}{bd}$ need not be in simplest form, even if $\frac{a}{b}$ and $\frac{c}{d}$ are in simplest form.
(d) Suppose that \( \gcd(b, d) = 1, \gcd(a, b) = 1, \) and \( \gcd(c, d) = 1. \) Prove that the fraction \( \frac{ad + bc}{bd} \) must be in simplest form.

20. Perform the following computations involving rational numbers.

(i) \([(-5, 7)] + [(2, 5)]\)

(ii) \([-5, 7)] + [(0, 1)]\).

21. Use Definition 6 from Chapter 6, Section 7 to verify that \([a, b] + [c, b] = [(a + c, b)]\).

22. For rational numbers \([a, b]\) and \([c, d]\), suppose we propose that \([a, b] \oplus [c, d] = [(a + c, b + d)]\).

(a) Using this proposed definition of addition, compute one-half \(\oplus\) three-fourths in two different ways.

(Choose two different representatives for the equivalence classes of \(\frac{1}{2}\) and \(\frac{3}{4}\)).

(b) Show that the two results you obtained in part (a) are not equal, and conclude that the proposed definition of addition is not well-defined.

23. Show that addition of rational numbers is well defined.

Real numbers

24. Use a linear model to explain why if \(a < b\) and \(c \leq d\), then \(a + c < b + d\).

25. Prove that the sum of a rational number and an irrational number is irrational. (Hint: assume the sum can be written as a fraction, and look for a contradiction.)

26. Find two irrational numbers whose sum is irrational. Find two irrational numbers whose sum is rational.

27. In the text, we have discussed what the sum of two real numbers means. We can define the sum of three terms by the rule \(x + y + z = (x + y) + z;\) using this, we can define the sum of four terms by the rule \(w + x + y + z = (w + x + y) + z,\) etc. Why can’t we define the sum of infinitely many terms by this strategy?

28. Show that if \((x_i)\) and \((y_i)\) are Cauchy sequences of rational numbers, then so is \((x_i + y_i)\).
29. Prove that if \((x_i)\) and \((x'_i)\) are equivalent Cauchy sequences of rational numbers, and \((y_i)\) and \((y'_i)\) are equivalent Cauchy sequences of rational numbers, then \((x_i + y_i)\) and \((x'_i + y'_i)\) are equivalent Cauchy sequences of rational numbers.

30. If we define real numbers as Dedekind cuts, then how should the sum of two real numbers be defined?

2. Multiplication in Number Systems

Let’s Go 3. What explanations can you give for the fact that \(2 \times 3 = 6\)? What explanations can you give for the fact that \(\frac{1}{2} \times \frac{1}{3} = \frac{1}{6}\)? Upon what meaning of multiplication do your explanations rely?

Let’s Go 4. How many five-character license plates, consisting of a list of either five digits (0 through 9) or five letters (A through Z), are possible?

2.1. Whole number multiplication. In this section, we continue our study of operations in number systems by reviewing multiplication, beginning with whole numbers.

Our first experience with multiplication likely involved repeated addition. For example, we might think of \(4 \times 3\) as 3 added to itself four times, or \(3 + 3 + 3 + 3\). Since the definition of multiplication goes back to addition, we can use our favorite models of addition to model multiplication. For example, we can illustrate “\(4 \times 3 = 12\)” by a discrete model (joining four non-overlapping collections of three objects each) or by a linear model (joining four line segments, each of length three), as shown in Figure 7. The “repeated addition by threes” connects with the multiplication tables we all learned in elementary school. For example, the row corresponding to multiplication by three can be viewed as “counting by threes.” (See Figure 8.)

One of the most useful ways to model multiplication comes from stacking the addends in a repeated addition to form an array (naturally, we will call this the array model of multiplication). For example, again consider the product 4 \(\times 3\) (see Figure 9). Allowing the dots in the array model to become squares of size 1
unit by 1 unit, we arrive at area model\(^6\) of multiplication (see Figure 10).

Modeling multiplication using Cartesian products is not very different from the array model. Given sets \(A\) and \(B\), we define the Cartesian product of \(A\) and \(B\), denoted \(A \times B\), to be the collection of ordered pairs

\(^6\)For whole numbers, the array model and the area model seem virtually identical, but the area model has the advantage of generalizing readily to the real numbers.
where the first slot is filled by elements of $A$, and the second by elements of $B$:

$$A \times B = \{(a, b) : a \in A \text{ and } b \in B\}.$$  

For example, if $A = \{a, b, c, d\}$ and $B = \{x, y, z\}$, then

$$A \times B = \{(a, x), (a, y), (a, z), (b, x), (b, y), (b, z), (c, x), (c, y), (c, z), (d, x), (d, y), (d, z)\},$$

which may naturally be written as an array:

$$(a, x) \quad (a, y) \quad (a, z)$$

$$(b, x) \quad (b, y) \quad (b, z)$$

$$(c, x) \quad (c, y) \quad (c, z)$$

$$(d, x) \quad (d, y) \quad (d, z)$$

which looks suspiciously like the array model for $4 \times 3$. The punchline is that if $A$ and $B$ are finite sets with $a$ and $b$ elements, respectively, then the whole number $ab$ equals the number of elements in the set $A \times B$.

Finally, the intuitive idea of multiplication as repeated addition is the basis for a rigorous recursive definition of multiplication. Using the definition of the whole numbers via the Peano axioms (see Chapter 6, Definition 2), for whole numbers $n, m$ we define

\begin{equation}
(4) \quad n \cdot 0 = 0 \text{ and } n \cdot S(m) = n + (n \cdot m).
\end{equation}

As with the rigorous definition of whole number addition, we fully expect multiplication to be commutative and associative, and to distribute over addition (i.e., $a \cdot (b + c) = a \cdot b + a \cdot c$), but these properties must be established (see Section 5, Exercises 7 through 10).

2.2. Positive rational numbers. To multiply rational numbers we multiply numerators and denominators of corresponding fractions: $\frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}$. But what does this mean?

The word “of” helps motivate this definition. As an example, how would an elementary school student make sense of “half of a box of twelve cookies”? She likely would divide the cookies into two equal piles, and count that each pile contains six cookies. This explains why “half of twelve is six,” and is consistent with the proposed rule for multiplying fractions:

$$\frac{1}{2} \times \frac{12}{1} = \frac{1 \times 12}{2 \times 1} = \frac{12}{2} = 6.$$  

It’s conceptually harder to work with examples where neither factor is a whole number. Let’s look at several models for the problem “What is three sevenths of two fifths (of something)?” Intuitively, we would
have a picture of a unit. The unit would be divided into five equal pieces, and two of them would be retained. The resulting object would be divided into seven equal pieces, and three of them would be retained. The resulting object is \( \frac{2}{5} \) of \( \frac{2}{7} \) of the unit. To justify our rule for multiplying fractions, we need to understand why this is \( \frac{6}{35} \) of the original object. We will do just that, with linear and area models.

In the linear model, we begin with a segment of length 1, and then it’s easy to color in a segment of length \( \frac{2}{5} \) (see Figure 11). We now take this segment of length \( \frac{2}{5} \), divide it into seven equal segments, and take three of the pieces. Here is where our friend subdivision returns to visit us. We realize that it would be helpful to divide each segment of length \( \frac{1}{5} \) into seven equal pieces, which amounts to subdividing the original unit into \( 5 \times 7 = 35 \) pieces (see Figure 12). At this point, we can take 3 out of 7 of the tiny segments between 0 and \( \frac{1}{5} \), and 3 out of 7 of the tiny segments between \( \frac{1}{5} \) and \( \frac{2}{5} \). We have \( 2 \times 3 = 6 \) parts of the unit, which had been divided into \( 5 \times 7 = 35 \) parts (see Figure 13). Thus, three sevenths of two fifths of a segment equals six thirty-fifths of the segment, and we’ve seen the role that multiplying the numerators, and multiplying the denominators, plays in this calculation.
Let’s investigate the same calculation using an area model. Beginning with a unit square, we indicate two fifths of the unit by vertical dividers, and shade in two of the five strips to indicate “two fifths of the unit” (see Figure 14). The region representing two fifths is now supposed to be divided into seven equal pieces, which we can do conveniently with horizontal dividers (see Figure 15). We have darkly shaded three out of seven pieces of the lightly shaded region. To finish our work, we realize that that the the darkly
shaded region contains $3 \times 2 = 6$ small squares, out of a total of $5 \times 7 = 35$ parts into which the original square was divided.

2.3. Multiplication with negative numbers. We have all memorized “The product of two positives is positive. The product of two negatives is positive. The product of two numbers of opposite sign is negative.” How can we justify these principles? We will limit our examples to integers, but these arguments could also be extended to negative rational and real numbers.\(^6\)

We have already explained why a positive integer times a positive integer should be a positive, since multiplication was defined in terms of repeated addition. This definition also explains why a positive times a negative should be negative. For example, $3 \times -4$ should be shorthand for the repeated addition $-4 + -4 + -4 = -12$.

It is a lot tougher to argue why $-4 \times 3$ should be $-12$. Most importantly, we should realize that it’s not another example of repeated addition, since on the face of it, “adding 3 to itself $-4$ times” has no meaning. We need to regard the statement that “a negative times a positive is negative” as a definition . . . but one that can be motivated. Here are some heuristic explanations:

\(^6\)The Greek mathematician Diophantus of Alexandria (2nd or 3rd century CE) was the first to perform arithmetic with negative numbers, and devised the multiplicative rule of signs preserving the distributive property: $(+)(+) = (+), (+)(-) = (-), (-)(+) = (-),$ and $(-)(-) = (+)$.\)
(i) We don’t want order to matter in multiplication (we want $a \cdot b = b \cdot a$ for all numbers), so $-4 \times 3$ should equal $3 \times -4$, which is $-12$ by repeated addition.

(ii) As a special case, consider $-1$ times a number. We decree that $-1 \times a$ should be $-a$, the opposite of $a$ (this is plausible just by the notation involved). Hence multiplication by a negative should involve a change of signs.

(iii) $-4 \times 3$ should be the opposite of $4 \times 3$, hence it should be $-12$. Note that this argument is based on wanting $(-a)b = -(ab)$ to hold.

Note that all of these arguments are rooted in the insistence that a particular algebraic property hold. They don’t appear to come from any real-world model.

Other justifications rely on “mathematical patterns.” For example, in the sequence of facts $0 \times 4 = 0$, $1 \times 4 = 4$, $2 \times 4 = 8$, $3 \times 4 = 12\ldots$, we see the outputs increasing by four at each step. Aesthetically, it is then reasonable to continue the pattern backwards, making $(-1) \times 4 = -4$, etc.

**Your Turn 3.** *What is the most convincing explanation you can give for the product of two negative numbers being a positive number? Does your explanation appeal to some desired mathematical axiom, to a real-world model, or to something else? (You might wish to consult textbooks for ‘mathematics for elementary school teachers.’)*

### 2.4. Exercises.

**Whole numbers**

1. Illustrate the following products using both the array model and the area model.
   
   (a) $5 \times 4$.
   
   (b) $4 \times 5$.
   
   (c) $1 \times 4$.

2. Recalling that there are ten digits (0 through 9) and twenty six letters (A through Z), invent story problems where the correct answers are the following:

   (a) $10 + 26$
2. MULTIPLICATION IN NUMBER SYSTEMS

(b) \(10 \times 26\)

(c) \((10 + 26) \times (10 + 26)\)

(d) \((10 \times 26) + (26 \times 26)\)

(You might use Let’s Go 4 for inspiration.)

3. What does a triple product \(a \times b \times c\) of whole numbers mean? (It counts the number of elements in what sort of “array”?)

4. Why do we use \(\times\) for products when we first learn about multiplication, and later switch to juxtaposition when we study algebra?

5. We have described multiplication as a sort of shorthand for repeated addition. In turn, how can one represent repeated multiplication concisely?

6. What set-theoretic fact (involving Cartesian products) helps explain why \(0 \times b = 0\) for all whole numbers \(b\)?

7. What set-theoretic fact (involving Cartesian products) helps explain why if \(a\) and \(b\) are nonzero whole numbers, then \(a \times b\) is nonzero?

8. What set-theoretic fact (involving Cartesian products) helps explain why \(1 \times b = b\) for all whole numbers \(b\)?

9. We have modeled multiplication of whole numbers using the Cartesian products of sets. Is it necessary that the two sets in the product be disjoint? Explain.

10. Suppose that \(A, B, A',\) and \(B'\) are sets, with \(A \sim A'\) and \(B \sim B'\). Prove that \(A \times B \sim A' \times B'\).

Integers, and issues involving negative numbers

11. Prove that \(|xy| = |x||y|\) for all real numbers \(x\) and \(y\).

12. Assume that \(y > z\).
(a) If \( x > 0 \), argue why \( xy > xz \). What models and properties are you using?

(b) Similarly, if \( x < 0 \), argue why \( xy < xz \).

13. In this exercise,\(^7\) we outline a model for multiplication. It resembles the area model and helps account for “signs.” Working in the \( xy \)-plane, we consider the solid rectangle with corners \((0, 0)\), \((0, b)\), \((a, 0)\), and \((a, b)\).

(a) Draw several examples, with a variety of signs (±) for \( a \) and \( b \).

(b) How is the product \( ab \) related to the area of the rectangle?

(c) In which quadrants should we consider area to be positive, and in which quadrants should we consider area to be negative?

14. In this exercise, we consider the rigorous definition of the integers given in Chapter 6, Section 6.

(a) What is the reasonable way to define \([ (a, b) ] \cdot [ (c, d) ] \)?

(b) Using your answer to (a), prove that multiplication of integers is well-defined.

**Rational numbers**

15. Use both the linear model and the area model to illustrate and compute the following products. Be sure to indicate the final result.

(a) \( \frac{1}{3} \times \frac{5}{6} \).

(b) \( \frac{5}{6} \times \frac{1}{3} \).

(c) \( \frac{7}{6} \times \frac{2}{5} \).

(d) \( \frac{7}{6} \times \frac{3}{2} \).

16. Find two irrational numbers whose product is rational. Find two irrational numbers whose product is irrational.

17. What can you say about the product of a rational number and an irrational number? Formulate a precise statement and prove it.

\(^7\) Readers familiar with manipulatives may wish to discuss how this Exercise is related to the use of ‘algebra tiles.’
18. We have defined multiplication of rational numbers by the rule \([\frac{a}{b} \cdot \frac{c}{d}] = \frac{ac}{bd}\). Prove that multiplication is well-defined.

19. Recall that any integer \(a\) can be thought of as a rational number \([\frac{a}{1}]\) (our fancy way of writing the fraction \(\frac{a}{1}\)). This raises an interesting question: *If we add (or multiply) two integers, does it make a difference whether we think about them as integers (and use the rules for adding or multiplying integers), or whether instead we think of them as rational numbers (and use the rules for adding or multiplying rational numbers)?* Let’s check that it doesn’t:

Let \(a, b \in \mathbb{Z}\). Show that

(a) \([\frac{a}{1}] + [\frac{b}{1}] = [\frac{a+b}{1}]\).

(b) \([\frac{a}{1}] \cdot [\frac{b}{1}] = [\frac{ab}{1}]\).

How do these formulas answer our original question?

**Real numbers**

20. Let \(x\) and \(y\) be numbers with \(x < y\). Let \(t\) be a real number between 0 and 1. What can you say about the quantity \(t x + (1 - t)y\); is it less than \(x\), between \(x\) and \(y\), or greater than \(y\)? What explanations can you give for your conclusion?

21. In the text, we did not give a definition of the product of two real numbers!

(a) Use an intuitive model to explain how to multiply real numbers. (Make certain that your explanation works for both positive and negative numbers.)

(b) How would you use Cantor’s model (using Cauchy sequences) to define the product of two real numbers?

(c) How would you use Dedekind’s model (using Dedekind cuts) to define the product of two real numbers?

**3. Division in Number Systems**

**Let’s Go 5.** Invent several real world problems whose solution would require computing \(15 \div 3\). Which of these problems can be adapted to model \(15 \div \frac{3}{2}\) instead?
The purpose of this section is to make sense of the notion of \( a \) divided by \( b \), where \( a \) and \( b \) are numbers. This involves a circle of ideas, often reflected in the range of notation we use for division: \( 21 \div 3 \) versus \( \frac{21}{3} \) versus \( 21 \cdot 3^{-1} \). The circle of ideas includes the formal and algebraic perspective on division, along with the applications and contexts in which we first learned about the topic (like “dividing 21 cookies among 3 people” or “distributing 21 cookies, 3 to each person”). In this section, we will start with the more formal facts about division, and move backwards in time toward the models which we learned first.

Along the way, it will be instructive to compare the relationship between addition and subtraction with the relationship between multiplication and division.

### 3.1. A definition of division.

Let’s recall one of our interpretations of subtraction: \( a - b \) was the unique solution to the equation \( x + b = a \). Likewise, we shall declare \( a \div b \) to be the solution (assuming it exists and is unique!) to the equation \( bx = a \). In words: \( a \div b \) should represent the (unique) number which, when multiplied by \( b \), equals \( a \). Of course, we can’t make \( a \div b \) exist and be unique simply by wishing; we must investigate.

For uniqueness, suppose that \( x_1 \) and \( x_2 \) are both solutions to \( bx = a \). Then \( bx_1 = a = bx_2 \), so \( b(x_1 - x_2) = 0 \). When \( b \neq 0 \) we have \( x_1 = x_2 \), so if a solution exists it must be unique. Your Turn 4 asks us to explore solutions to \( bx = a \) when \( b = 0 \):

**Your Turn 4.** Show that it is not the case that \( 0x = a \) has exactly one solution. Does your argument depend on the value of \( a \)?

Your solution to Your Turn 4 should convince you that \( a \div b \) does not exist if \( b = 0 \) so we will assume \( b \neq 0 \). Further, we already know that there is at most one solution to \( bx = a \) when \( b \neq 0 \), so all that remains is to establish the existence of a solution.

Let’s start in the case where \( a \) and \( b \) are integers (and \( b \neq 0 \)). In this case the fraction \( \frac{a}{b} \) is a solution to \( bx = a \), due to the definition of multiplication of rational numbers:

\[
\frac{b \cdot a}{b} = \frac{b \cdot a}{b} = \frac{ba}{b} = \frac{b \cdot a}{b} = 1 \cdot a = a.
\]

Having found a solution to \( bx = a \) when \( a \) and \( b \) are integers (always, with \( b \neq 0 \)), let’s do the same when \( a \) and \( b \) are rational.
Your Turn 5. Write \( a = \frac{c}{d} \) and \( b = \frac{e}{f} \).

(a) Show that \( de \neq 0 \). (Why is this important?)

(b) Show that \( \frac{cf}{de} \) is a solution of \( \frac{ex}{f} = \frac{c}{d} \).

(c) What procedure do you customarily use for dividing fractions? Is it consistent with (b)?

(d) Suppose that \( \frac{c}{d} \) and \( \frac{e}{f} \) are replaced by equivalent fractions. Show that the same rational number is obtained as the solution to \( bx = a \).

(e) Why is the solution to \( bx = a \) unique? (You might review the whole number case.)

Thus, we have obtained a unique solution to \( bx = a \) when \( b \) and \( a \) are rational and \( b \neq 0 \). Hence \( \frac{a}{b} \) is defined if \( a \) and \( b \) are rational numbers.

Finally, since we can approximate real numbers by rational numbers, it is not too hard to confirm that \( bx = a \) has a solution for any real numbers \( a \) and \( b \) (\( b \neq 0 \)). Thus, we have considered \( a \div b \) for all nonzero \( b \). We now turn to some other ways of understanding \( a \div b \).

3.2. Division and multiplicative inverses. Let's recall one of our views of subtraction: “\( a \) minus \( b \) equals \( a \) plus the additive inverse of \( b \),” or in symbols, \( a - b = a + \overline{b} \). A very similar thing happens for division: \( a \) divided by \( b \) equals \( a \) times the multiplicative inverse of \( b \). We must clarify some ideas and terms here. Recall that zero was an additive identity, since it doesn’t change any quantity when added to it (\( c + 0 = c \) for all \( c \)). We can also say that \( \text{one is a multiplicative identity} \), since it doesn’t change any quantity when multiplied by it (\( c \cdot 1 = c \) for all \( c \)). Now on to inverses: the numbers \( b \) and \( \overline{b} \) are additive inverses of each other since their sum is the additive identity. Likewise, we say that \( \text{numbers } b \text{ and } b' \text{ are multiplicative inverses of each other if their product is the multiplicative identity } 1 \). For example, \( \frac{5}{3} \) and \( \frac{3}{5} \) are multiplicative inverses of each other.

So, let’s examine our statement, “\( a \div b \) equals \( a \) times the multiplicative inverse of \( b \).” Why does this work?

Your Turn 6. Suppose \( b' \) is the multiplicative inverse of \( b \). Show that \( a \cdot b' \) is a solution of \( bx = a \).

What properties or definitions are needed in your argument?
Your Turn 7. Using Equation (5) for motivation, show that \( \frac{x}{2} = \frac{1}{2} \cdot x \). (This is a common source of difficulty for algebra students.)

Having boldly discussed multiplicative inverses, we should check whether they actually exist! In Exercise 10, you will verify that 0 doesn’t have a multiplicative inverse (this supports our argument that \( a \div 0 \) never exists). Likewise, you will give an argument that every real number has a unique multiplicative inverse.

3.3. Models of division. Let’s pretend that we are back in the elementary grades, and are trying to understand \( 21 \div 3 \). Part of our understanding at that point should have been computational: to compute \( 21 \div 3 \), we needed to find a number whose product with 3 is 21. By our familiarity with multiplication tables, we probably could have seen that \( 21 \div 3 = 7 \). Indeed, our array or area models help us understand the division problem: \( 21 \div 3 \) is the measure of the unknown side, such that the array or rectangle with other side 3 gives a total measure or area of 21 (see Figure 16).

Figure 16. Division and the array model

Your Turn 8. Figure 16 looks suspiciously like a figure one might draw to illustrate a multiplication fact, namely, \( 3 \times 7 = 21 \). How would the use or construction of the figure be different (depending on whether one were illustrating a multiplication or division problem)?
But more than as a solution to an abstract problem, “division” is a method which is appropriate for many real-world problems. We discuss these now.

First, we consider the *partitive model of division*. In this model, \(21 \div 3\) is the solution to the problem, “If 21 objects are divided (or ‘partitioned’) into 3 equal piles, how many objects are in each pile?” Even more concretely, we have the following model for \(21 \div 3\): a dealer has a deck of 21 cards, and is dealing them to 3 players; then each player receives \(21 \div 3\) cards. In terms of equations, if we think of \(x\) as the number of objects in each pile, we are solving \(x + x + x = 21\); again we remember that multiplication is repeated addition, we are back to the equation \(3x = 21\), the very equation whose solution is \(21 \div 3\).

However, there is another way to deal cards, and another way to interpret \(3x = 21\) as repeated addition. Imagine a dealer with a deck of 21 cards, attempting to serve a long line of players. In the game, each player is to receive 3 cards. Then in this second model, \(21 \div 3\) represents the number of players who will receive cards. More generally, “\(21 \div 3\) is the number of times 3 must be added to itself \(^8\) to obtain 21.” In terms of repeated addition, in this interpretation, the equation \(3x = 21\) becomes \(3 + 3 + 3 + \ldots = 21\).

This second model is called the *measurement model of division*. To see why, imagine a wall, whose length is in fact 21 feet, but which we are trying to measure with a yardstick (1 yard = 3 feet). Imagine laying down the yardstick, end over end, from one end of the wall to the other. Then, the number of times that we must position the yardstick is a measurement of the length of the wall (in yards). How many times must we do this? The answer is the number of times we must add 3 to itself to obtain 21.

Now \(21 \div 3\) is one of the simpler division problems we encountered in elementary school, since it “comes out even,” meaning that \(21 \div 3\) is an integer. How do we make sense of a problem like \(23 \div 3\)?

- In the partitive model, we picture a dealer attempting to deal 23 cards to 3 people. Of course, each person gets 7 cards, but there are 2 cards left over, which is problematic.

- If we instead imagine the dealer distributing doughnuts, we see that each person gets 7 doughnuts; then the remaining two doughnuts are cut, and each person gets an additional \(\frac{2}{3}\) doughnut. Thus \(23 \div 3 = 7 + \frac{2}{3}\), commonly written as \(7\frac{2}{3}\), a mixed number. Of course, as an improper fraction, we simply obtain \(\frac{23}{3}\), reinforcing the notion that “fractions represent division.”

\(^8\)A minor variant of this interpretation involves repeated subtraction: \(21 \div 3\) is the number of times 3 must be subtracted from 21 to obtain 0.
In the measurement model, again we fail if we view it as dealing cards: the best we can say is that 7 people get a full hand of 3 cards, with an eight person getting $\frac{2}{3}$ of a full hand.

However, in terms of measuring a 23-foot wall with a yardstick, the measurement model gives a clear picture of a wall that is $7\frac{2}{3}$ yards long (see Figure 17).

---

**Figure 17.** The measurement model

**Your Turn 9.** Consider $\frac{7}{2} \div \frac{5}{4}$. The partitive model is difficult to apply in this example (what would it mean to divide something among $\frac{5}{4}$ persons?), but the measurement model is straightforward:

(a) What “measuring stick” problem inspires the division problem $\frac{7}{2} \div \frac{5}{4}$?

(b) Draw a diagram that helps solve the “measuring stick” problem. (Do you need subdivision?)

You will be asked a slightly more difficult problem in the exercises: modeling $7/4 \div 2/3$.

**3.4. Exercises.**

1. Create two word problems that illustrate $36 \div 12$, one using the partative model and one using the measurement model.

2. Following the example of $\frac{7}{2} \div \frac{5}{4}$ in the text, use the measurement model to illustrate $\frac{7}{4} \div \frac{2}{3}$.

3. One might illustrate $60 \div 12 = 5$ by saying “60 is five dozen.” Is this an example of the partitive model or the measurement model? Explain.
4. A patient has 48 pills and is supposed to take a pill every eight hours. How many days will the patient be taking pills? Give an explanation that a middle school student could understand.

5. In the text, we interpreted \( x = 21 \div 3 \) by repeated addition in two different ways: \( x + x + x = 21 \) and \( \frac{3 + 3 + 3 + \ldots}{x} = 21 \). How well do these two approaches work in understanding \( x = 23 \div 3 \)? Discuss.

6. How might an elementary school student interpret and solve a division problem like \( 15 \div 3 \)? How would a high school student simplify an expression like \( \frac{x^2 - 4}{x+3} \div \frac{x - 2}{4x+12} \)? Compare and contrast their approaches.

7. In algebra, juxtaposition usually indicates multiplication: for example, \( 3c \) means 3 times \( c \). In a mixed number like \( 7\frac{5}{8} \), what operation is implied between 7 and \( \frac{5}{8} \)?

8. Consider the word problem, *A box contains a large number of doughnuts. Sam takes three-fourths of them, and this amounts to 24 doughnuts. How many doughnuts were originally in the box?*

   (a) Give a strategy for solving the problem that would make sense to students who know some algebra.

   (b) How would you explain how to solve the problem to younger students, who don’t know algebra?

      Give a detailed answer. (You may want to draw figures too.)

9. Let \( b \) be a nonzero real number. Give a clear argument explaining why \( 1 \div b \) is the multiplicative inverse of \( b \).

10. Give an explanation of why 0 does not have a multiplicative inverse. (Don’t simply say that we can’t divide by zero. Instead, assume that 0 has a multiplicative inverse, and derive a contradiction.)

11. (a) Show that every nonzero rational number has a multiplicative inverse that is rational. (Don’t forget to check that your answer is well-defined.)

    (b) Sketch the graph of the function \( f : \mathbb{Q} \setminus \{0\} \to \mathbb{Q} \setminus \{0\} \) that takes every nonzero rational number to the multiplicative inverse you found in (a).

    (c) Give a heuristic argument, using (b), to show that each nonzero real number has a multiplicative inverse.
(d) What does your graph in (b) suggest about the existence of a multiplicative inverse of zero?

12. Let $b$ be a nonzero real number. Show that $b$ can’t have two different multiplicative inverses. (Hint: suppose that $b'$ and $b''$ are both multiplicative inverses of $b$. Consider $b' bb''$.)

Exercises 13–18 are concerned with the Division Algorithm:

**Theorem 1.** (Division Algorithm) Let $a, b \in \mathbb{Z}$ with $b \neq 0$. There exist unique integers $q$ and $r$ satisfying $a = qb + r$ and $0 \leq r < |b|$. The number $q$ is called the quotient, while $r$ is called the remainder.

As the name suggests, the Division Algorithm asserts that every division problem of the form $a \div b$ in the integers (with $b \neq 0$) has a unique solution, provided that we are willing to accept both an integer quotient $q$ and an integer remainder $r$, subject to the conditions laid out in the theorem. Concretely, if $a$ and $b$ are positive, the Division Algorithm expresses the notion that $a$ objects can be pictured as $q$ groups of $b$ objects with $r$ “left over.” This is how elementary school students may regard division before they have fractions or decimals at their disposal. It turns out that the Division Algorithm is also important in college-level mathematics, since it is a handy tool for proving some basic theoretical results.

13. Let $a = 96$ and $b = 17$. Find the unique integers $q, r$ satisfying the requirements of the Division Algorithm.

14. Again let $a = 96$ and $b = 17$. Find three pairs of integers $q, r$ such that $a = qb + r$. Does the fact that you can find three such pairs violate the Division Algorithm?

15. Let $a$ and $b$ be whole numbers. Explain how the Division Algorithm is useful in converting the fraction $\frac{a}{b}$ to a mixed number.

16. Given $a$ and $b$, find the unique $q$ and $r$ satisfying the requirements of the Division Algorithm. (This exercise shows that the quotient and remainder in the Division Algorithm can be slightly different from those in “long division” when negative numbers are considered.)

   (a) $a = 26, b = 6$
4. Algebraic Properties in Number Systems: An Intuitive View

Let’s Go 6. Suppose a student in an algebra class you are teaching believes that \((a+b)^2 = a^2 + b^2\). What mathematical arguments could you give to convince him that he is mistaken? Which of these arguments help give the correct expansion for \((a+b)^2\)?

The associative property, the commutative property, the distributive property... We may have trouble remembering which is which. This problem can be addressed by thinking carefully about what these words mean (see Exercise 5). Beyond keeping their names straight, other issues deserve our close attention: why these laws hold, and why it’s important to know the laws and their justifications. In this section, we are not interested in rigorous proofs involving successor sets, equivalence classes of ordered pairs, Cauchy sequences of rational numbers, and the various tools necessary for the rigorous definition of number systems. Instead, we are looking for arguments, rooted in everyday experience and the intuitive models of the number systems, that can be understood by middle and high school students, and that have the power to convince and to illuminate.

---

9 We will consider some of these notions in Section 5.
Let’s consider some approaches to justifying the commutative property of multiplication. A teacher might try the following strategy: *We know that* $5 \times 8 = 40$ *and* $8 \times 5 = 40$. *That’s an example of* $ab = ba$.

How well does this explanation succeed? It does give an example, but a single example can be a fluke, and it does not go very far in convincing, let alone proving. Suppose the teacher instead said, *We know from our multiplication table that* $2 \times 3 = 6$, *and* $3 \times 2 = 6$ *too. We see that* $3 \times 8 = 24$, *and* $8 \times 3 = 24$ *too. In fact, we see that the multiplication table is symmetric. The commutative property “*ab = ba*” is exactly the statement that the multiplication table is symmetric.* This is a far better explanation. It is more convincing, since it gives a host of examples, not just a single one. Moreover, it ties the commutative property to the multiplication table, a tangible, familiar object for students. Still, one might fault the explanation, since it doesn’t illuminate why the commutative property holds through a direct appeal to the meaning of multiplication.

This can be remedied by considering an array model. We know that $3 \times 2$ means $2+2+2$. Let’s represent this in a diagram with two dots in the first row, two in the second row, and two in the third row. (See Figure 18.) On the other hand, we know that $2 \times 3$ means $3 + 3$. We can represent this in a figure with three dots in the first row and three dots in the second row. Now we can we flip (or rotate) the first diagram to obtain the second one. We know that the number of dots doesn’t change when we flip. Therefore $2 \times 3 = 3 \times 2$.  

![Figure 18. An array model illustration of the commutative property](image-url)
The same argument works for any array, so $ab = ba$ for any positive integers. Observe that the array model, along with physical intuition, can be used to give a very convincing explanation of the commutative property of multiplication for positive integers. In Exercise 1, you will be asked to do something similar: to use the area model to justify properties involving positive real numbers.

Finally, students need to know the properties of addition and multiplication because they are tantamount to the rules of algebra. For example, the commutative and associative properties are exactly what is necessary to perform the simplification $(x^2y^5)(x^7y^3) = x^9y^8$ (try it!). In the exercises you will analyze some other algebraic arguments, to identify precisely which properties are used in each step.

4.1. **Properties of addition and multiplication in the real numbers.** Here are the properties that you will be considering in the Exercises, and which you should commit to memory and understanding:

**Theorem 2. Additive properties of the real numbers:**

(i) Closure property of the real numbers under addition: *If $x$ and $y$ are real numbers, then $x + y$ is a real number.*

(ii) Commutative property of addition in the real numbers: *If $x$ and $y$ are real numbers, then $x + y = y + x.$*

(iii) Associative property of addition in the real numbers: *If $x$, $y$, and $z$ are real numbers, then $(x + y) + z = x + (y + z).$*

**Theorem 3. Multiplicative properties of the real numbers:**

(i) Closure property of the real numbers under multiplication: *If $x$ and $y$ are real numbers, then $xy$ is a real number.*

(ii) Commutative property of multiplication in the real numbers: *If $x$ and $y$ are real numbers, then $xy = yx.$*

(iii) Associative property of multiplication in the real numbers: *If $x$, $y$, and $z$ are real numbers, then $(xy)z = x(yz).$*
Theorem 4. Distributive properties in the real numbers: If \( x, y, \) and \( z \) are real numbers, then 
\[(x + y)z = xz + yz \text{ and } z(x + y) = zx + zy.\]

Theorem 5. The role of 1: The number 1 is a multiplicative identity: if \( x \) is a real number, then \( 1 \cdot x = x \) and \( x \cdot 1 = x \). The number 1 is unique in the sense that there are no other multiplicative identities in \( \mathbb{R} \) other than 1.

Theorem 6. The role of 0:

(i) The number 0 is an additive identity: if \( x \) is a real number, then \( 0 + x = x \) and \( x + 0 = x \). The number 0 is unique in the sense that there are no other additive identities in \( \mathbb{R} \) other than 0.

(ii) If \( x \) is a real number, then \( 0 \cdot x = 0 \) and \( x \cdot 0 = 0 \).

(iii) If \( x \) and \( y \) are real numbers and if \( xy = 0 \), then either \( x = 0 \) or \( y = 0 \).

Theorem 7. Additive inverses: If \( x \) is a real number, then there exists a unique real number \( y \) such that \( x + y = 0 \). The number \( y \) is called the additive inverse of \( x \) and is usually written as \( -x \) or \( -x \).

Theorem 8. Multiplicative inverses: If \( x \) is a nonzero real number, then there exists a unique real number \( y \) such that \( xy = 1 \). The number \( y \) is called the multiplicative inverse of \( y \) and is written as \( \frac{1}{x} \), \( 1/x \), or \( x^{-1} \).

Theorem 9. Properties of additive inverses:

(i) For all real numbers \( a \), \( -(a) = a \).

(ii) For all real numbers \( a \) and \( b \), \( -(a + b) = (-a) + (-b) \).

(iii) For all real numbers \( a \) and \( b \), \( (-a)b = -(ab) \).

(iv) For all real numbers \( a \), \( -a = (-1) \cdot a \).

Theorem 10. Properties of multiplicative inverses:

(i) For all nonzero real numbers \( a \), \( (a^{-1})^{-1} = a \).

(ii) For all nonzero real numbers \( a \) and \( b \), \( (ab)^{-1} = a^{-1}b^{-1} \).

It is thought that the Babylonians’ (ca 1700 BCE) knowledge of the distributive property allowed them to factor a difference of two squares, which in turn led to solutions of quadratic equations in generality.
4.2. Exercises. In the following, the most critical exercise is Exercise 1, since it challenges us to make the properties of addition and multiplication believable, even obvious, for our students.

1. For the positive real numbers, let’s use the “addition of lengths” model for addition and the “area” model for multiplication. Using these models, give convincing arguments in support of the commutative and associative properties (for both addition and multiplication); the distributive property; and the roles of 1 and 0. Identify, as clearly as possible, the properties of lengths or areas that you are using in making your arguments.\(^{11}\)

2. A student attempts to justify the associative property of addition by saying, “\((x+y) + z\) has to equal \(x + (y + z)\) since they are both equal to \(x + y + z\).” What is the main reason that this argument is inadequate?

3. Give at least one example of a typical problem or calculation in high school algebra in which “Property (iii) for the role of 0” is used.

4. Complete the sentence: Dividing a quantity by \(x\) is the same as multiplying the quantity by ______.

5. Look up the common meanings of the words “commute” and “associate” in a dictionary. Explain why it is plausible to use these words to name the “commutative properties” and “associative properties.” Can you think of better words than “commutative” or “associative” to name these properties?

6. Suppose you need to add up the following scores in a game: 23, 8, 14, 12, 7. Add them up in your head. What strategy did you employ? Which properties of addition did you use?

7. How could you justify that \((x + y) + z = (y + z) + x\), if \(x, y,\) and \(z\) are real numbers? (Produce a string of equalities, and justify each step using one of the properties of addition.)

8. Read through all the theorems in this section, with “real number” replaced by “integer.” Which of the statements become false? Try this again with “positive integer,” and with “rational number.”

\(^{11}\)What physical model might you use when you are considering triple products like \((xy)z\) in the associative property for multiplication?
9. Carefully solve the equation $3x + 8 = 14$. In each step, explain what action you are taking, and which algebraic property justifies your work.

10. Repeat the previous exercise, using the equation $3x + 8 = 13$.

11. Accepting that zero is an additive identity, show that there is no other additive identity. (Suppose that $a$ is an additive identity, and consider $0 + a$.)

12. Show that no other number except 1 is a multiplicative identity.

13. Let $x$ be a nonzero real number and let $y$ denote its multiplicative inverse. Is it possible that $y = 0$? Explain.

14. Let $x$ be a nonzero real number. Explain why

   The multiplicative inverse of the multiplicative inverse of $x$ is $x$ itself.

Then, write the statement in symbols, first with reciprocals, then with exponents. (Which theorem are we discussing in this exercise?)

15. A teacher discusses $9 + 8 = 17$ in the following way:

   $9 + 8 = 9 + (1 + 7) = (9 + 1) + 7 = 17$.

   (a) Which arithmetic property is featured in this calculation?
   (b) What idea appears to motivates the teacher’s discussion?

16. Let $a, b, c, d$ be whole numbers, and recall the “FOIL” method of multiplying binomials: $(a + b) \times (c + d) = ac + ad + bc + bd$.

   (a) Use the area model of multiplication to illustrate $(a + b) \times (c + d) = ac + ad + bc + bd$.
   (b) Use the distributive property to prove that $(a + b) \times (c + d) = ac + ad + bc + bd$.

17. (a) Use algebra to expand $(a + b + c)^2$. 

(b) Draw an area model that helps illustrate your answer to (a). Explain briefly.

(c) Use algebra to expand \((a + b)^3\), draw a model that illustrates this calculation, and explain briefly.

18. Suppose \(x, y, z\) are whole numbers. Expand the product \(((x + y) + z)(xy + z)\). At each step, state which property of arithmetic is being used.

19. Solve the equation \(x^2 - 14x + 45 = 0\) in the whole numbers by factoring. Show your work, and point out the place where you use properties of zero.

20. Create numerical examples that show that subtraction is neither commutative nor associative.

21. Create numerical examples that show that division is neither commutative nor associative.

5. Algebraic Properties in Number Systems: A Survey of Some Rigorous Arguments

With a great deal of effort, we could prove all of the theorems stated in the last section. That would be a course in itself: a long compendium of projects, some engaging, some merely tedious. In this section we will single out some of the most interesting issues. We begin with remarks on strategy:

On the one hand, let’s recall that integers are defined using whole numbers, rational numbers are defined using integers, and real numbers are defined using rational numbers. Hence, in large measure, one needs to prove the properties first for the whole numbers, then the integers, then the rational numbers, then the real numbers. Proofs of the properties often depend on the number system (\(\mathbb{Z}, \mathbb{Q}, \mathbb{R}\)) being investigated, and a given property may be much harder to prove in one number system than in another. Below, we will give examples involving each of these number systems.

On the other hand, there are some general principles that come through across these number systems. For example, why is the additive inverse of a number unique? Suppose that \(a'\) and \(a''\) are both additive inverses of \(a\). By the associative property, we have that

\[
a'' = 0 + a'' = (a' + a) + a'' = a' + (a + a'') = a' + 0 = a'.
\]

Hence \(a' = a''\). Note that this argument is “universal”: it applies in any context where a binary operation is associative and there is an identity element. These sorts of universal statements/arguments, which are very important in higher mathematics, are often the main focus of an abstract algebra course.
5.1. Whole numbers. Recall that in Chapter 6, Definition 2 we defined the whole numbers in terms of the Peano axioms, and this definition in turn inspired definitions of addition and multiplication of whole numbers given in Equations (1) and (4), respectively. Properties of these operations are proved almost entirely using the principle of induction given in Chapter 6, Definition 2 (see Exercises 3 through 13).

For example, to verify that whole number addition is associative we use the principle of induction on \( n \) to show \( a + (b + n) = (a + b) + n \) for given whole numbers \( a, b \). According to the definition of addition given in (1), we have

\[
a + (b + 0) = a + b = (a + b) + 0,
\]

so associativity holds when \( n = 0 \). Now suppose the result holds for \( n = k \in \mathbb{W} \). Using the induction hypothesis and (1) we obtain

\[
a + (b + S(k)) = a + S(b + k) = S(a + (b + k)) = S((a + b) + k) = (a + b) + S(k),
\]

so associativity also holds for \( n = S(k) \). We conclude by the principle of induction that \( a + (b + n) = (a + b) + n \) for all \( n \in \mathbb{W} \).

5.2. Integers. Once we accept the properties of the whole numbers, it’s straightforward (albeit tedious) to verify the properties for the integers, since one does not need to go back to the deep mysteries of set theory. For example, let’s prove that addition of integers is commutative. Let \([a, b]\) and \([c, d]\) be integers (equivalence classes of pairs of whole numbers, as in Chapter 6, Section 6.1). On the one hand, \([a, b] + [c, d] = [(a + c, b + d)]\), and on the other hand, \([c, d] + [a, b] = [(c + a, d + b)]\). Thus, commutativity of integers under addition follows from commutativity of whole numbers under addition.

5.3. Rational numbers. Now accepting the properties for integers, it’s another step to prove the properties for rational numbers. As an example, we will prove the associative property for multiplication. We will represent the three rational numbers as equivalence classes of ordered pairs: \( x = [(a, b)] \), \( y = [(c, d)] \), and \( z = [(e, f)] \). On the one hand,

\[
(xy)z = \left( [(a, b)] \cdot [(c, d)] \right) \cdot [(e, f)] = [(ac, bd)] \cdot [(e, f)] = [(ac)e, (bd)f],
\]
and on the other hand,

\[ x(yz) = [(a, b)] \cdot \left( ([c, d]) \cdot ([e, f]) \right) = [(a, b)] \cdot ([ce, df]) = [(a(ce), b(df))]. \]

These are equal, since \((ac)e = a(ce)\) and \((bd)f = b(df)\), by the associative property of multiplication of integers. So in this example, we see that the property for rational numbers follows quite quickly from the property for integers.

As a slightly harder example, let’s prove the distributive property. Write \(x = [(a, b)], y = [(c, d)], \) and \(z = [(e, f)] \). On the one hand,

\[ x(y + z) = [(a, b)] \cdot \left( ([c, d]) + ([e, f]) \right) = [(a, b)] \cdot ([cf + de, df]) = [a(cf + de), b(df)], \]

and on the other hand,

\[ xy + xz = [(a, b)] \cdot ([c, d]) + [(a, b)] \cdot ([e, f]) = [(ad + bc, bd)] + [(af + be, bf)] = [(ad + bc)(bf) + (bd)(af + be), (bd)(bf)]. \]

To verify that these are equal, we must show that

\[ (a(cf + de)) \cdot ((bd)(bf)) = (b(df)) \cdot ((ad + bc)(bf) + (bd)(af + be)). \]

This final equation is certainly true, but verifying it requires use of numerous algebraic properties of the integers.

5.4. Real numbers. By expressing real numbers as (equivalence classes of) Cauchy sequences of rational numbers, we can prove the properties for real numbers.

As an easy example, let’s check the commutative property of multiplication. We consider real numbers \(x \) and \(y\), and think of \(x\) as the equivalence class \([(x_i)]\) and \(y\) as the class \([(y_i)]\), where each \(x_i\) and each \(y_i\) is rational. Then \(x + y\) is the equivalence class of the sequence \((x_i + y_i)\) and \(y + x\) is the class of the sequence \((y_i + x_i)\). Since \(x_i + y_i = y_i + x_i\) for all \(i\) (a fact we are assuming about rational numbers), we conclude that \(x + y = y + x\).

As a harder example, let’s show that if \(x \neq 0\) and \(y \neq 0\), then \(xy \neq 0\). We can write \(x = [(x_i)]\), where \((x_i)\) is a Cauchy sequence of rational numbers. Since \(x \neq 0\), we know that there is a positive rational number\(^{12}\) \(X\) such that \(|x_i| \geq X\) for all \(i\) greater than or equal to some \(N(i)\), and likewise, there is a positive rational

---

\(^{12}\)If no such \(X\) exists then \(\lim_{i \to \infty} x_i = 0\), contradicting the fact that \(x\) is a nonzero real number.
number $Y$ such that $|y_i| \geq Y$ for all $i$ greater than or equal to some $M(i)$. Then, for all $i$ greater than or equal to the maximum of $N(i)$ and $M(i)$, we have that $|x_iy_i| \geq XY$. Hence $xy \neq 0$.

5.5. Exercises.

1. Show that multiplicative inverses are unique. (Hint: review the argument that additive inverses are unique.)

Whole numbers

2. Translate each of the following statements about whole numbers to statements about sets.

   (a) If $x$ and $y$ are whole numbers, then $x + y = y + x$.

   (b) If $x$, $y$, and $z$ are whole numbers, then $(xy)z = x(yz)$.

   (c) If $x$, $y$, and $z$ are whole numbers, then $x(y + z) = xy + xz$.

Exercises 3 through 13 explore the recursive definitions of whole number addition and multiplication based on the Peano axioms (see Equations (1) and (4) together with Chapter 6, Definition 2).

3. For each whole number $n \in \mathbb{W}$, show that either $n = 0$ or $n = S(k)$ for some whole number $k$. (Use the principle of induction given in Chapter 6, Definition 2.)

4. Prove that $S(n) = n + 1$ for all whole numbers $n$. (Use the principle of induction is given in Chapter 6, Definition 2.)

5. Prove that $k + S(n) = S(k) + n$ for all whole numbers $k,n$. (Hint: Again, use the principle of induction on $n$.)

6. Prove that addition is commutative: Use induction on $n$ to show $k + n = n + k$ for all whole numbers $n,k$. (Hint: Use the principle of induction on $n$ together with Exercise 5.)

7. Prove that $n \cdot 0 = 0 \cdot n = 0$ for all whole numbers $n$. (Use the principle of induction given in Chapter 6, Definition 2.)
8. Prove that $n \cdot 1 = 1 \cdot n = n$ for all $n \in \mathbb{W}$. (Recall that $1 = S(0)$, and use the principle of induction is given in Chapter 6, Definition 2.)

9. For given whole numbers $a$ and $b$, show that
   i. $a \cdot (b + n) = a \cdot b + a \cdot n$, and
   ii. $(a + b) \cdot n = a \cdot n + b \cdot n$
   for all whole numbers $n$. As always, use induction on $n$.

10. Prove that multiplication of whole numbers is commutative. (Hint: By Exercise 7 you already know $a \cdot 0 = 0 \cdot a$ for all $a \in \mathbb{W}$. It then suffices to show, for a given $m \in \mathbb{W}$, that $n \cdot S(m) = S(m) \cdot n$ for all $n \in \mathbb{W}$. You guessed it–use induction on $n$.)

11. For a given whole number $a$, show that $a$ is comparable to $n$ for all $n \in \mathbb{W}$. (Note: $a$ is comparable to $n$ means either $a < n$, $a = n$, or $a > n$, where $<$ is defined in Equation (2). Do your proof–all your proofs, and everyone else’s too–by induction on $n$.)

12. Show that the order on whole numbers given in Equation (2) is transitive, that is, show that $a < b$ and $b < c$ implies $a < c$.

13. Show that for a given $a \in \mathbb{W}$, exactly one of $a < n$, $a = n$, and $n < a$ holds for each $n \in \mathbb{W}$. (Hint: As always, do this by induction on $n$.)

**Integers**

14. Using the alternate definition of integers, prove the distributive property (for integers).

**Rational numbers**

15. Prove that addition of rational numbers is associative. (You may use properties of integers in your proof.)

**Real numbers**
16. Again, regard a real number as an equivalence class of Cauchy sequences of rational numbers. Prove that every nonzero real number has a multiplicative inverse.

6. Complex Numbers

Many of us greet the complex numbers with skepticism and resistance. Having spent several years believing that the real numbers really are all the numbers, it is easy to regard complex numbers as an entirely artificial construction. We may be told that $i = \sqrt{-1}$, but we certainly don’t see $-1$ having a square root in our graph of the squaring function (see Figure 19). Part of our objection is on linguistic grounds:

aren’t numbers that are “imaginary” actually fake, so how can they be of any use in the real world? Finally, by the time we reach high school, our notion of number is so closely tied to the number line and to length, that it is hard to see the usefulness of a system like the complex numbers, since complex numbers don’t occur along a line, and they don’t appear to be measuring the length of anything.

In this section, we will consider several uses of complex numbers that justify studying them. Before we get to that, we will recall their definition and basic properties.

---

13The complex numbers were first developed by Rafael Bombelli (1526-1572) as an outgrowth of his attempts to understand the method of solving cubic equations given by Girolamo Cardano (1501-1576) in his *Ars Magna*. For more on this, see Chapter 12.
6.1. The complex numbers: set and operations. Just as we modeled the real numbers on a line, we model the complex numbers on the plane. We all know that to describe a point in the plane, we need an ordered pair of real numbers. Thus, we could define a complex number to be any point \((a, b)\) in the \(xy\)-plane. This is not the usual notation for complex numbers, and to make the transition to the usual notation, we need to recall some operations with vectors. Recall that

\[(a, b) = (a, 0) + (0, b) = a(1, 0) + b(0, 1).\]

We have an easy way of interpreting this geometrically. The vectors \((1, 0)\) and \((0, 1)\) are the unit vectors (vectors of length one) along the positive \(x\)- and \(y\)-axes, respectively. The vectors \((a, 0)\) and \((0, b)\) are vectors along the \(x\)- and \(y\)-axes, respectively, and are obtained by “stretching” \((1, 0)\) and \((0, 1)\) by the lengths \(a\) and \(b\), respectively.\(^{14}\) Finally, the vector \((a, b)\) is formed by adding the vectors \((a, 0)\) and \((0, b)\), completing the rectangle with vertices \((0, 0)\), \((a, 0)\), and \((0, b)\). (See Figure 20.)

\[\text{Figure 20. } (a, b) \text{ and } a + bi\]

To obtain the usual description of complex numbers, we just give new names to the unit vectors \((1, 0)\) and \((0, 1)\). We call \((1, 0)\) simply “1” and we give \((0, 1)\) the new name “\(i\).” Completing our earlier calculation, we have

\[(a, b) = (a, 0) + (0, b) = a(1, 0) + b(0, 1) = a1 + bi.\]

\(^{14}\)If \(a\) or \(b\) is negative, a flip is involved too.
We then abbreviate $a_1$ simply as $a$, and we obtain our usual description “$a + bi$” for a complex number.\footnote{It may seem incorrect to replace an ordered pair $(1, 0)$ with a single real number 1. We justify this by imagining the real numbers themselves as being embedded in the plane as the $x$-axis. Hence the point 1 on the number line really does get identified with the ordered pair $(1, 0)$ on the $x$-axis.}

We adopt this description in the definition below, but never forget that complex numbers can always be pictured as points in the plane!

**Definition 11.** A complex number is any expression of the form $a + bi$, where $a$ and $b$ are real numbers. The set of all complex numbers is denoted $\mathbb{C}$.

For the complex number $z = a + bi$, if $b = 0$, we think of $z$ as simply a real number. If $a = 0$, we call $z$ an imaginary number. We abbreviate as much as possible, so $4 + 0i$ is usually written as 4, $0 + 7i$ is usually written as $7i$, $6 + 1i$ is usually written as $6 + i$, etc. We also allow ourselves to write $a + ib$ instead of $a + bi$.

We now turn to operations. Addition of complex numbers is defined in order to correspond to vector addition in the plane (see Figure 21). Specifically, we add complex numbers ‘component-wise’:

\[(a + bi) + (c + di) = (a + c) + (b + d)i\]
Multiplication is more unexpected and more interesting. Almost certainly, in a linear algebra course, you did not define the product of two vectors. Indeed, for vectors in \( \mathbb{R}^n \), one generally can’t make any reasonable definition of vector multiplication.\(^{16}\) Nevertheless, something special happens when \( n = 2 \).

The definition we all learn is

\[
(a + bi)(c + di) = (ac - bd) + (ad + bc)i
\]

**Your Turn** 10. Why should we believe that this definition of complex multiplication is a good one? Let’s look at some consequences.\(^{17}\)

(a) First, consider what happens when both complex numbers are actually real numbers (meaning, \( b = 0 = d \)). Compute \((ac - bd) + (ad + bc)i\). (We interpret this as the definition of multiplication for complex numbers extends the definition of multiplication of real numbers.)

(b) Second, what is the product of a real number and a complex number? (Take \( b = 0 \) and compute \((ac - bd) + (ad + bc)i\).) How would you interpret this geometrically?

(c) Compute the product of \( i \) with itself. (This explains why \( i \) is sometimes written as \( \sqrt{-1} \).)

(d) What is the product of \( i \) with another complex number? Show that \((0+1i)(c+di) = -d+ci\). Observe that one obtains \(-d+ci\) by applying a 90° rotation to \( c + di \) (see Exercise 10), so multiplication by \( i \) produces rotation by 90°.

Below, we will generalize (d) to reveal the geometric interpretation of complex multiplication.

### 6.2. Algebraic properties

Another justification for our definitions of complex addition and multiplication is that the usual algebraic properties of the real numbers (the associative property, the distributive property, etc.) are valid for the complex numbers! These properties are listed in Proposition 13. The proofs of these properties are straightforward, and are wonderful exercises since they force us to revisit the analogous properties of the real numbers.

\(^{16}\)You may have studied the “inner product” or “dot product” of vectors in \( \mathbb{R}^n \), but the inner product of two vectors is a real number, not a vector. You may also have learned about the cross product of vectors in \( \mathbb{R}^3 \), which is a sort of multiplication that does not obey the commutative property. There is an interesting multiplication in \( \mathbb{R}^4 \), called the quaternionic multiplication, which you may study in an abstract algebra course.

\(^{17}\)Bombelli, likely the first European mathematician to examine the work of Diophantus closely, produced a rule of signs for operating with \( \pm \sqrt{-1} \) similar to the one Diophantus produced for \( \pm 1 \). Thus complex arithmetic was born.
Before we get to Proposition 13, we will recall a few important notions involving complex numbers. Let $z = a + bi$ be a complex number. We call $a$ the real part of $z$ and write $a = \Re(z)$; likewise $b$ is the imaginary part of $z$, written $\Im(z)$. (If this fancy calligraphy is too difficult to reproduce by hand, feel free to write $\text{Re}(z)$ and $\text{Im}(z)$ instead.) The complex conjugate of $z$, denoted $\overline{z}$, is the complex number $a - bi$. Geometrically, the complex conjugate of $z$ is obtained by reflecting $z$ across the real axis. Finally, we denote the distance from $z$ to the origin by $|z|$. The Pythagorean Theorem assures us that $|z| = \sqrt{a^2 + b^2}$. The quantity $|z|$ is known by several names: the absolute value of $z$, the magnitude of $z$, the length of $z$, the norm of $z$, and the modulus of $z$.

The following important properties should be investigated and committed to memory:

**Lemma 12.** Let $z = a + bi$ and $w = c + di$ be complex numbers.

(i) $z + \overline{z} = 2\Re(z)$, and $z - \overline{z} = 2i\Im(z)$.

(ii) $\overline{z \overline{w}} = z \overline{w}$.

(iii) $\overline{z + w} = \overline{z} + \overline{w}$.

(iv) $|z|^2 = z\overline{z}$.

(v) $|z| = |\overline{z}|$.

(vi) $|z|$ is a nonnegative real number, and $|z| = 0$ if and only if $z = 0$.

(vii) $|z||w| = |zw|$.

(viii) $|z + w| \leq |z| + |w|$ (the triangle inequality)

**Proof.** We prove part (i), leaving the remaining parts as exercises. (The most interesting proofs are for items (iii), (vii), and (viii).)

Let $z = a + bi$, with $a, b \in \mathbb{R}$. Then

$$z + \overline{z} = (a + bi) + (a - bi) = 2a = 2\Re(z),$$

while

$$z - \overline{z} = (a + bi) - (a - bi) = 2bi = 2i\Im(z).$$

\qed
We are now ready to present the basic properties of addition and multiplication in \( \mathbb{C} \).

**Proposition 13.** Let \( z, w, \) and \( v \) be complex numbers.

(i) \( z + w = w + z \) and \( zw = wz \). (Addition and multiplication are commutative in \( \mathbb{C} \).)

(ii) \( (z + w) + v = z + (w + v) \) and \( z(wv) = (zw)v \). (Addition and multiplication are associative in \( \mathbb{C} \).)

(iii) \( z(w + v) = zw + zv \). (The distributive property holds in \( \mathbb{C} \).)

(iv) \( z + 0 = 0 + z = z \). (0 = 0 + 0i is an additive identity in \( \mathbb{C} \).)

(v) Suppose \( z = a + bi \) with \( a, b \in \mathbb{R} \). If we put \( -z = -a + (-b)i \), then \( z + (-z) = 0 \). (Each element of \( \mathbb{C} \) has an additive inverse.)

(vi) \( z \cdot 1 = 1 \cdot z = z \). (1 = 1 + 0i is a multiplicative identity in \( \mathbb{C} \).)

(vii) Suppose \( z \neq 0 \). There is an element \( z^{-1} \) in \( \mathbb{C} \) such that \( z z^{-1} = z^{-1} z = 1 \). (Each nonzero element of \( \mathbb{C} \) has a multiplicative inverse.)

(viii) The product of 0 and any complex number is 0.

(ix) If \( zw = 0 \), then either \( z = 0 \) or \( w = 0 \).

**Proof.** Most of these properties follow from the corresponding properties for \( \mathbb{R} \) (and we will omit most of the proofs). As an example, we will prove commutativity of addition. If \( z = a + bi \) and \( w = c + di \) with \( a, b, c, d \in \mathbb{R} \), then the commutativity of addition in \( \mathbb{R} \) gives

\[
z + w = (a + c) + (b + d)i = (c + a) + (d + b)i = w + z.\]

Thus addition is commutative in \( \mathbb{C} \).

For the rest of the proof we focus on item (vii). Since \( z \neq 0 \), by Lemma 12 we know that \( z \overline{z} \) is a positive real number, so \( \frac{1}{z} \overline{z} \) is also a positive real number (it’s the multiplicative inverse of \( z \overline{z} \)). Observe

\[
z \cdot \left( \frac{1}{z \overline{z}} \right) = z \cdot \left( \overline{z} \cdot \frac{1}{z \overline{z}} \right) = (z \cdot \overline{z}) \cdot \frac{1}{z \overline{z}} = 1,
\]

so \( z \) has a multiplicative inverse \( z^{-1} = \frac{1}{z \overline{z}} \).

\[\square\]

6.3. Polar decomposition. It’s crucial that we have a good geometric ways of understanding complex numbers themselves. We already have one such way: if \( z = a + bi \), then \( a \) and \( bi \) are the orthogonal
projections of $z$ onto the real and imaginary axes, respectively. In this section, we examine an alternate
geometric description that is closely tied to the notion of polar coordinates in the plane.

Let $z$ be a nonzero complex number. We already understand $|z|$, the length of $z$. The other important
ingredient we need is the argument of $z$ by which we mean the angle $\theta$ that $z$ makes with the positive real
axis, with the restriction that $-\pi \leq \theta \leq \pi$. For example, in Figure 22, we see a complex number $z$ whose
length is 2 and whose argument is $-\pi/4$, and a complex number $w$ whose length is 1.5 and whose argument
is $2\pi/3$.

**Figure 22. A geometric view of complex multiplication**

In geometric terms, a complex number $z$ is the unique point of intersection of

- A circle centered at the origin (consisting of all complex numbers having the same magnitude as $z$)
- A ray emanating from the origin (consisting of all complex numbers with the same argument as $z$)

Note that the point $\frac{z}{|z|}$ has magnitude one (why?) and lies on the same ray emanating from the origin
as $z$. If we let $\theta$ denote the argument of $z$, then by the definition of the trigonometric functions, we have
$\frac{z}{|z|} = \cos \theta + i \sin \theta$. 
Finally, we see that a nonzero complex number $z$ ‘factors’ as the product $|z| \cdot \frac{z}{|z|}$, where $|z|$ is a positive real number, and $\frac{z}{|z|}$ is a complex number of magnitude one. We refer to the factorizations

$$z = |z| \cdot \frac{z}{|z|} \quad \text{and} \quad z = |z| \cdot (\cos \theta + i \sin \theta)$$

as the polar decomposition of $z$. The complex numbers arising in the polar decomposition of $z$ are illustrated in Figure 23.

**Figure 23.** Polar decomposition of a nonzero complex number $z$

### 6.4. The geometric meaning of complex multiplication.

The polar decomposition of complex numbers is the key for understanding the geometry lying behind multiplication of complex numbers:

**Proposition 14.** Geometric interpretation of complex multiplication.\(^\dagger\) Let $z$ and $w$ be complex numbers.

(i) The length of $zw$ is the product of the lengths of $z$ and $w$.

(ii) The argument of $zw$ is obtained by adding the arguments of $z$ and $w$ and then replacing the result (if necessary) with a co-terminal angle between $-\pi$ and $\pi$.

\(^\dagger\)As you will verify in Exercise 20, the geometric interpretation of complex multiplication yields a quick proof of the identities for $\cos(\theta + \phi)$ and $\sin(\theta + \phi)$ that we learned in high school trigonometry. Most often, the mathematics is done in the reverse order: the sum formulas for cosine and sine and proved, and Proposition 14 is deduced from them. We take this approach in the chapter Trigonometry, where two proofs of the sum formulas are presented.
Proof. The tools in the proof are the fact that multiplying a complex number by a real number merely stretches the number, and multiplying a complex number by $i$ rotates the complex number counterclockwise by an angle of measure $\pi/2$. (See Your Turn 10.)

Suppose that $z$ and $w$ have arguments $\phi$ and $\theta$, respectively. We also will need to write $w$ as $a + bi$. (See Figure 24.) We add to Figure 24 in Figure 25:

![Figure 24. Geometric interpretation of multiplication, part 1](image)

- We mark the location of the complex numbers $az$ and $biz$. Note that $az$ lies on the same line through the origin as $z$, whereas $biz$ lies on the perpendicular line through the origin (why?).
- We mark the location of $(a + bi)z = wz$.
- Consequently, $\angle(biz, 0, az)$ is a right angle, and $\Delta(biz, 0, az)$ is right triangle.
- Moreover, since $|biz| = |z||bi|$ (why?) and $|az| = |z||a|$ (why?), we see that $\Delta(biz, 0, az)$ is similar to $\Delta(bi, 0, a)$, with the side lengths of latter triangle being stretched by a factor of $|z|$ to obtain the side lengths of the former triangle.

---

19 This argument is taken from *Mathematical Connections: A Companion for Teachers and Others*, Al Cuoco, Mathematical Association of America, 2005. Dr. Cuoco cites *Visual Complex Analysis*, Tristan Needham, Oxford University Press, 1998 as another location this idea has appeared in print, though he learned the idea from a group of teachers at the Park City Mathematics Institute who discovered it on their own.

20 If the real number is negative, then the complex number is stretched and its orientation is reversed.
Since $\Delta(biz,0,az)$ is congruent to $\Delta(wz,az,0)$, and $\Delta(bi,w,a)$ is congruent to $\Delta(w,a,0)$, we have that $\Delta(wz,az,0)$ is similar to $\Delta(w,a,0)$, with a similarity factor of $|z|$. Consequently:

(i) The length of $zw$ is $|z|$ times the length of $w$.

(ii) The measure of $\angle(wz,0,az)$ equals the measure of $\angle(w,0,a)$. Since angles $\angle(z,0,a)$ and $\angle(wz,0,az)$ sum to the angle $\angle(wz,0,a)$, part (ii) of the proposition follows.

Note that the first part of Proposition 14 is a restatement of Lemma 12(vii).

How does Proposition 14 help in finding products? Returning to Figure 22, by the Proposition, the length of $zw$ should be $2 \cdot 1.5 = 3$, and the argument of $zw$ should be $\pi/4 + 2\pi/3 = 11\pi/12$. Finally:

**Your Turn 11.** Show that we can locate $zw$ precisely, at the intersection of a particular circle centered at the origin and a particular ray. (What is the radius of the circle? What angle does the ray make with the positive real axis?) Then, plot this circle and ray in Figure 22, and mark $zw$.

There is also a geometric method for finding $z^{-1}$ given a nonzero complex number $z$ (see Exercise 12).
6.5. Why complex numbers? If you teach precalculus in high school, one of your challenges will be to persuade your students that complex numbers are worth learning about. Indeed, the complex numbers are ubiquitous in mathematics and its applications, but the fact that complex numbers are not used for counting, measurement, or other daily-life activities means that you will have to expend some effort to motivate them. Here is a sample of applications of complex numbers.

(a) When one thinks of the points in the plane as complex numbers, many computations become much easier. As an example, suppose we take a point like \((5, 7)\) in the plane and we want to rotate it around the origin by \(30^\circ\). This is not so easy to do! Using complex numbers, we realize that we need to “add an angle of \(30^\circ\) to \(5 + 7i\) but not change its length.” By Proposition 14, this means multiplying \(5 + 7i\) by a number of length 1 and argument \(\frac{\pi}{6}\), which is \(\frac{\sqrt{3}}{2} + \frac{1}{2}i\). Hence, the rotated point we seek is \((\frac{\sqrt{3}}{2} + \frac{1}{2}i)(5 + 7i) = \frac{5\sqrt{3} - 7}{2} + 5\frac{\sqrt{3}}{2}i\). Scientists and engineers routinely think of the plane as the complex numbers in order to simplify computations.

(b) Prior to learning about complex numbers, we tend to think that positive real numbers have two square roots, and negative real numbers have no square roots. However, once we know about imaginary numbers, we realize that negative real numbers also have two square roots; for example, the square roots of \(-1\) are \(\pm i\), the square roots of \(-4\) are \(\pm 2i\), and the square roots of \(-3\) are \(\pm i\sqrt{3}\). Of course, we can also view these roots as the roots of the polynomials \(x^2 + 1\), \(x^2 + 4\), and \(x^2 + 3\).

A bit more generally, we can see that any degree-two polynomial with real coefficients has two roots in \(\mathbb{C}\). How so? Starting with the quadratic polynomial \(ax^2 + bx + c\), the quadratic formula gives the two roots \(\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\). If \(b^2 - 4ac > 0\), we obtain two real roots. If \(b^2 - 4ac < 0\), we obtain two complex (nonreal) roots.

There is a far-reaching generalization of this phenomenon, known as the Fundamental Theorem of Algebra. This theorem states that any polynomial of degree \(n\), with complex coefficients, has exactly \(n\) roots in the complex numbers. More precisely, if the polynomial is \(a_nx^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_2x^2 + a_1x + a_0\), then there are exactly \(n\) complex numbers\(^{22}\) \(r_1, \ldots, r_n\) such that the polynomial factors as

\(^{21}\)If you have studied linear algebra, you may know how to do this using a rotation matrix, but a high school student might not know this.

\(^{22}\)There may be some repetitions among the roots.
6. COMPLEX NUMBERS

\(a_n(x - r_1)(x - r_2)\ldots(x - r_n)\). This is in marked contrast to what happens over the real numbers, where a polynomial of degree \(n\) might have fewer than \(n\) real roots. For this reason, in areas of higher mathematics where one needs to think about roots of polynomials, it’s considered easier to use complex numbers than real numbers. Even when one is interested in the real roots, one often starts by thinking about all the complex roots, and then using extra techniques to figure out which of them are actually real.

(c) In calculus, many of the functions that we study are given by convergent power series. For example, we know that \(e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k\), \(\cos x = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k}\), and \(\sin x = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1}\). It turns out that power series make sense when the variable is allowed to be a complex number. This means we can define functions from \(\mathbb{C}\) to \(\mathbb{C}\) using power series; an example would be the function \(e^z\) given by the power series above. Ideas like “differentiability” from calculus can be extended to the complex numbers as well. The behavior of “calculus over the complex numbers” has been studied extensively for several centuries. The ideas help to clarify connections among the most important functions in precalculus and calculus. We will discuss this in the chapter Transcendental Functions and Complex Numbers, but see Exercises 28 and 29 for a taste of things to come.

6.6. Exercises.

1. Compute the following sums and differences. Also, provide a sketch of the geometric interpretation of each computation.

(a) \((2 - 3i) + 5\)

(b) \((2 - 3i) + (3 + i)\)

(c) \((2 - 3i) - (3 + i)\)

(d) \((2 + 3i) + (2 - 3i)\)

(e) \((2 + 3i) - (2 - 3i)\)

(f) \((2 + 3i)(3 + i)\)

(g) \((2 + 5i)i\)

2. Let \(z = a + bi \neq 0\). Find an expression for \(z^{-1}\) in terms of \(a\) and \(b\).
3. Often one writes \(1/(a+bi)\) to denote \((a+bi)^{-1}\), and \((a+bi)/(c+di)\) to denote \((a+bi)(c+di)^{-1}\).

With this in mind, write the following expressions in the form \(a+bi\).

(a) \(\frac{1}{3+2i}\).
(b) \(\frac{\overline{z}}{2-i}\).
(c) \(\frac{3+2i}{2+i}\).
(d) \(\frac{1-i}{1+i}\).

4. True or false:

(a) The real part of a complex number is a real number.
(b) The imaginary part of a complex number is an imaginary number.
(c) \(z = \Re(z) + i\Im(z)\)
(d) \(|z| = \sqrt{\Re(z)^2 + \Im(z)^2}\)

5. How can you simplify the expression \(\overline{z}\)?

6. Find expressions for \(\Re(z)\) and \(\Im(z)\) in terms of \(z\) and \(\overline{z}\).

7. Find an expression for \(|z|\) in terms of \(z\) and \(\overline{z}\).

8. For which complex numbers \(z\) is \(\overline{z} = z\)? For which \(z\) is \(\overline{z} = -z\)? For which \(z\) is \(\overline{z} = 3z\)?

9. Find an expression for the distance between two complex numbers \(z\) and \(w\). Your answer should be in terms of \(z\), \(\overline{z}\), \(w\), and \(\overline{w}\).

10. Give persuasive evidence that when \(c+di\) is rotated counterclockwise by 90° around the origin, one obtains the complex number \(-d+ci\).

11. Suppose \(z \neq 0\). What is the relationship between the modulus of \(z\) and the modulus of \(z^{-1}\)?

12. Suppose \(z\) is a nonzero complex number.

(a) Let \(\theta\) be the argument of \(z\). What is the argument of \(z^{-1}\)?
(b) Let \( z \) be a nonzero complex number. Explain why \( z \) and \( z^{-1} \) lie on the same ray through the origin.

(c) Parts (a) and (b) indicate how to plot the location of \( z^{-1} \) in the complex plane given the location of \( z \). Plot \( z = 3 + 4i \) and then plot \( z^{-1} \) without computing \( z^{-1} \) explicitly.

13. Consider the complex number \( z \) sketched in Figure 26. Estimate the modulus and argument of \( z \). Then, plot the locations of \( z, z^2, z^{-1}, \) and \( iz \).

14. Consider the complex numbers \( z \) and \( w \) in Figure 27. Estimate the modulus and arguments of \( z \) and \( w \). Then, plot the locations of \( zw, z^2, w^2, z/w, \) and \( w/z \).

15. Let \( z \) be the complex number of modulus 3 and argument \( \frac{3\pi}{4} \).

(a) According to the geometric rule for multiplying complex numbers, what should the modulus and argument of \( z^2 \) be?

(b) Using your answer to (a), write \( z^2 \) in the form \( a + bi \).

(c) Using trigonometry, write \( z \) in the form \( a + bi \).

(d) Write \( z \) in the form \( a + bi \) and then compute \( z^2 \) by computing \( (a + bi)(a + bi) \). You should get the same answer as in (b).
16. If the point \((3, 7)\) is rotated counterclockwise by \(45^\circ\), what point is obtained? What point is obtained if \((3, 7)\) is rotated clockwise by \(45^\circ\) instead?

17. Write each part of Lemma 12 in words. (For example, (iii) could be written, “The product of complex conjugates is the complex conjugate of the product.”)

18. Find all complex numbers that are roots of the polynomial \(x^2 + 6x + 20\). Find all complex numbers that are roots of the polynomial \(x^2 + 6x - 20\). (Use the quadratic formula.)

19. Let \(ax^2 + bx + c\) be a degree-two polynomial with real coefficients. Using the quadratic formula, show that if the two roots are not real, then they are complex conjugates of each other.

20. Let \(z = \cos \theta + i \sin \theta\) and \(w = \cos \theta + i \sin \theta\).

(a) What are the magnitudes of \(z\) and \(w\)? Why?
(b) What are the arguments of \(z\) and \(w\)?
(c) According Proposition 14, what is the magnitude of \(zw\)?
(d) According to Proposition 14, what is the argument of \(zw\)?
(e) According to (c) and (d), what is the polar form of \(zw\)?
6. COMPLEX NUMBERS

(f) Using the definition of multiplication of complex numbers, compute the product \( zw \).

(g) Comparing your answers to (e) and (f), deduce the formulas for the cosine and sine of a sum.

21. Prove parts (ii) through (viii) of Lemma 12.


23. This exercise models an investigation into the geometric meaning of complex multiplication.

(a) Compute and plot the vectors in \( \mathbb{R}^2 \) associated to the complex numbers \( i, \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i, \) and \( 1 - i \).

(You may plot all of these on the same set of coordinate axes if you wish.)

(b) For each vector in part (a), find the length of the vector as well as the angle from the positive real axis to the vector.

(c) Plot vectors associated to the complex numbers \( 2, 1 + 2i, -2 + 2i, \) and \(-2\) on a single set of coordinate axes. Then, multiply each of these numbers by \( i \) and plot the products on another set of coordinate axes.

(d) Repeat the second step of part (c), but multiply by \( \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i \) and \( 1 - i \) instead of \( i \).

(e) Consider the plots you made in parts (c) and (d). Geometrically, what seems to be the effect of multiplying a given complex number by \( i \)? By \( \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i \)? By \( 1 - i \)?

24.

(a) Give an example of a degree-three polynomial with real coefficients that has three real roots.

(b) Give an example of a degree-three polynomial with real coefficients that has one real root and two nonreal roots.

(c) Is there a degree-three polynomial with real coefficients that has two real roots and one nonreal root? Explain.

25. By factoring \( x^4 - 1 \), find all the complex numbers that are fourth roots of 1. Do the same for the third roots of 1.

26. Factor \( x^8 - 1 \) into linear factors. (This is tantamount to finding all the complex numbers whose eighth powers equal 1.)
27. Let \( a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \) be a polynomial with real coefficients. Show that if the complex number \( r \) is a root, then \( \overline{r} \) is also a root. What properties of complex numbers do you use in your proof?

28. Using power series, show that \( e^{iz} = \cos(z) + i \sin(z) \).

29. 
   (a) Show that \( \frac{2}{z^2 + 1} = \frac{i}{z+1} - \frac{i}{z-1} \).
   
   (b) By integrating, use your answer to (a) to give an expression for \( \arctan z \) in terms of natural logarithms.
   
   (c) Confirm your answer to (b) with a computer algebra system.

30. A key feature of \( \mathbb{R} \) is that it is ordered. This means (roughly) that given two unequal elements of \( \mathbb{R} \), exactly one of the statements \( (a < b, b < a) \) is true, and some axioms hold (among them, \( \text{If } a > 0 \text{ and } b > 0, \text{ then } ab > 0 \)).

   Show that \( \mathbb{C} \) is not ordered. (Hint: show that the assumption \( i > 0 \) leads to a contradiction, and so does the assumption \( i < 0 \).)

31. Suppose we tried to invent a new number system, as follows. The number system consists of symbols \( a + bI \), where \( a \) and \( b \) are real numbers. Addition is defined by the rule \( (a + bI) + (c + dI) = (a + c) + (b + d)I \), and multiplication is defined by \( (a + bI)(c + dI) = (ac + bd) + (ad + bc)I \).

   (a) How do these operations differ from the operations in \( \mathbb{C} \)?
   
   (b) What is \( I^2 \)?
   
   (c) Show that Proposition 13 (vii) and (ix) fail in this number system.
CHAPTER 8

Topics in Number Systems

In Chapters 6 and 7, we emphasized the similarities of the number systems $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. We saw how each arises from an intuitive model, typically involving an activity of measurement. Each of these number systems allows for operations called addition and multiplication. There are many common properties that the operations in each of these number systems share (e.g., the distributive property). Common language (such as additive inverse and multiplicative identity) is used in the study of each system.

Despite these commonalities, there are important differences among the number systems. For example, consider the notion of factorization in the integers, the process where we might express 20 as $20 \times 1$, $10 \times 2$, or $5 \times 4$ (a product of integers). To deal with factoring larger integers, students require an understanding of primes and the uniqueness of prime factorization. By experience, students come to believe that any factorization (such as the three factorizations of 20 given above) can be prolonged (by factoring the factors) until the only factors are “primes;” moreover, the same primes occur when the process of factorization terminates, regardless of the preliminary steps in the factorization. Factorization, primes, and the Division Algorithm are discussed in Section 1.

Is factoring a useful notion in other number systems?

- In the rational numbers, we can factor using the ideas about factorization in the integers. For example, we could decompose $\frac{72}{250}$ as $\frac{2 \times 3 \times 3 \times 2}{5 \times 5 \times 5} = \frac{2^2 \times 3^2}{5^3}$.

- In the real numbers, factorization seems to be a useless project. What would it mean to factor $\pi$? We could write $\pi$ as $\pi \times 1$, as $\pi / 7 \times 7$, as $\pi \sqrt[3]{e} \times \sqrt[3]{e}$, or as $\frac{\pi}{7} \times r$ for any nonzero real number $r$. There is no compelling reason to prefer any of these factorizations, nor is there any notion of “prime” in the real numbers.
• Any nonzero complex number can be written uniquely as a positive real number times a complex
number of magnitude one.\textsuperscript{1} This factorization, the polar decomposition,\textsuperscript{2} is unrelated to factorization in the
integers, and there is no reasonable notion of prime in the complex numbers.

Thus, the behavior of the number systems tends to diverge, revealing unique and unexpected features. Conversely, differences that seem profound may turn out to be surmountable with the correct strategy. This is manifested in Sections 2 and 4, which discuss numeration systems—our ways of writing down numbers.

In Section 2, we overview some ancient numeration systems and analyze their merits and shortcomings. The base-ten system emerges as a system that enables us to represent large numbers compactly, while allowing for easy comparison of numbers (\textit{Which number is bigger?}) and lending itself to efficient algorithms for addition and multiplication. Remarkably, the strategy behind the base-ten notation of whole numbers (that of \textit{regrouping} using larger and larger powers of ten) can be turned on its head (into subdivision by powers of ten) to give the decimal system for representing real numbers (Section 4).

Choices of notational systems tend to make some problems easy and other problems hard. For example, when an integer is written as a base-ten numeral, it’s easy to determine whether the integer is a multiple of ten, but hard to determine (at a glance) whether it is a multiple of seven. In Section 3, we see how the classic divisibility tests from elementary school rely on features of the base-ten notational system.

We have focused on integers, rational numbers, real numbers, and complex numbers because of their crucial importance in the secondary school curriculum. A thorough investigation of other number systems is beyond the scope of this text, but in Section 5, we give a glimpse at systems of algebraic numbers, and use Liouville’s Theorem to produce an explicit example of a transcendental number.

The material in this chapter is accessible without reading Chapters 6 and 7.

1. Arithmetic in the Integers

Let’s Go 1.

\textsuperscript{1}The polar decomposition extends to a polar decomposition of matrices, as one may learn about in a linear algebra course or a Lie groups course.

\textsuperscript{2}See Section 6 of Chapter 7.
(a) Suppose $a$, $b$, and $c$ are positive integers. Suppose that $a$ is a divisor of the product $bc$. Can we conclude that $a$ must be a divisor of one of the factors $b,c$? Why or why not?

(b) Suppose $x$, $y$, and $z$ are positive integers. Suppose that $z$ is a multiple of $x$ and $z$ is a multiple of $y$. Can we conclude that $z$ is a multiple of the product $xy$? Why or why not?

1.1. Divisors and multiples. From elementary school, we are familiar with equations such as $4 \times 6 = 24$. We first see such equations as multiplication facts; later, perhaps written in the opposite order as $24 = 4 \times 6$, we recognize such equations as factorization exercises. We grow accustomed to terminology associated with these equations: we say that 24 is a multiple of both 4 and 6, and that 4 and 6 are divisors or factors of 24. In general:

**Definition 1.** Let $a,b \in \mathbb{Z}$. If there exists an integer $q$ such that $a = bq$, then we say that $b$ divides $a$, that $b$ is a factor or divisor of $a$, and that $a$ is a multiple of $b$.

It’s also customary to write $b|a$ as shorthand for the statement “$b$ divides $a$.” This is very convenient notation, though it sometimes causes confusion, since the vertical bar can be misinterpreted as a fraction symbol. We emphasize that $4|24$ is a statement, not a number: it’s the (true) statement that 4 divides 24, or in other words, 4 is a divisor of 24. Likewise $24|4$ is the (false) statement that 24 is a divisor of 4. Both of these are very different from the expressions $4/24$ and $24/4$. The latter aren’t “statements” at all—they are just the numbers $1/6$ and 6, respectively.

We also write $b \not| a$ as shorthand for “$b$ does not divide $a$.”

**Your Turn 1.** Which integers divide zero? Which integers does zero divide?

**Your Turn 2.**

(a) List the positive divisors of 30, in increasing order.

(b) How do the items in your list produce factorizations $30 = a \cdot b$ with $a \leq b$? Why is $a \leq \sqrt{30}$ and $b \geq \sqrt{30}$?

(c) How would this discussion change if we considered 36 instead of 30?

---

3If you have taken a course in abstract algebra and are getting a different answer here than you expect, see Exercise 13.
Very often, one is interested in deducing certain “divisibility facts” from other “divisibility facts.” For example, after checking that $3 \mid 27642$ and $3 \mid 981$, one can automatically conclude that $3$ divides their sum $28623$, without performing any calculation. This is an illustration of the following fact:

**Proposition 2.** Let $a, b, c \in \mathbb{Z}$. If $a \mid b$ and $a \mid c$, then $a \mid (b + c)$.

**Proof.** Since $a \mid b$, there is an integer $q_1$ such that $b = aq_1$. Since $a \mid c$, there is an integer $q_2$ such that $c = aq_2$. Then $b + c = aq_1 + aq_2 = a(q_1 + q_2)$. Note that $q_1 + q_2$ is the sum of two integers so it is also an integer. Finally, since we’ve verified that $b + c$ is the product of $a$ and some integer, we conclude that $a \mid (b + c)$. \[\square\]

**Your Turn 3.** In a similar way, prove that if $x$, $y$, and $z$ are integers, and if $x \mid y$ and $y \mid z$, then $x \mid z$.

**Your Turn 4.** Illustrate Your Turn 3 with several numerical examples.

The following proposition contains a very general divisibility fact, which you will be asked to prove in Exercise 10.

**Proposition 3.** Let $x, a, b, s, t \in \mathbb{Z}$. If $x \mid a$ and $x \mid b$, then $x \mid (sa + tb)$.

**Your Turn 5.** Show that Proposition 3 includes Proposition 2 as a special case.

### 1.2. Greatest common divisors and least common multiples.

How might a middle school student find the greatest common divisor of 60 and 36? One strategy might go like this. The student would find the (positive) divisors of each:

- **{positive divisors of 60} = \{1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60\}**
- **{positive divisors of 36} = \{1, 2, 3, 4, 6, 9, 12, 18, 36\}**

The “common” divisors are the divisors that 60 and 36 have “in common,” namely, \{1, 2, 3, 4, 6, 12\}. The greatest of these is twelve.

This strategy is not the most efficient one, but we highlight it since it so closely reflects the definition of the greatest common divisor of a pair of integers:
Definition 4. Let \( a, b \) be integers, at least one of them nonzero. The greatest common divisor of \( a \) and \( b \), denoted \( \gcd(a, b) \), is the largest integer that divides both \( a \) and \( b \).

Your Turn 6. Let’s check that this definition makes sense:

(a) Why do \( a \) and \( b \) necessarily have any divisors in common?
(b) Why must the collection of common divisors have a largest element? (Not every collection of integers has a largest element!)

Your Turn 7.

(a) Let \( a \) be a nonzero integer. What is \( \gcd(a, 0) \)? Explain your reasoning.
(b) In the definition, we avoided giving a meaning to \( \gcd(0, 0) \). Why?

We now turn to a related notion, the least common multiple of two integers.

Your Turn 8. A teacher starts introducing the idea of least common multiple to her middle school students. She asks them to find the least common multiple of 6 and 4. She gets a variety of answers, including 24, 12, and 0. One precocious student even says there is no answer, since one can find common multiples that are as small as you like. What might each of these students have been imagining as the (unstated) definition of “least common multiple”?

Definition 5. Let \( a \) and \( b \) be nonzero integers. We define the least common multiple of \( a \) and \( b \), denoted \( \text{lcm}(a, b) \), to be the smallest positive integer that is a multiple of both \( a \) and \( b \).

Your Turn 9. Using our calculation of \( \gcd(60, 36) \) as a model, find \( \text{lcm}(60, 36) \). (Write down positive multiples of 60 and 36, and find the smallest number common to both lists.) How does this process differ from the process of finding a gcd?

Your Turn 10. Why does \( \text{lcm}(a, b) \) exist, and why did we require both \( a \) and \( b \) to be nonzero in Definition 5?
1.3. Primes.

Definition 6. A positive integer \( p \) is prime if \( p > 1 \) and the only positive divisors of \( p \) are 1 and \( p \). If \( p > 1 \) and \( p \) is not prime, we say that \( p \) is composite. (The integers 0 and 1 are deemed neither prime nor composite.)

Primes play a fundamental role in number theory. Questions about how the primes are distributed among the positive integers are among the most fundamental (and difficult) in mathematics. Primes are just as relevant to students in pre-college mathematics: for example, they underlie strategies that a middle school student might use in writing \( \frac{72}{48} \) as a fraction in simplest form. For us, the key fact we will need about prime numbers is:

Proposition 7. Let \( a \) and \( b \) be integers, and let \( p \) be prime. If \( p \mid (ab) \), then either \( p \mid a \) or \( p \mid b \).

Your Turn 11. Construct several numerical examples that illustrate Proposition 7.

A proof of Proposition 7 is given at the end of Section 1.6.2. Also, we can generalize Proposition 7 (see Exercise 11):

Corollary 8. If a prime \( p \) divides a product \( a_1a_2\ldots a_n \), then it divides one of the factors.

The notion of being prime is an absolute concept: given a positive integer \( x \), either it’s prime or it’s not. There is a relative concept that is also useful:

Definition 9. Let \( a \) and \( b \) be integers. We say that \( a \) and \( b \) are relatively prime (to each other) if \( \gcd(a, b) = 1 \).

As examples, 28 and 45 are relatively prime (since \( \gcd(28, 45) = 1 \)), but 28 and 20 are not relatively prime (since \( \gcd(28, 20) = 4 \)).

---

4What about negative integers? One approach—the one that works best for generalizations in abstract algebra classes—is to say that for a negative integer \( x \), then \( x \) is prime if \( |x| \) is prime and \( x \) is composite if \( |x| \) is composite. This makes \(-7\) a prime number, for example. However, having negative primes would needlessly complicate some statements in this book, so we will require primes to be positive.
Your Turn 12. If \( p \) and \( q \) are primes with \( p \neq q \), do they have to be relatively prime to each other? Justify your conclusion.

Your Turn 13. List the numbers 1, 2, 3, \ldots, 40. Which of them is relatively prime to 40? Which of them is prime? Clearly explain your strategies.

1.4. The Fundamental Theorem of Arithmetic. In school, many of us learned to factor positive integers using “factor trees.” In Figure 1, we show several factor trees and the factorization of 60 produced by each.

From experience, we came to believe in two realities:

- *The process of constructing a factor tree always terminates.* If one follows any descending strand in a factor table, the numbers form a decreasing sequence. Hence, in any factor tree for 60, the longest possible strand would have length no more than sixty.\(^5\) Since each strand has finite length, the process of constructing a factor tree is guaranteed to terminate.

- *No matter which factor tree is used to factor a particular integer, the same prime factorization is obtained in the end (the same primes occur, with the same multiplicities)*, aside from the obvious fact that

---

\(^5\)In fact, the length is always much less than sixty.
the primes may be listed in a different order. We see this in the factor trees for 60: in the end, the only prime factorization for 60 seems to be \(60 = 2 \cdot 2 \cdot 3 \cdot 5\).

These two facts together constitute the Fundamental Theorem of Arithmetic:

**Theorem 10.** (Fundamental Theorem of Arithmetic for Whole Numbers\(^6\)) Any positive integer \(n\) may be written as a product of primes.\(^7\) This factorization is unique in the sense that any two factorizations of \(n\) agree except possibly on the order in which the factors occur.

**Proof.** Let an integer \(n > 1\) be given. We must prove both the existence and the uniqueness of a prime factorization for \(n\).

We begin with the existence of a factorization. Either \(n\) is prime, or \(n\) is composite. If \(n\) is prime, then we are finished. If \(n\) is composite, let \(p_1\) be the smallest divisor of \(n\) satisfying \(p_1 > 1\). Then \(p_1\) is prime (why?), and \(n = p_1n_1\) for some whole number \(n_1\) satisfying \(1 < n_1 < n\). If \(n_1\) is prime, then we are finished. If not, repeat the previous process to find a prime \(p_2\) (the smallest prime dividing \(n_1\)) and an integer \(n_2\) satisfying \(n_1 = n_2p_2\), and conclude that

\[n = p_1p_2n_2, \quad \text{with } 1 < n_2 < n_1.\]

The strictly decreasing list \(n, n_1, n_2, \ldots\) of positive integers cannot continue indefinitely (why?), and thus \(n_k\) must be prime for some \(k \geq 1\). Therefore we obtain the prime factorization \(n = p_1 \cdots p_{k-1}n_k\).

Next we handle uniqueness of the factorization. Suppose \(n\) has prime factorizations

\[n = p_1 \cdots p_s \quad \text{and} \quad n = q_1 \cdots q_r\]

with the primes arranged so that \(p_1 \leq p_2 \leq \cdots \leq p_s\), and \(q_1 \leq q_2 \leq \cdots \leq q_r\). Since \(p_1\) divides \(n\), by Corollary 8 we conclude that \(p_1 = q_k\) for some \(k \geq 1\), and thus \(p_1 \geq q_1\). Similarly, one may show that \(q_1 \geq p_1\), hence

---

\(^6\)The Fundamental Theorem of Arithmetic is an old theorem. It was known to the ancient Greeks, debuting in Euclid’s *Elements* as Proposition 31 of Book VII together with Proposition 14 of Book IX. One often associates Euclid with geometry, but Euclid devoted Books VII, VIII, and IX of his *Elements* to number theory. These propositions state “Any composite number is measured by some prime” and “If a number be the least that is measured by primes, it will not be measured by any other prime except those originally measuring it,” respectively. Unfortunately, Euclid’s *Elements* lacks a complete proof of the Fundamental Theorem of Arithmetic, which was finally given (in the more general setting of Gaussian integers) by Carl Friedrich Gauss in the early 19th century.

\(^7\)We must extend the notion of product, so that if \(n\) itself is a prime, then \(n\) is a “product” with only one factor, and if \(n = 1\), then \(n\) is a “product” in which no primes actually occur.
1. ARITHMETIC IN THE INTEGERS

$q_1 = p_1$. Dividing by this common factor gives

$$p_2 \cdots p_s = q_2 \cdots q_r$$

and if we knew that $r = s$, we could just repeat this process to show that $p_k = q_k$ for all $k$. So why is $r = s$?

Suppose that $r > s$. Then, by repeatedly dividing as above, we would obtain

$$1 = q_{s+1} q_{s+2} \cdots q_r,$$

which is impossible. Similarly $r$ can not be less than $s$, therefore $r = s$ and $p_k = q_k$ for $1 \leq k \leq s$. □

We conclude this section with some remarks on notation. Take for example the number 360. Its prime factorization is $2 \cdot 2 \cdot 3 \cdot 3 \cdot 5$. For brevity’s sake, we usually write this as $2^3 \cdot 3^2 \cdot 5^1$. Likewise, we write the prime factorization of 1400 as $2^3 \cdot 5^2 \cdot 7^1$.

However, if we had reason to think about the prime factorizations of 360 and 1400 at the same time, there is good reason to include all of the relevant primes (2, 3, 5, and 7) in both factorizations. This is easy to do: we write 360 = $2^3 \cdot 3^2 \cdot 5^1 \cdot 7^0$ and 1400 = $2^3 \cdot 3^0 \cdot 5^2 \cdot 7^1$.

How do we indicate the prime factorization of a general, unspecified positive integer? This is messy since we don’t know the specific primes that occur in the factorization, we don’t know the multiplicities (powers) with which these primes occur, and we don’t know how many distinct primes occur in the factorization. Hence, all these quantities must be indicated with letters. A typical way to represent the factorization is $p_1^{n_1} \cdot p_2^{n_2} \cdots p_s^{n_s}$. It’s understood that the primes $\{p_i\}$ are distinct, and there is no harm in assuming that $p_1 < p_2 < \cdots < p_s$. The exponents $\{n_i\}$ are understood to be nonnegative integers.

**Your Turn 14.** Give a good reason why 1 should not be considered a prime number. (Look carefully at FTA.)

1.5. Applications of FTA.

1.5.1. Divisors, multiples, and prime factorizations. As we saw at the beginning of this section, any statement about “divisors” or “multiples” ultimately relies on looking at a product like $ab = c$. We’ll assume that $a$, $b$, and $c$ are positive integers, and consider their factorizations into primes.
On the one hand, suppose that \( a = p_1^{n_1} \cdots p_s^{n_s} \) and \( b = p_1^{m_1} \cdots p_s^{m_s} \). By the properties of exponents, we know that \( c = ab = p_1^{N_1} \cdots p_s^{N_s} \), where \( N_i = n_i + m_i \). Since each \( m_i \geq 0 \), we know that \( N_i \geq n_i \). In words: if \( c \) is a multiple of \( a \), then in the prime factorization of \( c \), each prime appears at least as many times as a factor as it does in the prime factorization of \( a \). Equivalently, in each divisor of \( c \), the prime factors of \( c \) occur to the same or to smaller powers that they occur in \( c \).

Conversely, suppose we started with \( a = p_1^{n_1} \cdots p_s^{n_s} \) and \( c = p_1^{N_1} \cdots p_s^{N_s} \), and suppose we knew that \( n_i \leq N_i \) for all \( i \). We can show easily that \( a \) is a divisor of \( c \) (and \( c \) is a multiple of \( a \)), since if we define \( b = p_1^{N_1-n_1} \cdots p_s^{N_s-n_s} \), then \( b \) is an integer (why?) and \( c = ab \).

This gives us a complete picture for all the positive divisors of a positive integer, if we happen to know its prime factorization:

**Corollary 11.** Let \( c \) be a positive integer with prime factorization \( p_1^{n_1} \cdots p_s^{n_s} \), where \( p_1 < \cdots < p_s \) are (distinct) primes. The positive divisors of \( c \) are exactly the numbers of the form \( p_1^{m_1} \cdots p_s^{m_s} \), where \( 0 \leq m_i \leq n_i \) for \( 0 \leq i \leq s \).

As an example, let's see how we can find the positive divisors of 24 without testing all the numbers in the list \( 1, 2, 3, \ldots, 23, 24 \). Since \( 24 = 2^3 \times 3^1 \), Corollary 11 assures us that the (positive) divisors of 24 are the numbers of the form \( 2^i 3^j \), where \( 0 \leq i \leq 3 \) and \( 0 \leq j \leq 1 \). In a list, these numbers are \( 2^0 3^0 = 1, 2^0 3^1 = 3, 2^1 3^0 = 2, 2^1 3^1 = 6, 2^2 3^0 = 4, 2^2 3^1 = 12, 2^3 3^0 = 8, \) and \( 2^3 3^1 = 24 \). These factors may be arranged in a diagram, with primes on the bottom and strokes to indicate divisors (see Figure 2).

**Your Turn 15.** Let \( p \) and \( q \) be distinct primes. Find all the positive divisors of \( pq \) and of \( p^3 \). Draw diagrams as above. Check your examples with a specific numerical choice of \( p \) and \( q \).

1.5.2. **Prime factorization, greatest common divisors, and least common multiples.** Earlier, we showed how to compute gcds and lcms in a tiresome way, namely, by listing lots of divisors and multiples. Here is an approach that uses prime factorizations:

**Proposition 12.** Let \( a \) and \( b \) be positive integers. Suppose that the primes that occur in their prime factorizations are \( p_1 < \cdots < p_s \). Write \( a = p_1^{n_1} \cdots p_s^{n_s} \) and \( b = p_1^{m_1} \cdots p_s^{m_s} \), where we allow some of
1. ARITHMETIC IN THE INTEGERS

the exponents to be zero. Then \( \gcd(a, b) = p_1^{\min\{n_1, m_1\}} \cdot \ldots \cdot p_s^{\min\{n_s, m_s\}} \) and \( \lcm(a, b) = p_1^{\max\{n_1, m_1\}} \cdot \ldots \cdot p_s^{\max\{n_s, m_s\}} \).

Your Turn 16. Use Proposition 12 to find the \( \gcd \) and \( \lcm \) of \( 2^7 \cdot 5^8 \cdot 13^6 \) and \( 2^{10} \cdot 7^1 \cdot 13^4 \). Also use it to find the \( \gcd \) and \( \lcm \) of 60 and 36.

Your Turn 17. Explain how Proposition 12 follows from Corollary 11. (By Corollary 11, what can the prime factorization of a common divisor (or a common multiple) of \( a \) and \( b \) be?)

Your Turn 18. Use Proposition 12 to explain why \( \gcd(a, b) \cdot \lcm(a, b) = ab \), if \( a \) and \( b \) are positive integers.

1.5.3. Infinitely many primes.

Corollary 13. There are infinitely many prime numbers.

Proof. For a contradiction, suppose there are only finitely many prime numbers. Let \( p_1, \ldots, p_n \) denote the totality of these primes, and consider \( N = p_1 p_2 \cdots p_n + 1 \). By the Fundamental Theorem of Arithmetic, \( N \) is a product of primes, and hence is divisible by some prime, say \( p_1 \). Yet \( p_1 p_2 \cdots p_n \) is also divisible by...
Hence, it divides their difference, which is 1. This is impossible, since each prime is greater than 1. Therefore there are infinitely many prime numbers. \[\square\]

1.5.4. Existence of irrational numbers. FTA can be used to show that certain real numbers are irrational. (See Section 4 of Chapter 6.)

1.6. The Division Algorithm. In the previous section, we’ve seen that prime factorizations provide a conceptually easy way to understand \(\text{gcds}\) and \(\text{lcm}s\).

What may not be obvious is that there are much simpler ways to compute \(\text{gcds}\) and \(\text{lcm}s\). In this section, we’ll explore these techniques, which also turn out to be important theoretically.

1.6.1. Is it easy to compute \(\text{gcds}\)?

**Your Turn 19.** Find the greatest common divisor of 246 and 229.

What makes Your Turn 19 time-consuming to solve? The burden is not implementing Proposition 12. The problem is the work in finding the prime factorizations of the numbers. Let’s consider a similar example: finding the greatest common divisor of 1927 and 1763. As we try to factor 1927, we test 2, 3, 5, 7... and don’t find a prime divisor until we reach 41. Thus it takes a long time to discover \(41 \cdot 47\) as the prime factorization of 1927, and likewise that \(1763 = 41 \cdot 43\) (though it’s easy, given the prime factorizations, to conclude that their gcd is 41).

The difficulty in finding gcds is not so obvious when we deal with numbers less than a hundred or so, since from the multiplication tables we memorized long ago, we probably know which of these numbers is prime, and we know how to factor the composites. On the other hand, it’s often very difficult to factor a large integer—or even to know whether that number is prime.\(^8\)

Fortunately, there is a different mathematical technique, called the Euclidean Algorithm, that bypasses the factorization problem and allows us (or our computers) to compute gcds with relative ease. The Euclidean Algorithm is itself based on a fundamental fact about integers, the Division Algorithm.\(^9\)

---

\(^8\)Imagine the difficulty you would have had with Your Turn 19 if the two numbers had been ten-digit numbers, or fifty-digit numbers. Indeed, RSA encryption (a standard way of encrypting data) relies exactly upon the difficulty of factoring a large number into a product of primes. The RSA cryptosystem was developed by Rivest, Shamir, and Adleman in 1977 (hence the name ‘RSA’).

\(^9\)The Division Algorithm appeared previously in the exercises in Section 3 of Chapter 7.
Theorem 14. (Division Algorithm) Let \( a, b \in \mathbb{Z} \) with \( b \neq 0 \). There exist unique integers \( q \) and \( r \) satisfying \( a = qb + r \) and \( 0 \leq r < |b| \). The number \( q \) is called the quotient, while \( r \) is called the remainder.

Concretely, if \( a \) and \( b \) are positive, the Division Algorithm expresses the notion that \( a \) objects can be pictured as \( q \) groups of \( b \) objects with \( r \) “left over.” For example, if \( a = 96 \) and \( b = 17 \), we can easily compute that \( 96 = 5 \cdot 17 + 11 \), so \( q = 5 \) and \( r = 11 \) in this example. We interpret \( 96 = 5 \cdot 17 + 11 \) as, “96 is five groups of 17, with eleven left over.”

Your Turn 20. Let \( a \) and \( b \) be positive integers. How is the Division Algorithm useful in converting the fraction \( \frac{a}{b} \) to a mixed number?

Your Turn 21. Develop a procedure for finding \( q \) and \( r \) using a calculator and only the \( + \times - \div \) functions. (Try \( a = 96, b = 17 \) as an example.)

While it is not universally taught in schools, the Euclidean Algorithm for computing greatest common divisors already appeared in Euclid’s Elements. The theoretical basis for the algorithm is contained in the following lemma.

Lemma 15. Let \( a, b \) be integers with \( b \neq 0 \), and suppose \( a = qb + r \) as in the Division Algorithm. Then \( \gcd(a, b) = \gcd(b, r) \).

Proof. It suffices to show that the set \( D_{a,b} \) of common divisors of \( a \) and \( b \) is equal to the set \( D_{b,r} \) of common divisors of \( b \) and \( r \). We begin by verifying \( D_{b,r} \subset D_{a,b} \). If \( d \in D_{b,r} \), then automatically \( d|b \), and since \( a = qb + r \), we may apply Proposition 3 to conclude that \( d|a \). Thus \( d \in D_{a,b} \). For similar reasons \( D_{a,b} \subset D_{b,r} \), hence the sets are equal. \( \square \)

The Euclidean Algorithm itself is based on repeated application of the Division Algorithm. Let’s consider the problem of finding \( \gcd(420, 154) \). If we apply the Division Algorithm, we obtain \( 420 = 2 \cdot 154 + 112 \), so by Lemma 15, \( \gcd(420, 154) = \gcd(154, 112) \). This looks easier, since the numbers are smaller, but why stop here? Applying the Division Algorithm another time, we have \( 154 = 1 \cdot 112 + 42 \), so \( \gcd(154, 112) = \)
gcd(112, 42). The entire list of divisions we must perform is:

\[
\begin{align*}
420 &= 2 \cdot 154 + 112 \\
154 &= 1 \cdot 112 + 42 \\
112 &= 2 \cdot 42 + 28 \\
42 &= 1 \cdot 28 + 14 \\
28 &= 2 \cdot 14 + 0.
\end{align*}
\]

It’s handy to keep a running tab of the original numbers and the succession of remainders: 420, 154, 112, 42, 28, 14, 0. By Lemma 15,

\[
gcd(420, 154) = gcd(154, 112) = gcd(112, 42) = gcd(42, 28) = gcd(28, 14) = gcd(14, 0) = 14.
\]

Thus \(gcd(420, 154) = 14\). In general, the last nonzero remainder is the greatest common divisor of the two original numbers. This is the content of the Euclidean Algorithm:

**Proposition 16.** (Euclidean Algorithm) Let \(a, b\) be positive integers with \(a \geq b\). If \(b\) divides \(a\) then \(gcd(a, b) = b\). If \(b\) does not divide \(a\), then apply the Division Algorithm repeatedly:

\[
a = q_0b + r_0 \quad 0 < r_0 < b \\
b = q_1r_0 + r_1 \quad 0 < r_1 < r_0 \\
r_0 = q_2r_1 + r_2 \quad 0 < r_2 < r_1 \\
r_1 = q_3r_2 + r_3 \quad 0 < r_3 < r_2,
\]

et cetera. After finitely many iterations, a remainder of 0 is obtained. That is, there exists a whole number \(t\) with \(r_{t-2} = q_tr_{t-1} + r_t\) where \(0 < r_t < r_{t-1}\), and \(r_{t-1} = q_{t+1}r_t + 0\). Then \(gcd(a, b) = r_t\).

**Your Turn 22.** How can we use the Euclidean Algorithm to find the least common multiple of two positive integers (without computing their prime factorizations)?

1.6.2. The Division Algorithm, beyond computing greatest common divisors. We have seen that the Euclidean Algorithm is an efficient method for computing greatest common divisors. In fact, the Euclidean Algorithm is good for even more.
As we shall show in Proposition 17, the Euclidean Algorithm gives us a way to express the greatest common divisor \( \gcd(a, b) \) as an integer linear combination of \( a \) and \( b \), that is, in the form

\[
\gcd(a, b) = ax + by
\]

for specific integers \( x \) and \( y \). That may not strike the reader as particularly interesting; it is a fact about greatest common divisors that was probably not mentioned in middle school. But it turns out to be crucial in many issues in number theory. In particular, it’s the key tool in proving *If a prime divides a product, then it divides one of the factors* (Proposition 7), which was itself the key to proving the Fundamental Theorem of Arithmetic.

We continue with the example, \( \gcd(420, 154) = 14 \):

**Example 1.** Recall that our goal is to write 14 as \( 420x + 154y \) for some integers \( x \) and \( y \) (which we need to find). We do this by “backsolving” through this list of equations, starting at the bottom and working up:

\[
14 = 42 - 28 \\
= 42 - (112 - 2 \cdot 42) = (-1) \cdot 112 + 3 \cdot 42 \\
= (-1) \cdot 112 + 3 \cdot (154 - 1 \cdot 112) = 3 \cdot 154 + (-4) \cdot 112 \\
= 3 \cdot 154 + (-4) \cdot (420 - 2 \cdot 154) = (-4) \cdot 420 + 11 \cdot 154
\]

This accomplishes our goal, with \( x = -4 \) and \( y = 11 \).

**Proposition 17.** Let \( a, b \) be integers, not both zero, and let \( d = \gcd(a, b) \). There exist integers \( x \) and \( y \) satisfying \( ax + by = d \). Further, \( d \) is the smallest positive integer which may be written in the form \( au + bv \), where \( u, v \) are integers.

**Proof.** If either \( a \) or \( b \) is zero, the proposition follows readily (why?). So we assume neither \( a \) nor \( b \) is zero, and without loss of generality we may assume that \( a \) and \( b \) are both positive (why?).

We let \( I_{a,b} \) denote the collection of integer linear combinations of \( a \) and \( b \), that is,

\[
I_{a,b} = \{ au + bv : u, v \in \mathbb{Z} \}.
\]
Note that $I_{a,b}$ includes positive integers (why?), so there is a smallest positive element of $I_{a,b}$, which we denote $e$. Proposition 17 is the statement $d = e$. We will prove this by proving that $e \geq d$ and $e \leq d$.

First, we prove $e \geq d$. We begin by taking an arbitrary element $au + bv \in I_{a,b}$. By the Division Algorithm, there exist integers $q$ and $r$ such that $au + bv = qd + r$ and $0 \leq r < d$. Since $d$ divides $a$, $b$, and $qd$, we conclude that $d$ divides $r$; but since $0 \leq r < d$, we conclude further that $r = 0$. Hence $au + bv$ is a multiple of $d$; in particular, $e$ is a (positive) multiple of $d$. Hence $e \geq d$.

Second, we prove that $e \leq d$. Again by the Division Algorithm, there exist integers $q'$ and $r'$ such that $a = q'e + r'$, with $0 \leq r' < e$. We have $e \in I_{a,b}$ by assumption, and $a \in I_{a,b}$ (why?), so $r' \in I_{a,b}$ (why?). Since $e$ was the smallest positive element of $I_{a,b}$, we conclude that $r' = 0$, showing that $e$ is a divisor of $a$. For similar reasons, $e$ is a divisor of $b$, so $e$ is a common divisor of $a$ and $b$. As such, it cannot be larger than $d$, the greatest common divisor of $a$ and $b$.

We conclude with the long-anticipated proof of Proposition 7. Suppose that $p$ is a prime and $p$ divides the product $ab$. We must show that either $p|a$ or $p|b$. If $p|a$, then we are done. Suppose, then, that $p \nmid a$; we must show that $p|b$. Let’s ask: what is the greatest common divisor of $a$ and $p$? Since $p$ is a prime, its only positive divisors are 1 and $p$, so $\gcd(a, p)$, as a divisor of $p$, could only be 1 or $p$. However, it can’t be $p$, since if it were, then $p$ would be a divisor of $a$ (which we’ve assumed is not the case). Hence $\gcd(a, p) = 1$.

Here Proposition 17 comes to our aid: we can write $1 = xa + yp$, for two integers $x$ and $y$. Multiplying by $b$, we have $b = xab + ypb$. Now $p$ divides $xab$ since it already divides $ab$ by assumption, and $p$ also divides $ypb$ since we explicitly see $p$ as a factor. Hence $p$ divides the sum $xab + ypb$, so $p$ divides $b$, concluding the proof.

1.7. Exercises.

Terminology and notation:

1. Suppose $a$ and $b$ are integers. Complete the definition: We say that $a$ is a divisor of $b$ if there exists

2. Mark each statement as true or false.

(a) $6|48$
(b) $48|6$
(c) $5|13$
(d) $13|5$
(e) $1|237$
(f) $237|1$
(g) $-1|9$
(h) $-1|-9$

3. Write each of the following statements using the symbol $|$ or $\nmid$.

(a) $u$ does not divide $r$.
(b) $x$ is a multiple of $y$.
(c) $v$ is a factor of $w$.
(d) $f$ is not a divisor of $g$.

4. Motivated by the equation $15 = 3 \times 5$,

(a) Write two correct sentences using the phrase “is a multiple of.”
(b) Write two correct sentences using the phrase “is a divisor of.”

5. Rewrite the statement of Your Turn 3 in words, using the word *multiple* and avoiding the words *divides* and *divisor*.

*Divisibility:*

6. List the divisors of $-12$.

7. In the following questions, find *small positive integer* values of $a$, $b$, and $c$ that meet the requirements.

(a) Find numerical values for $a$, $b$, $c$ such that $a|b$, $b|c$, and $a|c$.
(b) Find numerical values for $a$, $b$, $c$ such that $a|b$, $b \nmid c$, and $a|c$.
(c) Find numerical values for $a$, $b$, $c$ such that $a|b$, $b \nmid c$, and $a \nmid c$.
(d) Find numerical values for $a$, $b$, $c$ such that $a \nmid b$, $b|c$, and $a|c$. 
(e) Find numerical values for \( a, b, c \) such that \( a \nmid b, b \mid c \), and \( a \nmid c \).

(f) Find numerical values for \( a, b, c \) such that \( a \nmid b, b \mid c \), and \( a \mid c \).

(g) Find numerical values for \( a, b, c \) such that \( a \nmid b, b \nmid c \), and \( a \nmid c \).

(h) There is one remaining case. What is it, and why does no example exist?

8. Explain why the following is true: if \( x \) divides \( a \), then \( x \) divides the sum of \( a \) and (any) multiple of \( a \).

9. Suppose that \( a \mid b \). Prove that \( a \mid (-b) \), \( (-a) \mid b \), and \( (-a) \mid (-b) \).

10. Let \( x, a, b, s, t \in \mathbb{Z} \). Prove that if \( x \mid a \) and \( x \mid b \), then \( x \mid (sa + tb) \).

11. Under what conditions does it happen that \( a \mid b \) and \( b \mid a \)? Formulate and prove a conjecture.

12. Prove or disprove: If \( a > b > 0 \), then \( a \) has more divisors than \( b \).

13. In abstract algebra courses, one calls a nonzero quantity \( x \) a divisor of \( 0 \) if there exists a nonzero quantity \( y \) such that \( xy = 0 \). (In many algebraic structures, divisors of zero exist. For example, in the context of modular arithmetic, we have that \( 3 \times 4 \equiv 0 \pmod{12} \), so \( 3 \) and \( 4 \) are both divisors of zero in \( \mathbb{Z}/12 \).)

   (a) How does this definition differ from Definition 1?

   (b) According to the definition stated in this exercise, what are the divisors of zero in \( \mathbb{Z} \)?

   Factorization, gcds, and lcm's (computational exercises):

14. Find \( \gcd(4, 6) \), \( \gcd(3, 12) \), \( \gcd(8, 15) \), \( \gcd(1, 6) \), and \( \gcd(0, 12) \).

15. For the following choices of \( a \) and \( b \), compute \( \gcd(a, b) \), \( \text{lcm}(a, b) \), and \( \gcd(a, b) \cdot \text{lcm}(a, b) \). Compare this to \( a \cdot b \).

   (a) \( a = 2, b = 12 \).

   (b) \( a = 6, b = 45 \).

   (c) \( a = 24, b = 36 \).

   (d) \( a = 10, b = 21 \).

17. Create diagrams indicating the divisors of 60, as in Figure 2.

18. Let $a = 6$ and $b = 16$. By trial and error, find integers $x$ and $y$ such that $ax + by = 2$.

19.

(a) List all the positive divisors of 100. List all the positive divisors of 72. By comparing the two lists, find gcd(100, 72).

(b) Use the Euclidean Algorithm to find gcd(100, 72).

(c) Using your calculations in (b), find integers $x$ and $y$ such that $gcd(100, 72) = 100x + 72y$.

(d) Find lcm(100, 72), and explain several strategies for making this computation.

(e) Verify that $gcd(a, gcd(a, b)) = 1$ if $a = 100$ and $b = 72$. (This idea recurs in Exercise 27.)

20. Find the greatest common divisor and the least common multiple of $2^3 \cdot 3^2 \cdot 7^2$ and $3^3 \cdot 5^1 \cdot 7^1$.

21. Use the Euclidean Algorithm to compute gcd(1927, 1763).

22. Find the greatest common divisor and the least common multiple of 360 and 675, first using prime factorizations, and then using the Euclidean Algorithm and Your Turn 18.

23. Find the greatest common divisor and the least common multiple of 2,744, 431 and 1,603, 687.

24. Find the prime factorization of 14651. Explain your strategy, including any shortcuts you take.

25. The number 6 can be written as a product of two positive integers in four different ways, namely as $1 \cdot 6, 2 \cdot 3, 3 \cdot 2,$ and $6 \cdot 1$. In how many such ways can 16 be written as a product? Try the same question for 12 and for 60.

Factorization, gcds, and lcm's (theoretical exercises):

26. Generalize Exercise 25: in how many ways can $p_1^{n_1} \cdots p_s^{n_s}$ be written as the product of two positive integers?
27. Prove that \( \gcd \left( \frac{a}{\gcd(a, b)}, \frac{b}{\gcd(a, b)} \right) = 1 \). (This does not require prime factorizations, the Euclidean Algorithm or Proposition 17, only the definition of the greatest common divisor.)

28. Use Proposition 12 to solve Exercise 27.

29. To what extent does it make sense to talk about “unique factorization of a positive rational number into a product of primes”?

30.
   (a) How many positive divisors does \( p_1^{n_1} \cdots p_s^{n_s} \) have? (Use Corollary 11.)
   (b) Find a number \( a \) with exactly four positive divisors.
   (c) Find a number \( b \) with exactly three positive divisors.
   (d) Find a number \( c \) that is divisible by exactly three primes and has exactly eight positive divisors.
   (e) Find a number \( d \) that is divisible by exactly two primes and has exactly eight positive divisors.
   (f) Find a number \( e \) that is divisible by exactly two primes and has exactly nine positive divisors.

31. Find all positive integers \( x \) such that \( \text{lcm}(12, x) = 120 \).

32. Prove that there are infinitely many positive integers \( x \) such that \( \gcd(100, x) = 10 \).

33. Prove or disprove: if \( c \mid (ab) \) and \( \gcd(a, b) = 1 \), then \( c \mid a \) or \( c \mid b \).

34. Theorem: If \( a, b, \) and \( c \) are positive integers, if \( a \mid c \) and \( b \mid c \), then \( \text{lcm}(a, b) \mid c \). (Compare to Let’s Go I(a).)
   
   (a) Illustrate the theorem with a well-chosen example.
   (b) Using prime factorizations, prove the theorem.
   (c) Give a proof that does not use prime factorizations. (Apply the Division Algorithm to \( c \div \text{lcm}(a, b) \).)

35. Theorem: If \( a, b, \) and \( c \) are positive integers, if \( a \mid c \) and \( b \mid c \), and if \( \gcd(a, b) = 1 \), then \( (ab) \mid c \).
   (Compare to Let’s Go I(b).)

   (a) Illustrate the theorem with a well-chosen example.
(b) Prove the theorem, using Exercise 34 and Your Turn 18.

(c) Prove the theorem directly. (Start with the fact that since $\gcd(a, b) = 1$, we can write $1 = ax + by$ for two integers $x, y$.)

36. In the beginning of Section 1.2, we saw that the common positive divisors of 36 and 60 are 1, 2, 3, 4, 6, 12; hence $\gcd(36, 60) = 12$. We could have noticed something stronger: not only is 12 greater than the other common divisors, it is actually a multiple of each of them. In general: If $a$, $b$, and $c$ are positive integers, if $c|a$ and $c|b$, then $c|\gcd(a, b)$.

(a) Give a proof that uses prime factorizations.

(b) Give a proof that does not use prime factorizations. (Start by applying the Division Algorithm to $\gcd(a, b) \div c$. You also will need to use Proposition 17.)

37. In showing that 1777 is prime, do we need to check all numbers between 2 and 1777 as potential factors of 1777? For example, do we need to check 6? Do we need to check 101? Explain.

38. Prove that 41 is a prime. Show your work.

*Induction:*

39. Let $p$ be prime and suppose that $p|a_1 \cdots a_n$ for some integers $a_1, \ldots, a_n$. Use induction and Proposition 7 to show that $p|a_k$ for some $k$.

40. Revisit the proof of the Fundamental Theorem of Arithmetic in the text. Show that both the “existence” and “uniqueness” proofs can be shortened by using induction as a proof technique.

*Applications to fractions:*

41. One sometimes hears people talking about “greatest common denominators” and “least common denominators.” Do either of these phrases make sense? Is either of them useful? Explain.

42. Using prime factorizations of the numerator and denominator, find a simplest-form fraction that is equivalent to $\frac{120}{36}$. 
43. Find a simplest-form fraction that is equivalent to \( \frac{123.889}{9.085.117} \).

44. Prove that the fraction \( \frac{a}{b} \) is equivalent to \( \frac{a}{\gcd(a,b)} \cdot \frac{b}{\gcd(b, a)} \), which is in simplest form. (What other exercise is helpful?)

45.

(a) Compute the sum \( \frac{13}{84} + \frac{33}{70} \), using \( 84 \cdot 70 \) as a common denominator. Then, convert your answer to a fraction in simplest form.

(b) Compute the sum using \( \text{lcm}(84, 70) \) as a common denominator.

46. Find the easiest example you can to show that the right-hand side of the equation \( \frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \) need not be in simplest form, even if \( \frac{a}{b} \) and \( \frac{c}{d} \) are in simplest form.

47. Suppose that \( \frac{a}{b} \) and \( \frac{c}{d} \) are in simplest form, and one uses \( \text{lcm}(b, d) \) as a common denominator for addition. Must the sum be in simplest form? Give a proof or a counterexample.

48. Suppose that \( \gcd(b, d) = 1 \), \( \gcd(a, b) = 1 \), and \( \gcd(c, d) = 1 \). Prove that the right-hand side of the equation \( \frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \) must be in simplest form.

**Irrational numbers:**

49. Let \( a \) and \( b \) be relatively prime positive integers. Prove that \( \log_a b \) is irrational.

50. Let \( a \) and \( b \) be positive integers. Under what circumstances is \( \log_a b \) a rational number? (It may help to consider Exercise 49 first.)

51. Let \( p \) be a positive prime number. Show that \( \sqrt{p} \) cannot be written as a quotient of integers.

2. Systems of Numeration for Whole Numbers

**Let’s Go 2.** As English speakers, we have inherited the alphabet of the Romans, but not their system of numeration. What do you think accounts for this asymmetry?
2. Systems of Numeration for Whole Numbers

2.1. Numbers and numerals. What is the distinction between number and numeral? In essence, a number is an abstract quantity, that which characterizes the “number of elements in a collection of objects.” In contrast, a numeral is a symbolic way of representing a number in written form. Thus, “sixteen” (in English), “16” (in base-ten Hindu-Arabic numerals), and “XVI” (in Roman numerals) are numerals that represent the number of dots in the collection ••••••••••. From time to time, we use each of these systems. We use the first two (English and Hindu-Arabic) every time we write a check. We use Roman numerals when a sense of permanent significance is intended, as on the cornerstone of a building we expect to endure, or in numbering monarchs and Superbowls.

By what criteria should we judge a numeration system? Here are some key considerations:

- The system should allow numbers to be written concisely and accurately.
- It should be easy to compare which of two numerals represents the larger number.
- Key computations (such as addition and multiplication) should be easy to perform in the numeration system.

As we will see, the base ten system is phenomenally successful in meeting all these criteria.

Why is it worth making a distinction between numeral and number? First, our long experience with base-ten numerals might lead us to believe that they are the only “real” way of representing numbers, as though any culture, isolated from the rest of the world, would quickly discover them. By examining historical systems of numeration, we see that this is not the case. Second, our ingrained habit of using them might also make us forget that on occasion, they are not the best way to write down certain numbers. Let’s consider an example: suppose we want to know the number of ways (orders) of arranging six people in a list. It is easy to see that the answer is \(6 \times 5 \times 4 \times 3 \times 2 \times 1 = 6!\) (six factorial). Here we might compulsively multiply out to obtain 720 as the “answer.” However, it’s arguable whether this is the best thing to do. In fact, one sees the significance of the number when it is written as 6!, not when it is written as a base ten number. Likewise, in number theory, it often is more helpful to have a number written as a product of primes than as a base ten numeral, since prime factorizations reveal the divisors of a number more transparently than do base ten numerals.
2.2. Ancient numeration systems. Let’s imagine a culture just beginning to develop a system of numeration. Two people want to communicate the number of sheep, or sacks of flour, or pairs of shoes being sold. Suppose there are twenty-three such objects. The most natural thing to do is to record the number by drawing 23 strokes or dots or circles or some other simple figure:

||| |

Your Turn 23. In the last section, we gave a list of criteria that a good numeration system should possess. Which of these hold for this system?

It’s hard to stare at a list of twenty-three strokes and quickly ascertain the number of strokes in the list. So it would be natural for our merchant to group the strokes together—perhaps into groups of ten, if he were accustomed to counting on fingers. This would give the “numeral”

||| |

From here, it’s not much of a leap to think of substituting another symbol for the list of ten, achieving some brevity in the numeral:

**|||

Something like this happens in many numeration systems. For example, consider the numeration system used by the ancient Egyptians.\(^1\) Like our system, the Egyptian system is a base ten system. That is, it employs powers of ten, with the figures |, ∩, @, , , and representing the numbers one, ten, one hundred, one thousand, ten thousand, one hundred thousand, and one million, respectively. Whole numbers were expressed by combining these symbols. For example, where we would write 2537, they might write any one of the numerals in Figure 3 in the Egyptian system. We observe an important distinction between the numeration systems: in the Egyptian system, one may swap the order of symbols without affecting the number they represent. This is certainly not true in our system; for example, 456 and 546 represent different

\(^{10}\)The *Rhind papyrus*, an Egyptian mathematical document dating from 1650 B.C., is one of the few sources of information concerning the nature and scope of ancient Egyptian mathematics.
Figure 3. 2537 in Egyptian numbers. Specifically, in the first numeral, the 4 represents four hundred, while in the second it represents forty. Thus we say that our system has \textit{place value} (i.e., the location of a symbol within a numeral determines the numerical value of the symbol), whereas the Egyptian system does not.

\textbf{Your Turn 24.} We’ve seen that the Egyptian system relies on a short list of “fundamental symbols.” Our numeration system also relies on a short list of fundamental symbols, namely, 0,1,2,3,4,5,6,7,8,9. How do the roles of these symbols differ in the two numeration systems?

\textbf{Your Turn 25.} In what sense does the system of Roman numerals have place value? In what ways is the system of place value different in the Roman and base-ten systems?

How does a lack of place value affect Egyptian arithmetic? First, it is difficult to express very large numbers efficiently. For example, to express the number 100,000,000 requires one hundred little kneeling men in the Egyptian system (or a brand new symbol). This is because the Egyptian symbols do not have flexible values that can vary depending on their position in the numeral. Second, addition of whole numbers in the Egyptian system is relatively easy, and is accomplished by pushing symbols together. For example, 67 + 48 would become

However, to express the answer efficiently requires us to \textit{trade} ten tens for a one hundred, just as in our Hindu-Arabic number system, but apparently the Egyptians lacked an efficient method for trading.
Your Turn 26. Show how to use trading to finish the addition problem above. Use Egyptian, not Hindu-Arabic numerals. (Your strategy will be analogous to the use of manipulatives in many elementary school classrooms.)

Finally, the Egyptian method of multiplication, while beautiful in its own right, is at the same time rather tedious. (We will show the strategy while using our familiar Hindu-Arabic numerals.) In order to compute \(a \times b\), one starts with the pair \((1, b)\) and generates a list of pairs by repeatedly doubling each of \(1, b\). Then, one identifies the elements of the list whose first components add to \(a\) (this boils down to expressing \(a\) in base two). To finish, one adds the second components of these distinguished elements to obtain the product \(a \times b\). The whole process is known as the method of duplication, since it relies on repeated doubling.\(^{11}\)

The following computation of \(24 \times 27\) illustrates the Egyptian method of multiplication. The symbol \(\backslash\) indicates that a product is being selected for use in the total product.

\[
\begin{array}{c|c}
1 & 27 \\
2 & 54 \\
4 & 108 \\
\backslash 8 & 216 \\
\backslash 16 & 432 \\
\hline
\text{total} & 648
\end{array}
\]

Your Turn 27. Discuss a method for computing the product of IX and XCVIII (Roman numerals).

2.3. Hindu-Arabic numeration. The Hindu-Arabic numeration system originated in 7th century India\(^{12}\) and was further adapted by the Arabs\(^{13}\) who conquered parts of northern India in the 7th century. By the 8th century the system was in use throughout the Islamic world, from Turkey through North Africa to Spain. The Hindu-Arabic system was introduced to western Europeans in the 10th century by Gerbert d’Aurillac, who learned it while studying in Spain. However, it was not until the early renaissance that

\(^{11}\)The duplication method of multiplication is a longstanding legacy of ancient Egyptian mathematics. It was routinely taught in schools as an alternate method of multiplication as recently as the 19th century.

\(^{12}\)The first known reference to the Hindu-Arabic system is given in A.D. 662 by a Syrian priest, who wrote that the Hindus had a remarkable means of calculation “using nine signs.” Presumably the priest omitted zero, which at that time in history was represented by a dot.

\(^{13}\)The Arabs were not merely guardians of the knowledge of the classical world—they made contributions of their own, including the beginnings of symbolic algebra. The word algebra derives from the Arabic al-\(\text{jabr}\), which roughly means “reunion and reduction.”
Hindu-Arabic numerals became widely used in Europe, at first mostly by merchants who found them to be convenient for calculations.

The Hindu-Arabic system is a base ten, place value system in which one may uniquely express any whole number whatsoever by using only the ten symbols (digits) 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Let’s recall how it works. The digits are used to represent successive groups of ten (i.e., powers of ten), in decreasing order. For example, 57406 represents 

\[(5 \times 10^4) + (7 \times 10^3) + (4 \times 10^2) + (0 \times 10^1) + (6 \times 10^0).\]

Any given digit in a Hindu-Arabic numeral has both a face value and a place value. For example, in 57406, the face value of 6 is six, while the place value of 7 is seven thousand.

By examining our standard algorithms for addition and multiplication, one may see that calculations in our system boil down to gathering powers of ten. Importantly, since a digit such as 6 may represent six, sixty, or six thousand, it’s a relief to realize that

- the additional and multiplication facts one must remember are restricted to the numbers zero through nine.

This is why we all learned addition and multiplication tables for the numbers zero through nine in school. These are precisely the arithmetic “facts” we need to compute any product or sum (along with keeping track of powers of ten). We may see all of this more clearly by producing “expanded” versions of our common algorithms which highlight the gathering of powers of ten. For example, consider the sum 5729 + 814 using the standard algorithm and an expanded algorithm (see Figure 4).

Next, consider the product 28 × 34. Figure 5 shows the scratchwork that most of us would record in computing the product.

**Your Turn 28.** Given an “expanded” version of the multiplication algorithm that illustrates and justifies the calculation of 28 × 34 in Figure 5.

Even after all this, due to its sheer familiarity, it may be difficult to study our number system objectively. One way to help us comprehend our system as well as our common arithmetic algorithms is to study them in a place-value system possessing a base other than ten. For example, suppose we wish to work in a base six
system. In this new system, we express whole numbers in terms of powers of six, using the digits 0, 1, 2, 3, 4, 5.

To the numerals thus formed we attach the subscript \( \text{six} \), to indicate that we are working in base six rather than base ten.

Counting in base six also helps us get used to the idea:

\[ 1\text{six} \quad 2\text{six} \quad 3\text{six} \quad 4\text{six} \quad 5\text{six} \quad 10\text{six} \quad 11\text{six} \quad 12\text{six} \quad 13\text{six} \quad 14\text{six} \quad 15\text{six} \quad 20\text{six} \quad 21\text{six} \quad 22\text{six} \quad 23\text{six} \]

Of course, there are more efficient ways of converting from base six to base ten and vice-versa than listing out the numerals. For instance, \( 253_{\text{six}} \) represents one-hundred five, since

\[
253_{\text{six}} = 2 \cdot 6^2 + 5 \cdot 6^1 + 3 \cdot 6^0 = 72 + 30 + 3 = 105.
\]
Thus, conversion from non-base-ten to base ten relies simply on an interpretation of place value.

Conversion from base ten to base six requires a bit more care. The main idea is:

- To write a number in base six, we must gather (disjoint) groups of six.

One may get used to the idea by considering figures. For example, consider the task of writing the 51 dots given below as a base six numeral.

We begin by circling groups of six dots. We see that there are eight groups of six with three dots left over. We summarize by writing $51 = 8 \times 6 + 3$.

Then, we may circle groups of six groups.

We see that there is one group of six groups with two groups of six and three dots left over. We summarize by writing $51 = 1 \times 6^2 + 2 \times 6 + 3$. Since there are not six full groups of thirty six, we declare ourselves done and write $54 = 123_{\text{six}}$.

The process for converting larger numbers to base six is essentially the same, relying on repeated division by six. For example, suppose we are to convert 5389 to base six. We first divide by six to obtain

$$5389 = 898 \cdot 6 + 1.$$
We have broken 5389 into groups of six, with a remainder of one. Now, we wish to break our 898 groups of six into further groups of six. Dividing 898 by 6, we obtain

\[ 5389 = (149 \cdot 6 + 4) \cdot 6 + 1 = 149 \cdot 6^2 + 4 \cdot 6 + 1. \]

We continue by dividing our 149 groups of \(6^2\) into groups of 6:

\[ 5389 = (24 \cdot 6 + 5)6^2 + 4 \cdot 6 + 1 = 24 \cdot 6^3 + 5 \cdot 6^2 + 4 \cdot 6 + 1. \]

Finally, we see that the 24 groups of \(6^3\) may be broken evenly into 4 groups of \(6^4\), and thus

\[ 5389 = (4 \cdot 6 + 0)6^3 + 5 \cdot 6^2 + 4 \cdot 6 + 1 = 4 \cdot 6^4 + 0 \cdot 6^3 + 5 \cdot 6^2 + 4 \cdot 6 + 1. \]

We conclude that \(5389 = 40541_{\text{six}}\).

**Your Turn 29.** With the discussion above as a guide, explain why every whole number can be represented by a base six numeral. (For the same reason, every whole number can be represented by a base ten numeral.)

### 2.4. Exercises.

1. How would the ancient Egyptians have expressed 24,536,972?

2. Find the face value and the place value of the digit 3 in the numeral 8,345,821.

3. Write the numbers from one to twenty in base two, base five, base ten, and base twelve. (For base twelve, let 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, T, E be the basic digits, with T and E representing ten and eleven, respectively.)

4. Interpret 100 as a base two, base seven, and base nine numeral. (What is the corresponding base ten numeral?)

5. Describe how 7777\text{eight}, 111111\text{two}, and 999\text{ten} are related to a power of the base.

6. Convert 4332\text{five}, E39\text{twelve}, and 100111\text{two} to base ten. (For base twelve, use the notation from Exercise 3.)

7. Consider the set of dots given in Figure 6.
(a) Circle as many (disjoint) groups of three dots as possible.

(b) After circling groups of three dots, circle as many (disjoint) groups of *three groups* as possible.

(c) Use your figures from parts (a) and (b) to write the number of dots as a base three numeral.

8. Suppose you had a large tub of pennies that you had to count. In a paragraph, explain how you could count them by first putting them in piles of ten, then gathering ten piles of ten into piles of one hundred, then gathering piles of one hundred into piles of one thousand, et cetera. How do you wind up with the base ten numeral for the number of pennies, from the piles of pennies at the end of the process?

9. Convert $240653_{\text{ten}}$ to base sixteen. (Use $A = \text{ten}, \ldots, F = \text{fifteen}$.)

10. Convert $166167_{\text{ten}}$ to base six.

11. What advantages accrue to using a large number as a base? What advantages accrue to using a small number as a base?

12. Instinctively, we can glance at base-ten numerals such as

$$6271 \quad 386854 \quad 79452 \quad 80662 \quad 6290$$
and arrange them in increasing order. In other words, it is easy to compare numbers written as base ten numerals.

(a) Write a complete set of instructions for how to determine the larger of two base-ten numerals.

(b) How can one compare two Egyptian numerals?

13. Convert 1287 to base two, to base four, to base eight, and finally to base sixteen. How are these numerals related to each other?

14. Earlier we mentioned that the ancient Egyptian method of multiplication involved a “conversion to base two.”

(a) Convert 58 to base two.

(b) Use the ancient Egyptian method of multiplication to compute $58 \times 234$. Explain how your answer to part (a) is useful here.

15. If we “count by threes” in base ten, we have the succession of numerals, 3, 6, 9, 12, 15, … Count by threes in base five, until you reach $201_{\text{five}}$. (Write the succession of numerals in base five.)

16. Create addition and multiplication tables in base six for the numbers 0, 1, 2, 3, 4, 5. (The sums and products should be written as base six numerals.)

17. By adapting familiar algorithms for addition, subtraction, and multiplication, perform the following computations without converting to base ten. (Use Exercise 16.)

(a) $1413_{\text{six}} + 453_{\text{six}}$

(b) $5432_{\text{six}} + 3045_{\text{six}}$

(c) $1413_{\text{six}} - 453_{\text{six}}$

(d) $40032_{\text{six}} - 2545_{\text{six}}$

(e) $42_{\text{six}} \times 35_{\text{six}}$

(f) $205_{\text{six}} \times 134_{\text{six}}$

18. Prove that the base ten numeral for any whole number is unique.
19.

(a) Use the pencil-and-paper algorithm you were taught in elementary school to perform the division problem $38475 \div 21$.

(b) How do you interpret the answer given by this algorithm?

(c) Give a justification that the algorithm gives the correct answer. Your response should make explicit reference to the base-ten notational system.

3. Divisibility Tests

There are easy and familiar tests for quickly checking whether a whole number $n$ is divisible by $a$, for certain choices of $a$ such as 2, 3, and 5. These are useful in shortening the work in factoring numbers, and in checking that numbers are prime. In stating and proving these “divisibility tests,” we see the interplay between properties of numbers (such as “the number is divisible by three”) and properties also of numerals (such as “the sum of the digits of the base-ten numeral is divisible by three”).

3.1. Familiar tests and their proofs. The divisibility tests we will discuss all take the following form: A number $n$ is divisible by $a$ if and only if another number $n'$ is divisible by $a$. What makes the tests useful is that it is usually much easier to check whether $n'$ is divisible by $a$ than to check whether $n$ is divisible by $a$. In proving the tests, the punch line always is that $a$ divides $n - n'$.

**Your Turn 30.** Why does $n = n' + a \cdot b$ imply that $n$ is divisible by $a$ exactly when $n'$ is?

Here’s the familiar test for divisibility by two:

**Proposition 18.** A whole number $n$ is divisible by 2 exactly when its ones digit $n'$ is divisible by 2.

**Your Turn 31.** Suppose $n = 2754$.

(a) What is $n'$?

(b) Carefully explain why $n - n'$ is a multiple of 2.

(c) Based on this example, give a proof of Proposition 18.

**Your Turn 32.** State and prove a similar test for divisibility by 5.
We will give another proof of Proposition 18, using some new notation. We “write out” all the digits of $n$. For example, if $n = 2754$, we write $n = a_3 a_2 a_1 a_0$, where $a_3 = 2$, $a_2 = 7$, $a_1 = 5$, and $a_0 = 4$. Note that $a_0$ is another name for $n'$. Also be careful to note that $a_3 a_2 a_1 a_0$ is not the product of the four numbers; instead, it stands for $(a_3 \times 10^3) + (a_2 \times 10^2) + (a_1 \times 10^1) + a_0$.

Here is our new proof of Proposition 18. Suppose $n = a_k a_{k-1} \ldots a_2 a_1 a_0$. We merely note that

$$n - n' = a_k a_{k-1} \ldots a_2 a_1 0 = 10 \times a_k a_{k-1} \ldots a_2 a_1 = 2 \times (5 \times a_k a_{k-1} \ldots a_2 a_1),$$

which is a multiple of two.

The test for divisibility by 3 is only slightly harder to prove.

**Proposition 19.** A whole number is divisible by 3 exactly when the sum of its digits is divisible by 3.

For example, to check whether 2,457,382 is divisible by 3, we can compute $2 + 4 + 5 + 7 + 3 + 8 + 2 = 31$, which is not divisible by 3. Hence 2,457,382 is not divisible by 3.

**Proof.** Suppose $n = a_k \ldots a_2 a_1 a_0 = (a_k \times 10^k) + \cdots + (a_2 \times 10^2) + (a_1 \times 10^1) + (a_0 \times 1)$. Let $n'$ be the sum of the digits, that is, $n' = a_k + \cdots + a_2 + a_0 + a_0$. Then

(1)  

$$n - n' = a_k \cdot (10^k - 1) + \cdots + a_2 \cdot (10^2 - 1) + a_1 \cdot (10 - 1) + a_0 \cdot (1 - 1).$$

We must show that this number is divisible by 3, and for this, it is enough to show that each $10^k - 1$ is divisible by 3. But this is clear, since $10^k - 1 = 9 \cdot 10^{k-1} + \cdots + 9 \cdot 10^2 + 9 \cdot 10 + 9 - 1 = 3(3 \cdot 10^{k-1} + \cdots + 3 \cdot 10^2 + 3 \cdot 10 + 3 - 1)$. □

Fortunately, we can combine many divisibility tests. Recall that if $a$, $b$, and $n$ are positive integers, if $a|n$ and $b|n$, and if $\gcd(a, b) = 1$, then $(ab)|n$ (Exercise 35 in Section 1.7). This helps us extend our divisibility tests. For example, a number is divisible by 6 if and only if it is divisible by both 2 and 3 (and we have easy divisibility tests for these).

**3.2. Exercises.**

1. Prove that $n$ is divisible by 4 exactly when the numeral obtained by erasing all but the tens and ones digits represents a number divisible by 4.

2. State and prove a test for divisibility by 8.
3. Prove that a number is divisible by 9 exactly when the sum of its digits is divisible by 9.

4. Prove that a whole number is divisible by 10 if and only if its ones digit is 0.

5. Give an argument, using divisibility tests, that 4,770 is divisible by 45.

6. With divisibility tests as an aid, find the prime factorization of 9,144,576,000.

7. State a divisibility test for divisibility by $2^t$, where $t$ is a positive integer.

8. The test for divisibility by 4 (given in Exercise 1) can be expressed as, $n$ is divisible by 4 exactly when its remainder upon division by 100 is divisible by 4. Why?

9. Prove or disprove: if a number is divisible by 27, then the sum of its digits is divisible by 27.

10. Prove or disprove: if the sum of the digits of a number is divisible by 27, then the number is divisible by 27.

11. Give an argument, using divisibility tests, that 402,948 is divisible by 6 and divisible by 4. Is it divisible by 24?

12. In this exercise, all numerals are in base twelve.

   (a) Prove that a number $n$ is divisible by three if and only if its last digit is divisible by three.

   (b) Prove that a number $n$ is divisible by eleven if and only if the sum of its digits is divisible by eleven.

   (c) How can one tell when $n$ is divisible by twelve? By $144_{\text{ten}}$?

In your work, label all numerals so it is clear whether they are in base ten or base twelve.

4. Decimals

Let’s Go 3.

(a) By long division, find decimals that represent the fractions $\frac{1}{5}$ and $\frac{1}{95}$.

(b) On the basis of your answers to (a), find decimals that represent the fractions $\frac{1}{7}$, $\frac{7}{31}$, and $\frac{28}{9999}$. 
Let’s Go 4. Interpret the repeating decimal 0.21212121 as the infinite series \( \frac{2}{10} + \frac{1}{100} + \frac{2}{1000} + \frac{1}{10000} + \ldots \). How could you find a simple expression for this infinite sum?

Let’s Go 5. One of your middle-school students asserts that \( \frac{2}{3} = 0.66 \). Another counters that \( \frac{2}{3} = 0.67 \). A third claims that \( \frac{2}{3} = 0.66666666 \). A fourth asserts that \( \frac{2}{3} \) can’t be written as a decimal at all. Are any of the students correct? What would you tell these students to help them understand writing \( \frac{2}{3} \) as a decimal?

Remarkably, the base ten Hindu-Arabic numeration system for whole numbers can be adapted to describe all real numbers. (That’s a long-winded way of reminding ourselves that “every real number can be given by a decimal.”) In this section, we review our intuitive understanding of decimals, recall their connection with infinite series, and describe the delicate interplay between fractions and decimals.\(^{14}\)

4.1. Making sense of decimals. An intuitive way to interpret a decimal is as an address on the number line, based on repeated subdivision by ten. For example, suppose we wished to “locate the number \( a = 85.73 \)” on the number line. The initial digit 8 tells us that \( a \) should be placed between 80 and 90 on the number line. The next digit, 5, tells us that \( a \) is between 85 and 86; what we really should focus on is that the placement of the integers 80, 81, . . . , 88, 89, 90 come from subdividing the interval \([80, 90]\) into ten subintervals of equal length. This gives direct inspiration to the notion of decimals: we subdivide by ten again (and again…). For example, the process of subdividing the interval \([85, 86]\) into ten subintervals gives rise to a natural placement of the numbers 85.0 = 85, 85.1, 85.2, . . . , 85.8, 85.9, 86.0 = 86. Further, the number \( a = 85.73 \) lies between 85.7 and 85.8. After one more subdivision by 10, we hit \( a \) “on the nose.”\(^{15}\)

It’s easy to recast the “address interpretation” more algebraically. For example, we can interpret 85.73 as the finite sum \( (8 \times 10^1) + (5 \times 10^0) + (7 \times 10^{-1}) + (3 \times 10^{-2}) \), or in fractional notation as \( (8 \times 10) + (5 \times 1) + (\frac{7}{10}) + (\frac{3}{100}) \). As a sum of rational numbers, it is a bona-fide rational number.

\(^{14}\) The decimal system was introduced in 1585 by Simon Stevin (1548-1620) in his pamphlet De Thiende. The subtitle of the 1608 English version of the pamphlet reads “Teaching how all computations that are met in business may be performed by integers alone without the aid of fractions.”

\(^{15}\) This notion of a decimal as an “address” is conceptually easy. In fact, students in countries using the metric system of measurement may encounter decimals before fractions, because the “address interpretation” of a decimal, which arises naturally via subdivision by ten, is so closely tied to the metric system of measurement. In contrast, the English system of measurement—with varying conversion factors, familiar to us by the rules like twelve inches in a foot, 5280 feet in a mile, eight fluid ounces in a cup—is more naturally suited to fractions than to decimals.
This handles the case of terminating decimals. But how are we to make sense of the decimal expansion for $\pi$, which begins as $3.1415926535897932385\ldots$ but never ends?

Let’s examine the “address interpretation” first. By considering successive truncations of the decimal (truncating farther and farther to the right), we obtain an infinite succession of inequalities:

$$3 \leq \pi \leq 4 \quad 3.1 \leq \pi \leq 3.2 \quad 3.14 \leq \pi \leq 3.15 \quad 3.141 \leq \pi \leq 3.142 \quad \ldots$$

Each of these inequalities can be thought of as giving upper and lower estimates for the value of $\pi$ ($\pi$ is between 3 and 4; better, $\pi$ is between 3.1 and 3.2; still better, $\pi$ is between 3.14 or 3.15, etc.). What is useful, of course, is the difference of the two approximations keeps going down by a factor of $\frac{1}{10}$, since $4 - 3 = 1, 3.2 - 3.1 = 1/10, 3.15 - 3.14 = 1/100, 3.142 - 3.141 = 1/1000\ldots$. Thus by taking sufficiently many decimal places, the difference between the overestimate of $\pi$ and the underestimate of $\pi$ can be made as small as we wish.\footnote{Archimedes (287-212 BCE) used regular 96-gons to approximate the circumference of a circle, thus showing that $3\frac{1}{7} < \pi < 3\frac{10}{71}$.}

Truncations also appear in another interpretation of (non-repeating) decimals, using infinite series. For example, we think of the decimal expansion for $\pi$ as an infinite series:

$$\pi = 3 + \frac{1}{10} + \frac{4}{100} + \frac{1}{1,000} + \frac{5}{10,000} + \frac{9}{100,000} + \ldots$$

and the value of this infinite series is (by definition) the limit of the sequence of partial sums

$$3, \quad 3 + \frac{1}{10}, \quad 3 + \frac{1}{10} + \frac{4}{100}, \quad 3 + \frac{1}{10} + \frac{4}{100} + \frac{1}{1,000}, \quad 3 + \frac{1}{10} + \frac{4}{100} + \frac{1}{1,000} + \frac{5}{10,000} \quad \ldots$$

which, if we switch over to decimal notation, is the sequence of (ever longer) truncated decimals for $\pi$:

$$3, \ 3.1, \ 3.14, \ 3.141, \ 3.1415, \ 3.14159, \ \ldots$$

4.2. Some technicalities. There are two main questions we should be thinking about:

Question 1: Why does every decimal give rise to a real number?

Question 2: Why can every (nonnegative\footnote{To avoid notational difficulties, we will restrict our attention to nonnegative real numbers. We all know how to write negative real numbers as decimals too.}) real number be written as a decimal?
We’ve hinted at these questions in the last section, but now is the time to settle them. Along the way, we will encounter one of the classic stumbling blocks for students: Why is $0.\overline{9} = 1$?

First let’s set up some notation for decimals:

**Definition 20.** A *decimal* is a formal expression $B.a_1a_2a_3\ldots$, where $B$ is the base-ten numeral for a nonnegative integer, and where $\{a_k\}^\infty_{k=1}$ is a sequence with each $a_k$ being one of the numerals in $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

We will now answer Question 1, *Why does every decimal give rise to a real number?* As we outlined in the last section, this is tied up with infinite series and with powers of ten. Starting with the decimal expression, $B.a_1a_2a_3\ldots$, we associate to it the real number$^{18}$

$$B + \sum_{k=1}^{\infty} \frac{a_k}{10^k}$$

It is worth considering why such an infinite series actually sums to a real number. After all, in Calculus, one encounters infinite series such as $1 + 1 + 1 + \ldots$ and $\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$ that do not sum to real numbers, whereas other series like $\frac{1}{1} + \frac{1}{3} + \frac{1}{9} + \frac{1}{27} + \cdots = \frac{1}{2}$ sum to a real number. Let’s recall the relevant theorem: *a series with nonnegative terms converges if the sequence of partial sums is bounded.*

So, we must think why “the partial sums of a decimal” are bounded. Our long experience tells us that a decimal is bounded by “one plus the whole number part,” e.g. $38.498636047\cdots \leq 39$. The “address model” for a decimal supports this. We will now show it rigorously. We just need one lemma:

**Lemma 21.** If $a_k \in \{0, 1, \ldots, 9\}$ then the series

$$\sum_{k=1}^{\infty} \frac{a_k}{10^k} = \frac{a_1}{10} + \frac{a_2}{10^2} + \cdots$$

converges to a real number in the interval $[0, 1]$.

**Proof.** First observe that the series

$$0.\overline{9} = \sum_{k=1}^{\infty} \frac{9}{10^k}$$

$^{18}$To avoid driving ourselves crazy, we are blurring the distinction between number and numeral here.
is a geometric series with first term $9/10$ and common ratio $1/10$. Therefore, this series converges to

$$\frac{\frac{9}{10}}{1 - \frac{1}{10}} = 1.$$ 

Since $0 \leq a_k \leq 9$ for each $k$, by the comparison test for non-negative termed series we conclude that the original series converges, and its value lies in the interval $[0, 1]$. □

Therefore, since $0.a_1a_2a_3\ldots a_N \leq 1$ for all $N$, we have that $B.a_1a_2a_3\ldots a_N \leq 1 + B$ for all $N$. This says that the sequence of partial sums for the decimal is bounded by $1 + B$. So, we have shown conclusively that every decimal gives rise to a real number.

Let’s move on to Question 2: Why can every real number be written as a decimal? Intuitively, if we picture real numbers as points on a line, we see how to obtain a decimal expansion by the “subdivision process” as we have explained above. Let’s write this down rigorously:

**Theorem 22.** Each real number possesses a decimal representation.

**Proof.** It suffices to prove the theorem for real numbers in the interval $[0, 1)$ (why?). Let $x$ be such a number. By partitioning $[0, 1)$ into subintervals of length $1/10$, we choose $a_1 \in \{0, 1, \ldots, 9\}$ with $a_1/10 \leq x < (a_1 + 1)/10$ (see Figure 7).

**Figure 7.** Subdivision into tenths

Then $x - \frac{a_1}{10}$ lies in $[0, 1/10)$, which in turn may be partitioned into subintervals of length $1/10^2$. So, there exists $a_2 \in \{0, 1, \ldots, 9\}$ satisfying $a_2/10^2 \leq (x - a_1/10) < (a_2 + 1)/10^2$ (see Figure 8).

**Figure 8.** Subdivision into hundredths
Then \( x - \frac{a_1}{10} - \frac{a_2}{10^2} \) lies in the interval \([0, 1/10^2)\). Continuing on with this procedure, we produce a sequence \( \{a_k\} \) satisfying

\[
0 \leq x - \sum_{k=1}^{n} \frac{a_k}{10^k} < \frac{1}{10^n}.
\]

We claim that \(0.a_1a_2\ldots\) is a decimal representation for \(x\). We must verify that the series

\[
\sum_{k=1}^{\infty} \frac{a_k}{10^k}
\]

converges to \(x\). That is, we must show that \( \lim_{n \to \infty} \left(x - \sum_{k=1}^{n} \frac{a_k}{10^k}\right) = 0 \). Fortunately, this follows immediately by applying the “squeeze theorem” from Calculus to Equation (2), that is, by considering limits as \(n \to \infty\).

Some useful notions about decimal representations may be gleaned from the proof of the previous theorem. First, the proof of the theorem gives an algorithm by which we may construct decimal representations. Representations constructed as such will be referred to as standard decimal representations.

Of course, if we are bothering to introduce a “standard” decimal representation of a real number, you should suspect that some real numbers have a “nonstandard” decimal representation. Indeed, you already know an example. Let \(x = 1\). If one follows the recipe above, the standard decimal representation is \(1.\overline{0}\), which most of us would write more simply as the terminating decimal \(1\). However, we have also shown that \(0.\overline{9}\) is a decimal representation for \(1\) (it’s just not the “standard” one).

It turns out that other “nonstandard” decimals are slight variations on this theme. First, we should observe that any positive number with a terminating decimal expansion can also be written as a non-terminating decimal, with repeating 9’s (this is what we will mean by a nonstandard decimal representation). For example, \(38.72 = 38.71\overline{9}\). However, if a number does not have a terminating decimal, then that expansion is the unique decimal for the number:

**Proposition 23.** Let \(x\) be a real number, and suppose that \(x\) does not have a terminating decimal expansion. Then the decimal expansion for \(x\) is unique.

**Proof.** We may assume that \(x \in [0, 1)\) (why?). Suppose \(0.a_1a_2\ldots\) and \(0.b_1b_2\ldots\) are two decimal expansions for \(x\). For a contradiction, we suppose that these expansions are distinct, and so there exists
a first positive integer \( N \) for which \( a_N \neq b_N \); suppose that \( a_N > b_N \). After multiplying both sides of the equation by a power of ten, we know

\[
a_1 \cdots a_N a_{N+1} a_{N+2} \cdots = b_1 \cdots b_N b_{N+1} b_{N+2} \cdots,
\]
and so subtraction gives

\[
(a_N - b_N) + 0.a_{N+1} a_{N+2} \cdots = 0.b_{N+1} b_{N+2} \cdots.
\]

Since \( a_N - b_N \geq 1 \), the left side lies in \([1, \infty)\), while Lemma 21 implies that the right side of Equation 3 lies in \([0,1]\). Thus, both sides of Equation 3 must equal one. However, since \( (a_N - b_N) \geq 1 \), it follows that \( 0.a_{N+1} a_{N+2} \cdots = 0 \), and so \( a_k = 0 \) for \( k > N \). This contradicts the fact that \( x \) does not have a terminating decimal expansion; therefore \( x \) has a unique decimal representation. \( \square \)

4.3. Rational versus irrational numbers and their decimal expansions. In this section, we recall how intrinsic information about a real number—specifically, whether it is rational—can be gleaned from its decimal expansion.

4.3.1. The flavors of decimals. From our experience, we know that decimals come in several flavors. The easiest ones are terminating decimals like 36.2074. They are conceptually easy, since we don’t have to know about infinite series to understand them. Moreover, as we learned before high school, it’s easy to add or multiply terminating decimals with a simple modification of the algorithms for adding or multiplying Hindu-Arabic numerals.

However, even a simple fraction like \( \frac{1}{3} \) fails to have a terminating decimal, which is fair warning that we have to deal with fairly complicated decimals even for everyday fractions. Now, a special feature of the decimal for \( \frac{1}{3} \) is that it is repeating. Specifically, a repeating decimal is a decimal whose tail is a finite block of one or more digits that are repeated endlessly. We are familiar with handy notation for repeating decimals, involving a bar placed over the repeating block. Some examples of repeating decimals are 35.59878787\( \cdots \) = 35.5\( \overline{987} \), 0.33333\( \cdots \) = 0.3, and 5984.698469846\( \cdots \) = 5984.\( \overline{6984} \). A terminating decimal is a special kind of repeating decimal, namely, one with a repeating block of zeroes, since we can regard a terminating decimal like 35.748 as shorthand for the repeating decimal 35.748\( \overline{0} \). On the other hand, it’s easy to create non-repeating decimals—a simply example would be 0.101001000100001000001\( \cdots \), which by
virtue of the ever-longer strings of zeroes, clearly can’t be written with a repeating block. It at least has
the property that its decimal expansion is cooked up following a simple recipe. A more typical example of a
non-repeating decimal would be the decimal expansion of \( \pi \), which does not follow any transparent pattern.

The main content of this section is that a real number is rational exactly when its decimal expansion is
repeating. In addition, we’ll discover criteria on the number that give it a terminating decimal expansion,
and we’ll come to an understanding of the length of the repeating block of a repeating decimal.

4.3.2. Terminating decimals. Let’s begin with an obvious point: Any terminating decimal represents a
rational number. For example, \( 35.728 = \frac{35728}{1000} \), and since this number is written as the ratio of two natural
numbers, it is a rational number. One can say much more than the fact that 35.728 is rational: it can be
written as a fraction where the denominator is a power of ten. It’s clear that any terminating decimal can
be written in such a way too. Conversely, any ratio of natural numbers in which the denominator is a power
of ten gives rise to a terminating decimal, by reversing the process. For example, the fraction \( \frac{36}{10000} \) gives
the same rational number as the terminating decimal 0.0036.

Thus, a real number has a terminating decimal expansion exactly when it has a representation as a
fraction whose denominator is a power of 10. Ultimately, we will need a slight variant of this criterion (see
Exercise 25):

**Proposition 24.** A positive real number has a terminating decimal expansion if and only if it can be
written as a fraction in which every prime divisor of the denominator is either 2 or 5.

4.3.3. Repeating decimals. The main fact about repeating decimals is the following:

**Theorem 25.** A positive real number has a repeating decimal expansion if and only if it is rational.

The ideas behind this theorem are surprisingly easy. Let’s first try to understand why rational numbers
have repeating decimals. Consider the rational number \( \frac{21}{34} \). In Figure 9, we extract a decimal that represents
\( \frac{21}{34} \).

Note that each step, a remainder is obtained; the list of these remainders is

\[ 21, 6, 26, 22, 16, 24, 2, 20, 30, 28, 8, 12, 18, 10, 32, 14, 4, 6, 26, 22 \ldots \]
We note that 6 occurs as a remainder (the second one), and occurs again. The key observation is that when a remainder occurs for a second time, the algorithm repeats, and a repeating block must occur in the decimal expansion. We conclude that

\[
\frac{21}{34} = 0.61764705882352941
\]

Why must this happen for every rational number \( \frac{a}{b} \)? We imagine the long-division algorithm being carried out. Each place after the decimal point produces a remainder. The remainder is at least 0 and less
than $b$. If a remainder of 0 ever occurs, then $\frac{a}{b}$ has a terminating decimal. If a remainder of zero never occurs, then all possible remainders are elements of the set $\{1, 2, 3, \ldots, b-1\}$. After $b$ iterations, a total of $b$ remainders will have been obtained, which means that the same remainder must have occurred twice by this point. Hence $\frac{a}{b}$ must have a repeating decimal expansion.

Conversely, why does every repeating decimal necessarily represent a rational number? Here, there is a handy algorithm for converting the repeating decimal into a fraction. Let’s illustrate the algorithm with the repeating decimal $x = 1.647806780678067806\cdots = 1.647806$. Something interesting happens when we multiply $x$ by $10^4$ (the fourth power of ten, which we choose since the repeating block has length four). Note that $10^4 x = 16478.06780678\cdots = 16478.067806$; then when we compute the difference $10^4 x - x$, we obtain $(10^4 - 1)x = 16,476.42$, which is a terminating decimal! Finally, if we multiply by $10^2$ (this comes from the two places to the right of the decimal point in $x$ before the repeating block starts), we have that

$$10^2(10^4 - 1)x = 1,647,642$$

which is a whole number. Hence $x = \frac{1647642}{10^2(10^4 - 1)} = \frac{1647642}{999999900}$. Since this is the ratio of natural numbers, it is a (positive) rational number. Students often practice this algorithm using the standard pencil-and-paper algorithm for subtraction, as in Figure 10. A version of this procedure is indicated in Your Turn 33, and is crucial for the following section.

**Figure 10.** Converting a fraction to a repeating decimal

**Your Turn 33.** Consider a particularly simple repeating decimal: one that is zero to the left of the decimal point, has $r$ zeros immediately to the right of the decimal point, and then has a repeating block of length $l$. Show that the decimal is given by $\frac{B}{10^r(10^l - 1)}$, where $B$ is the repeating block (an $l$-digit numeral).
4. DECIMALS

4.3.4. The length of the repeating block. Above, we showed that the decimal expansion of \( \frac{a}{b} \) is a repeating decimal, and the length of the repeating block is at most \( b - 1 \). In general, the length of the repeating block is shorter. In this section, we’ll give a description of its length. We’ll also be able to say how long the decimal continues to the right of the decimal point before the repeating block is reached.

Let’s first clarify the ideas about repeating decimals. Consider the decimal 45.6981273273273273 \( \cdots \) = 45.6981273. We will say that the decimal can be written as a repeating decimal whose repeating block begins after 4 digits to the right of the decimal point (in the notation that follows, we’ll say \( r = 4 \)), and we’ll say that the decimal can be written with a repeating block of length 3, that is, \( l = 3 \). As will be important later, note that the decimal is the sum of two other decimals:

\[
45.6981273 = 45.6981 + 0.0000273.
\]

The first is a terminating decimal, and the second is exactly the sort in Your Turn 33. Thus we have a quick way of turning the decimal into a fraction:

\[
45.6981273 = 45.6981 + 0.0000273 = \frac{456981}{10^4} + \frac{273}{10^4(10^3 - 1)}
\]

which simplifies to \( \frac{456981.9999}{9990000} + \frac{273}{9990000} = \frac{456524292}{9990000} \). In simplest form, this fraction is \( \frac{48043691}{832500} \).

Note here that \( r = 4 \) and \( l = 3 \) represent the minimal choices of \( r \) and \( l \). We can view the decimal as having bigger values of \( r \) and \( l \): for example, the decimal can be written (less efficiently) as 45.6981273273273273, and if we parse the decimal that way, we have \( r = 6 \) and \( l = 15 \). Of course, ultimately we will be interested in the smallest values of \( r \) and \( l \) that make sense for the decimal.

Now we move on to some notation involving fractions. Suppose that \( \frac{a}{b} \) is a positive fraction written in simplest form (meaning, the greatest common divisor of \( a \) and \( b \) is 1). Consider the prime factorization of \( a \) and \( b \). The prime 2 could be a factor of \( a \), or \( b \), or neither (but it can’t be a factor of both \( a \) and \( b \)). The same goes for the prime 5. We are going to be concerned about the number of times that either 2 or 5 occurs in the denominator of the fraction.

Thus, given a rational number \( x \), we are going to write \( x \) as \( \frac{k}{x^{\prime}, y^{\prime}, m} \), where we assume that the fraction is in simplest form (so \( m \) is divisible by neither 2 nor 5, and neither 2 nor 5 occurs in both numerator and
denominator). So, if \( r_1 \geq 1 \), then 2 is not a factor of \( k \), and the same goes for \( r_2 \) and 5. (If either 2 or 5 occurs in the numerator, as a divisor of \( k \), then we will have no special reason to pull it out as a factor.)

As an example, for the simplest form fraction \( \frac{38043691}{832500} \), when we peel off the factors of 2 and 5 in the denominator, we have

\[
\frac{38043691}{832500} = \frac{38043691}{2^2 \cdot 5^4 \cdot 333},
\]

so in our notation, \( k = 38043691 \), \( r_1 = 2 \), \( r_2 = 4 \), and \( m = 333 \).

Recall that if \( m = 1 \), then \( x \) has a decimal expansion that terminates (this is Proposition 24).

We are now ready to give the big theorem:

**Theorem 26.** Suppose that \( x = \frac{k}{2^{r_1}5^{r_2}m} \) is given as above. Let \( r \) and \( l \) be integers with \( r \geq 0 \) and \( l \geq 1 \). Then the following statements about \( x \), \( r \), and \( l \) are equivalent:

1. The fraction \( x \) has a decimal expansion beginning after \( r \) places to the right of the decimal point, and with repeating block of length \( l \).
2. \( 10^r \cdot (10^l - 1) \cdot x \) is an integer.
3. \( r \geq r_1 \), \( r \geq r_2 \), and \( m \) is a divisor of \( 10^l - 1 \).

We are mainly interested in the corollary:

**Corollary 27.** Take notation as in Theorem 26. In the most efficient way of writing \( x \) as a decimal, the decimal repeats after \( \max\{r_1, r_2\} \) places to the right of the decimal point, and the length of the repeating block is the smallest positive integer \( l \) such that \( 10^l - 1 \) is a multiple of \( m \).

So, working backwards in the example we’ve discussed, consider the fraction \( x = \frac{38043691}{832500} \). According to the corollary, in the most efficient parsing of the decimal expansion for \( x \), the decimal should repeat after \( \max\{2, 4\} = 4 \) places to the right of the decimal point. To find the length of the repeating block, we must find the first \( l \) such that 333 divides \( 10^l - 1 \). Clearly this is \( l = 3 \), since 333 divides 999 but not 99 or 9. This gives us complete information about the minimal \( r \) and \( l \) (and of course, it agrees with the explicit decimal expansion we had for \( x \)).

To end our discussion of decimals, we give a proof of Theorem 26:
First, we prove that (1) implies (2). By (1), we can write the decimal for $x$ as $c.a_1a_2\ldots a_rb_1b_2\ldots b_r$, where the $a$s and the $b$s are digits, and $c$ is the Hindu-Arabic base ten numeral for a whole number. It follows that

$$10^r \cdot (10^l - 1) \cdot x = 10^{r+l} \cdot x - 10^r \cdot x = ca_1a_2\ldots a_rb_1b_2\ldots b_l - ca_1a_2\ldots a_rb_1b_2\ldots b_l = ca_1a_2\ldots a_rb_1b_2\ldots b_l - ca_1a_2\ldots a_r,$$

which is an integer, showing (2). (In hindsight, we have played the usual game for turning a repeating decimal into a fraction: we multiply by some powers of ten and subtract, eliminating everything to the right of the decimal point.)

Next, we show that (2) implies (3). With the assumption (2), suppose that $10^r \cdot (10^l - 1) \cdot x$ is the integer $y$. This tells us that

$$y = \frac{10^r \cdot k(10^l - 1)}{2^r5^r} = \frac{2^r5^r \cdot k(10^l - 1)}{2^r5^r}.$$

By assumption, $y$ is a whole number. How can this happen?

Let’s first think about powers of 2. On the one hand, if $r_1 = 0$, then we have $r \geq r_1$ for free. On the other hand, suppose that $r_1 \geq 1$. Then, there must be at least $r_1$ factors of 2 in the numerator. Now $k$ is no help (recall that if $r_1 \geq 1$, then there are no factors of 2 in $k$), and likewise, there are no factors of 2 in $10^l - 1$ (why?). The only way out is for $r \geq r_1$, as desired. By a similar argument involving factors of 5, we can conclude that $r \geq r_2$. What about the factor of $m$ in the denominator? We already know that $m$ has no factors of 2 or 5, hence for $y$ to be a whole number, it must happen that $k(10^l - 1)$ is a multiple of $m$. However, the greatest common divisor of $k$ and $m$ is 1. So, it must happen that $10^l - 1$ is a multiple of $m$. We have proved (3).

Finally, we show that (3) implies (1). By (3), we have that $10^l - 1 = ms$ for some integer $s$. Hence

$$x = \frac{k}{2^r5^r m} = \frac{2^{r-r_1}5^{r-r_2}k}{10^r m} = \frac{2^{r-r_1}5^{r-r_2}sk}{10^r(10^l - 1)}.$$

By (3), we know that those powers of 2 and 5 are whole numbers (why?). Let’s rename the whole numerator as $z$ (which is a whole number!). The upshot is that $x = \frac{z}{10^r(10^l - 1)}$. 
We now apply the Division Algorithm to \( z \) and \( 10^l - 1 \). There must exist an integer \( Q \), and an integer \( R \) with \( 0 \leq R < 10^l - 1 \) (meaning, as a base ten numeral, \( R \) has at most \( l \) digits), and such that \( z = Q(10^l - 1) + R \), so

\[
x = \frac{Q}{10^r} + \frac{R}{10^r(10^l - 1)}.
\]

This should remind us very much of the way we wrote \( 45.6981273 \) as \( 45.6981 + 0.0000273 \). Indeed, \( x = \frac{Q}{10^r} \) is a terminating decimal that extends \( r \) places to the right of the decimal point. On the other hand, according to Your Turn 33, \( \frac{R}{10^r(10^l - 1)} \) is one of those “simple” repeating decimals—it’s merely a repeating block of the \( l \) digits\(^{19} \) of the number \( R \), starting after \( r \) digits to the right of the decimal place. This proves (1), since we’ve shown that \( x \) has a decimal expansion of the required form.

### 4.4. Exercises.

**Pedagogical exercises:**

1. What meanings or representations can you attach to the decimal symbol 3.74? How are these related to the number line, to fractions, or to a real-world notion?

2. When grading your students’ homework, you discover that one of your students believes that 0.0047 > 0.03. What might be the reasons for his confusion? What strategies might you employ to help him better understand decimals?

3. In the text, we used the formula for the sum of a geometric series to show that 0.5 = 1. Find at least one other argument that would be appropriate for high school students.

4. Use your calculator to compute a decimal approximation for \( \frac{21}{34} \). Can you tell from this approximation that the decimal expansion of \( \frac{21}{34} \) is a repeating decimal? How could you use your calculator to obtain the full repeating block (with certainty)?


\(^{19} \) If \( R \) happens to have fewer than \( l \) digits, we have to stick zeros as the first part of the block, as in the decimal 0.0000008 = 0.000008008008 \ldots.
(a) Use Corollary 27 to find the length \( l \) of the repeating block in the decimal expansions of each fraction. Also find the number of places \( r \) after the decimal point after which the repeating block begins. \((A \text{ calculator will be helpful in using Corollary 27 to find } l.)\)

(b) Use a calculator to obtain decimal approximations of each fraction. Is it easy to detect the repeating block?

(c) Use a computer algebra system to obtain decimal approximations of each fraction. Choose sufficiently many digits so that the repeating blocks (whose lengths you know from (a)) are apparent.

6. Carefully explain the role of the digit 0 in the decimal system. Why is it necessary?

7. Which is bigger, \( \frac{9}{11} \) or \( \frac{5}{6} \)? Which is bigger, \( \frac{744}{999} \) or 0.744? Which is bigger, 0.474 or 0.47? Compare the strategies you used in answering these three questions.

8. How is a meter stick helpful in obtaining the decimal representation for a length (in meters)? Compare this with measurement using a yardstick.

9. A student is converting a fraction to a repeating decimal. He computes the first five digits of the decimal as 0.53813. Since a “three” has recurred, he concludes that the decimal is 0.5318. Is he correct? Why or why not?

10. A student evaluates \( \frac{15561}{20007} \) on a calculator and obtains 0.7777777778. She concludes that \( \frac{15561}{20007} = 0.7 \). A second student evaluates \( \frac{7777777782}{100000000100} \) and gets 0.777777777. She concludes that \( \frac{7777777782}{100000000100} = 0.7 \) too. Is either student correct in her answer? Should either student have been confident in her conclusion?

11. Recall the standard algorithm for multiplying two terminating decimals: \emph{First, count the total number \( n \) of nonzero digits to the right of the decimal point in both numbers. Then, multiply the numbers as if they were both integers and insert a decimal point \( n \) places to the left of the last digit in the product.}

Compute 2.47 \( \times \) 0.28 using the standard algorithm. Then compute the product by converting each number to a fraction and multiplying the fractions. Use these computations to explain why the standard algorithm works in this case.
12. Recall that in the standard algorithm for dividing two terminating decimals, one may move the divisor’s decimal point in any way we choose, provided that we move the dividend’s decimal point by a like amount. Compute \(21.34 \div 0.76\) by using the standard algorithm. This will involve computing the equivalent quotient \(2134 \div 76\). By converting the numbers to fractions, explain why these two quotients are equivalent.

13. After Mrs. D. has carefully explained why \(0.9 = 1\), Jimmie raises his hand and conjectures, “So you mean \(0.7 = 0.8\)?”

(a) Why might Jimmie conclude that \(0.7 = 0.8\)?

(b) Is Jimmie correct or incorrect? How would you respond to Jimmie to provide mathematical insight?

14. Find a repeating, nonterminating decimal that begins as \(0.33333333333333333333\) (twenty 3s) but is not equal to \(0.3\). Then, write the decimal as a fraction.

15. Consider the terminating decimal \(0.a_1 \ldots a_n\) and let \(a_1 \ldots a_n\) denote the base 10 representation of the number \(a_110^{n-1} + a_210^{n-2} + \cdots + a_n10^{0}\). Carefully show that \(\frac{a_1 \ldots a_n}{10^n} = 0.a_1 \ldots a_n\).

16. It is known that \(\pi\) is approximately equal to \(3.1415926535897932385\). Using this, find the best decimal approximation you can for \(2\pi\). How large might your error be?

17. Estimate the product \(12.345578 \times 8.1243\) by truncating each multiplicand to three decimal places to the right of the decimal point (use a calculator!). Then, give a bound on the error incurred in your approximation. How might you obtain an exact (non-approximate) answer to this multiplication problem?

Basic computations:

18. Write each of the following terminating decimals as a rational number (integer divided by an integer).
You need not write the fraction in simplest form.

(a) 4.2

(b) 758

(c) 0.0001

(d) 365.112498
19. Find a terminating decimal for each of the following fractions. (*Replace the fraction with an equivalent fraction whose denominator is a power of ten. Don’t simply use a calculator to perform the division.*)

(a) \( \frac{15}{16} \)

(b) \( \frac{14}{125} \)

(c) \( \frac{63}{80} \)

(d) \( \frac{1234}{5} \)

(e) \( \frac{1}{5} \)

20. For each decimal \( x \) written below,

- Rewrite \( x \) in notation that shows the repeating blocks (e.g. rewrite \( x = 1.647806780678067806... \) as \( \frac{1647806}{999999} \)). You may assume that the “obvious” repeating pattern occurs.
- By inspecting the decimal, find the minimal values of \( r \) and \( l \) (as in the text).
- Write \( x \) as a rational number (integer divided by an integer).
- Find the whole number \( 10^r(10^l - 1)x \).

(a) 0.8333333333333333333333333...

(b) 0.7777777777777777777777777...

(c) 350.79429429429429429429429...

(d) 4.5793755755755755755755755...

(e) 1324.2432432432432432432432...

(f) 4.19999999999999999999999...

(g) 4.2000000000000000000000000...

(h) 758.0000000000000000000000000...

In (g) and (h), you will write the repeating blocks in two different ways. Which decimal better illustrates the minimal values of \( r \) and \( l \), the one with repeating zeros or the one with repeating nines?

21. For each fraction in the list below:

- Use Corollary 27 to find the minimum values of \( l \) and \( r \).
Report the decimal approximation to $\frac{4}{5}$ that your calculator gives. In which cases could you make a reasonable guess about the values of $r$ and $l$, purely from your calculator’s decimal approximation?

Use the standard algorithm for dividing whole numbers to find the repeating decimal expansion for each fraction $\frac{a}{b}$. Be certain to write your answer with an overbar to indicate the repeating blocks.

(a) $\frac{1}{6}$
(b) $\frac{59}{72}$
(c) $\frac{2}{7}$
(d) $\frac{5893}{7}$
(e) $\frac{1}{9}$
(f) $\frac{1}{9999}$
(g) $\frac{1}{99999}$
(h) $\frac{33338}{99999}$

22. For each fraction in Exercise 21, use Corollary 27 to find the length of the repeating block in the decimal expansion.

**Terminating decimals:**

23. Separate the fractions $\frac{1}{32}$, $\frac{3}{14}$, $\frac{22}{352}$, $\frac{18}{168}$, $\frac{77}{28}$, and $\frac{21}{384}$ into two categories: those having terminating decimal expansions and those that do not. Do not use a calculator.

24. For each terminating decimal, find a non-terminating decimal that gives the same real number.

(a) $45.67$
(b) $456700$
(c) $0.009$
(d) $359699$

25. Show that a rational number can be written as a fraction whose denominator is a power of ten if and only if it can be written as a fraction whose denominator is a power of 2 times a power of 5.
Additional exercises:

26. What strategies can you find for converting $\frac{0.0007}{0.006}$ to a fraction in simplest form? Which of these strategies are effective for $\frac{0.0007}{0.006}$?

27. Let $a$ and $b$ be positive integers whose product is 9999.

(a) Find a specific example of $a$ and $b$. With this choice, compute the decimal representations of $\frac{1}{a}$ and $\frac{1}{b}$. Explain how these decimals are related to $b$ and $a$.

(b) Give a clear mathematical explanation of the phenomenon you observed in (a).

28. Let $\frac{a}{b}$ be a fraction in simplest form, and assume that $b$ is divisible by neither 2 nor 5. What are all possible choices of $b$ such that the repeating decimal for $\frac{a}{b}$ has a repeating block of length 1? Then, answer the same question for repeating blocks of length 2, 3, and 4. (Interpret the question as asking for the minimal length of a repeating block, so that (for example) $\frac{1}{9} = 0.\overline{1}$ has repeating block of length 1 but not of length 2.)

29. Let $x$ and $y$ be real numbers with decimal representations $0.a_1a_2\ldots$ and $0.b_1b_2\ldots$, respectively. Devise conditions on $\{a_k\}$ and $\{b_k\}$ such that $x < y$ if and only if your conditions hold.

30. Change the conditions on $a_k$ given in Equation (2), so that one obtains the decimal representation 0.09999 for the number $\frac{1}{10}$.

31. Let $x = 0.\overline{72}$.

(a) Find a repeating decimal $y$, whose repeating block has length 2, such that the length of the repeating block in $x + y$ is two.

(b) Find a repeating decimal $y$, whose repeating block has length 2, such that the length of the repeating block in $x + y$ is one.

(c) Find a repeating decimal $y$, whose repeating block has length 2, such that $x + y$ can be written as a terminating decimal.
32. Suppose that $x$ and $y$ are decimals whose repeating blocks have length 2 and 3, respectively. What possibilities exist for the length of the repeating block of $x + y$?

33. Just as whole numbers can written in bases other than ten, the same is true for real numbers. In this exercise, we explore decimals written in base twelve.

(a) How would you interpret $7T.3E604$ as a sum? How would you physically produce a point along the number line from this decimal? (Here $T$ stands for ten and $E$ stands for eleven.)

(b) Which real numbers have terminating decimal expansions in base twelve? Why?

(c) Which real numbers have repeating decimal expansions in base twelve? Why?

(d) Find the (base twelve) decimal expansions for $1/E$ (one eleventh), $2/E$, ..., $T/E$, and explain several strategies for finding them.

(e) Convert the repeating base-twelve decimal $7.839$ to a fraction (written in base twelve).

(f) Conjecture analogues of Your Turn 33 and Corollary 27.

5. Algebraic and Transcendental Numbers

5.1. Definitions and examples. If you take a course in abstract algebra or number theory, you may learn that there are many number systems besides $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. In particular, a great deal is known about certain number systems that contain $\mathbb{Q}$ and are contained in $\mathbb{C}$. The text and exercises in this section give a first glimpse at the richness of these systems.

**Definition 28.** A real (or complex) number $\alpha$ is said to be algebraic if $\alpha$ is a root of a nonconstant polynomial with integer coefficients. Otherwise, we say that $\alpha$ is transcendental.\(^{20}\)

Observe that every rational number $a/b$ is necessarily algebraic, since $a/b$ is a root of the polynomial $bx - a = 0$, which has integer coefficients. In addition, many irrational numbers are algebraic. For example, $\sqrt{2}$ and $\sqrt{1 + \sqrt{3}}$ are both algebraic since they satisfy the equations $x^2 - 2 = 0$ and $x^4 - 2x^2 - 2 = 0$, respectively.

\(^{20}\)This terminology is due to Euler. In his words, non-algebraic numbers are to be called transcendental, because they “transcend the power of algebraic methods.”
Your Turn 34. Verify that \( \sqrt{1+\sqrt{3}} \) is a root of \( x^4 - 2x^2 - 2 = 0 \).

Thus the set of algebraic numbers contains the set of rational numbers as a proper subset. But where are the transcendental numbers? Perhaps the most famous transcendental numbers are \( e \) and \( \pi \). While the proofs that \( \pi \) and \( e \) are transcendental are beyond the scope of this book, we will show that transcendental numbers exist, and we construct specific examples. A key tool will be the use of decimals to represent numbers.

5.2. Liouville’s Theorem: explicit construction of some transcendental numbers. In this section, we give a method for producing transcendental numbers.

**Definition 29.** Given an algebraic number \( \alpha \) and a positive integer \( n \), we say that \( \alpha \) is of degree \( n \) if

(i) The number \( \alpha \) is a root of a polynomial of degree \( n \) with integer coefficients.

(ii) The number \( \alpha \) is not a root of any polynomial with integer coefficients of degree less than \( n \).

For example, \( \sqrt{3} \) is an algebraic number of degree 2, since \( \sqrt{3} \) satisfies \( x^2 - 3 = 0 \), yet \( \sqrt{3} \) is not a zero of any degree 1 polynomial with integer coefficients (why?).

**Lemma 30.** Let \( \alpha \) be a real number. There exists a sequence \( \{\alpha_m\} = \{p_m/q_m\} \) of rational numbers converging to \( \alpha \) with \( \lim_{m \to \infty} q_m = +\infty \).

**Proof.** We use truncations of the decimal expansion of \( \alpha \). Assume \( \alpha \in [0,1) \). By Theorem 22 from Section 4.2, \( \alpha \) has a standard decimal expansion \( 0.a_1a_2\ldots \). Put \( \alpha_m = 0.a_1\ldots a_m \). By equation (2) from the same section,

\[
|\alpha - \alpha_m| = |\alpha - 0.a_1\ldots a_m| < \frac{1}{10^m}.
\]

21These special numbers were shown to be transcendental in the late 19th century by Charles Hermite (1822-1901) and Ferdinand Lindemann (1852-1939), respectively. The fact that \( \pi \) is transcendental is of great historical significance, having played a major role in resolving a famous 2300 year old mathematical problem, namely: Is it possible to “square the circle” via straightedge and compass using Platonic rules? In this problem, given a circle, one attempts to construct a square of equal area via straightedge and compass while adhering to strict rules regarding the use of these tools. This essentially boils down to the problem of constructing a segment of length \( \sqrt{\pi} \) from a given segment of length 1. Algebraically, this last problem is solvable if and only if \( \pi \) is a zero of a certain type of polynomial with integer coefficients. Therefore, since \( \pi \) is transcendental, it is not the zero of any polynomial with integer coefficients, and so it is not possible to square a circle.
from which it follows that \( \{\alpha_m\} \) converges to \( \alpha \). Further, by Exercise 15, one may write

\[
\alpha_m = \frac{a_110^{m-1} + a_210^{m-2} + \cdots + a_m10^0}{10^m}.
\]

Let \( p_m \) and \( q_m \) denote the numerator and denominator of the right-hand side of (4), respectively; clearly \( q_m \) tends to +\( \infty \) as \( m \) gets large. \( \square \)

**Your Turn 35.** Modify the proof to accommodate the case \( \alpha \geq 1 \).

**Theorem 31.** (Liouville) Let \( \alpha \in \mathbb{R} \) be an algebraic number of degree \( n > 1 \) and let \( \{\alpha_m\} = \{p_m/q_m\} \) be a sequence of rational numbers converging to \( \alpha \) as in Lemma 30. There exists a positive integer \( n \) such that if \( m > n \), then

\[
|\alpha - p_m/q_m| > \frac{1}{(q_m)^{n+1}};
\]

Before proving Theorem 31, we demonstrate its usefulness in the construction of transcendental numbers. For example, consider the real number

\[
\alpha = \sum_{k=1}^{\infty} 10^{-k!} = 10^{-1!} + 10^{-2!} + 10^{-3!} + \cdots.
\]

The decimal representation of \( \alpha \) is

\[
\alpha = 0.1100010000000000000000001000\ldots,
\]

with an increasing number of zeros between nonzero digits. Since the decimal does not repeat, we conclude by Theorem 25 that \( \alpha \) must be irrational. We claim even more: that \( \alpha \) is transcendental. For a contradiction, assume that \( \alpha \) is algebraic of degree \( n \), and let

\[
\alpha_m = \sum_{k=1}^{m} 10^{-k!} = 10^{-1!} + \cdots + 10^{-m!},
\]

which may be written as \( \frac{p_m}{10^m} \) for some integer \( p_m \) (why?). The sequence \( \{\alpha_m\} \) satisfies the hypotheses of Lemma 30, and so by Liouville’s Theorem

\[
|\alpha - \alpha_m| > \frac{1}{10^{(n+1)m!}}
\]

for sufficiently large \( m \). However, by subtracting (6) from (5) one obtains

\[
|\alpha - \alpha_m| < \frac{10}{10^{(m+1)!}} \quad \text{(why?)},
\]
Combining Equations (7) and (8) gives
\[
\frac{1}{10^{(n+1)m!}} < \frac{1}{10^{(m+1)!-1}},
\]
so that \((n+1)m! > (m+1)! - 1\) for all sufficiently large \(m\). However, as you will show in Exercise 8, this is false for \(m\) larger than \(n\), giving a contradiction. Therefore \(\alpha\) is transcendental.

To prove Theorem 31, we will need the Factor Theorem for polynomials with rational coefficients:

**Theorem 32.** Suppose \(r \in \mathbb{Q}\) and \(f(x)\) is a nonconstant polynomial with rational coefficients. The number \(r\) is a zero of \(f(x)\) if and only if there exists a polynomial \(g(x)\) with rational coefficients such that \(f(x) = (x - r)g(x)\).

Observe \(f(x) = (x - r)g(x)\) implies that the degree of \(g(x)\) is one less than the degree of \(f(x)\).

We now sketch the proof of Liouville’s theorem (Theorem 31):

**Proof.** Suppose that \(\alpha\) (a real algebraic number of degree \(n\)) is a root of the polynomial \(f(x) = a_nx^n + \cdots + a_1x + a_0\) (with integer coefficients). As a first order of business, we claim that \(f(\alpha_m) \neq 0\) for each \(m\). For a contradiction, suppose \(f(\alpha_m) = 0\) for some \(m\). Since \(\alpha_m\) is rational, the Factor Theorem tells us that \(f(x) = (x - \alpha_m)g(x)\) for some polynomial \(g(x)\) of degree \(n - 1\) with rational coefficients. Further, since \(f(\alpha) = 0\) and \(\alpha \neq \alpha_m\) (why?), we conclude that \(g(\alpha) = 0\). However, this contradicts the fact that \(\alpha\) is algebraic of degree \(n\) (see Definition 29 and Exercise 7). We conclude that \(f(\alpha_m) \neq 0\) for all positive integers \(m\).

Now we attend to the inequality in Liouville’s Theorem. We may write
\[
f(\alpha_m) = f(\alpha_m) - 0 = f(\alpha_m) - f(\alpha) = a_1(\alpha_m - \alpha) + a_2(\alpha_m^{2} - \alpha^2) + \cdots + a_n(\alpha_m^n - \alpha^n).
\]
Using the fact that
\[
\frac{x^n - y^n}{x - y} = x^{n-1} + x^{n-2}y + x^{n-3}y^2 + \cdots + xy^{n-2} + y^{n-1},
\]
and assuming that \(m\) is sufficiently large to guarantee \(|\alpha_m - \alpha| < 1\), one obtains (via the Triangle Inequality and Exercise 9)

\[
\frac{|f(\alpha_m)|}{|\alpha_m - \alpha|} = |a_1 + a_2(\alpha_m + \alpha) + a_3(\alpha_2^2 + \alpha_m\alpha + \alpha^2) + \cdots + a_n(\alpha_m^{n-1} + \cdots + \alpha^{n-1})|
\]
\[
< |a_1| + 2|a_2|(|\alpha| + 1) + 3|a_3|(|\alpha| + 1)^2 + \cdots + n|a_n|(|\alpha| + 1)^{n-1}.
\]
If we let \( M \) denote the quantity given in (10), and we let \( m \) be large enough so that \( q_m > M \), then we have

\[
|\alpha_m - \alpha| > \frac{|f(\alpha_m)|}{M} > \frac{|f(\alpha_m)|}{q_m}.
\]

By writing \( \alpha_m = p_m/q_m \) and substituting into the rule for \( f(x) \), we obtain

\[
f(\alpha_m) = \frac{a_0q_m^n + a_1q_m^{n-1}p_m + \cdots + a_np_m^n}{q_m^n},
\]

where the numerator is a nonzero integer since \( f(\alpha_m) \) is nonzero. Therefore, in combination with Equation (11), one obtains

\[
|\alpha - \alpha_m| > \frac{|a_0q_m^n + a_1q_m^{n-1}p_m + \cdots + a_np_m^n|}{q_m^{n+1}} \geq \frac{1}{q_m^{n+1}},
\]

concluding the proof. \( \square \)

5.3. Exercises.

1. Let \( n, a \) be positive integers. Show that \( n\sqrt{a} \) is algebraic.

2. Give an example of a transcendental number that is not real.

3. Show that the following real numbers \( \alpha \) are algebraic by finding polynomial functions \( f(x) \) with integer coefficients and positive degree satisfying \( f(\alpha) = 0 \).
   (a) \( \alpha = \sqrt{2} + \sqrt{3} \)
   (b) \( \alpha = \sqrt{1 + \sqrt{2}} \)
   (c) \( \alpha = \sqrt{3} - \sqrt{6} \)

4. Show that if \( \alpha \) is algebraic, then so is \(-\alpha\). (If you know a polynomial \( p(x) \) with integer coefficients that has \( \alpha \) as a root, how must you modify \( p(x) \) to obtain a polynomial that has \(-\alpha \) as a root?)

5. Show that if \( \alpha \) is algebraic (and nonzero), then so is \( 1/\alpha \).

6. In the text, we argued that every rational number is algebraic. Does the argument work for the rational number zero? Why or why not?
7. Formulate new definitions from Definitions 28 and 29 by replacing all instances of the word “integer” with the word “rational”. Show that these new definitions are equivalent to the original definitions. *(To show the equivalence of Definition 28 and its modification, show that a number is algebraic according to the original definition if and only if it is algebraic according to the modified definition.)*

8. Prove that \((n + 1)m! \leq (m + 1)! - 1\) whenever \(m > n\).

9. Verify equation (9) and inequality (10).

10. Show that if \(\alpha\) and \(\beta\) are algebraic, then so are \(\alpha + \beta\) and \(\alpha\beta\). *(This exercise probably requires some experience in abstract algebra, and you may want to look up the solution in an abstract algebra or number theory book. Note that in the language of abstract algebra, this exercise (together with Exercises 4 and 5) shows that the set of algebraic numbers forms a subfield of the complex numbers.)*

11. *This exercise is intended to give a taste of the issues involving factorization that arise in abstract algebra. Let \(S = \{a + bi\sqrt{5} : a, b \in \mathbb{Z}\} \subset \mathbb{C}\).*

   (a) Show that the additive inverse of any element of \(S\) is again an element of \(S\).

   (b) Show that the sum and the product of any two elements of \(S\) are again elements of \(S\).

   (c) Regarding \(\mathbb{C}\) as a plane, how would you describe elements of \(S\) (in graphical terms)? We define \(N : S \to \mathbb{Z}, N(a + bi\sqrt{5}) = a^2 + 5b^2\).

   (d) Prove that for any elements \(s_1\) and \(s_2\) in \(S\), we have \(N(s_1s_2) = N(s_1)N(s_2)\).

   (e) List all the elements \(s \in S\) such that \(0 \leq N(s) \leq 9\).

We now are ready to illustrate a profound peculiarity of factorization in \(S\). Note that \(6 = 6 + 0i\sqrt{5} \in S\) can be factored in two different ways, namely as \(6 = 2 \cdot 3\) and as \(6 = (1 + i\sqrt{5}) \cdot (1 - i\sqrt{5})\). In itself, this is not surprising. For example, in \(\mathbb{Z}\), we know that 60 can be factored as \(4 \cdot 15\) but also as \(6 \cdot 10\). Of course, 4, 15, 6, and 10 can be factored further: we have \(4 \cdot 15 = (2 \cdot 2) \cdot (3 \cdot 5) = 2^23^15^1\), and \(6 \cdot 10 = (2 \cdot 3) \cdot (2 \cdot 5) = 2^23^15^1\) as well. Hence the two factorizations \(4 \cdot 15\) and \(6 \cdot 10\) can be refined to yield a common factorization of 60.

   (f) In contrast, show that in \(S\), the two factorizations \(6 = 2 \cdot 3\) and \(6 = (1 + i\sqrt{5}) \cdot (1 - i\sqrt{5})\) cannot be refined to give a common factorization of 6. *(Your answers to (d) and (e) will help.)*
12. The Gaussian integers are defined to be \( \{a + bi : a, b \in \mathbb{Z}\} \subset \mathbb{C} \). While this set might seem very similar to the set \( S \) defined in 11, it has remarkably different properties.

Consult an abstract algebra or number theory text that discusses the Gaussian integers. Using the text, write a report that discusses factorization in the Gaussian integers. In particular, are there examples of the phenomenon in Exercise 11(f)? (Your research may lead you to the topic of Euclidean rings, an axiomatic system inspired by properties the integers.)

6. Exercises Involving Student Work: Fractions and Decimals

For the Problem below,

- Complete the given task yourself. Compare your responses with those of a partner or small group.
- Read through the student responses provided.
- Comment on the quality of each student response. Identify ways in which the student’s thinking is correct and ways in which the student’s thinking is incorrect or incomplete.

One way to compare fractions is to convert them into decimals. For instance, comparing \( \frac{3}{8} \) and \( \frac{2}{5} \) can be done by first converting \( \frac{3}{8} \) to 0.375 and \( \frac{2}{5} \) to 0.4 and then comparing the decimals.

Suppose that Mary is attempting to compare the fractions \( \frac{9443}{9444} \) and \( \frac{9444}{9445} \) using this method. When she converts each of these fractions to its decimal form, the number 0.9998941 is displayed on her calculator. Therefore, she concludes that the two fractions are equal.

(a) Is Mary’s conclusion correct?

(b) If Mary’s conclusion is correct, explain why it is correct. If Mary’s conclusion is not correct, explain why it is incorrect and identify which fraction is larger and why.
CHAPTER 9

Exponentiation

The use of *exponents* is a recurring theme in secondary mathematics. As we begin our expedition into exponentiation, let's take stock of what we know\(^1\).

**Let’s Go 1.** *Consider the following expressions of the form \(z^w\).*

(A) *For each expression, if you know what its value should be, give its value and explain why.*

(B) *Use a scientific calculator to obtain a value (or approximation) for each expression. (Most scientific calculators can handle complex numbers.)*

\[
\begin{align*}
(a) & \; 9^2 \\
(e) & \; 9^{5/2} \\
(i) & \; 0^0 \\
(m) & \; i^3 \\
(q) & \; (-1)^{1/2} \\
(u) & \; (-1 + 0i)^{1/3} \\
(y) & \; (i^i)^i \\
(b) & \; 9^{-2} \\
(f) & \; 9^0 \\
(j) & \; 2^{1/2} \\
(n) & \; i^4 \\
(r) & \; (-1)^{1/4} \\
(v) & \; i^{1/2} \\
(z) & \; (2^i)^i \\
(c) & \; 9^{1/2} \\
(g) & \; 0^3 \\
(k) & \; e^\pi \\
(o) & \; i^{-1} \\
(s) & \; (-1)^{1/3} \\
(w) & \; e^{i\pi/2} \\
(d) & \; 9^{-1/2} \\
(h) & \; 0^{-2} \\
(l) & \; e^{-\pi/2} \\
(p) & \; (1 + i)^3 \\
(t) & \; (-1)^{2/6} \\
(x) & \; i^i \\
\end{align*}
\]

Early in our education we learn that

\[4^3 = 4 \times 4 \times 4 = 64,\]

while years later we may encounter another equation involving exponents, the famous *Euler Formula:*

\[e^{\pi i} = -1.\]

On the face of it, Euler’s Formula is utterly baffling. We understand why \(4^3 = 4 \times 4 \times 4\) because we know how to multiply a number with itself three times, but what does it mean to multiply a number with itself \(\pi\) times, let alone \(\pi i\) times? How can the left side of Euler’s Formula involve non-real complex numbers, if the right side is a real number?

\(^1\)This chapter will clarify items (a)–(p) of Let’s Go 1. The key ideas lying behind items (q)–(z) will be developed in Chapter 11.
We don’t learn $4^3 = 64$ on Day One and then $e^{\pi i} = -1$ on Day Two; there is a long journey between these two experiences of exponentiation. Ultimately, a goal of the next three chapters is to define and make sense of the expression $z^w$ where $z$ and $w$ can be any complex numbers. This is a long process, proceeding through special cases ($w$ is a whole number, then an integer, then the reciprocal of a whole number, then a fraction, then a real number, and finally a complex number). Some of the steps will be very familiar to you, while others will be new. We will see that when the base $z$ is a positive real number, everything works out much more easily than in the general case—this is a theme of the present chapter.

Exponentiation is interesting and important because it is connected to so many of the ideas in the secondary curriculum and beyond. Here are some of the highlights:

- Exponentiation produces some of the most important functions in the secondary curriculum. For example, if the base is a variable and the exponent is a constant, we obtain the power functions that send $x$ to $x^2, x^3, x^{-1}$ and $x^{1/3}$. When the base is a constant, we obtain the exponential functions such as $x \mapsto 2^x$ and $x \mapsto e^x$.

- Pairs of inverse functions occur in the context of exponentiation, e.g. $f(x) = x^3$ and $g(x) = x^{1/3}$, and $f(x) = 2^x$ and $g(x) = \log_2 x$.

- Functions like $x \mapsto x^{1/2}$ and $x \mapsto x^{1/3}$ are related to our notion of roots (e.g. square roots and cube roots, sometimes denoted $\sqrt{x}$ and $\sqrt[3]{x}$), and also to the more general notion of the “root” of a polynomial.

- Exponentiation is intimately connected with trigonometry. This is a profound fact, though it is not transparent to students at the high school level. For instance we have

$$e^{i\theta} = \cos \theta + i \sin \theta$$

which is the full statement of Euler’s Formula. In particular, when $\theta = \pi$, we recover $e^{i\pi} = -1$.

- Power series, as studied in calculus, are a surprisingly helpful tool for studying exponential functions. Using power series, the connections among exponential, trigonometric, and hyperbolic-trigonometric functions become clear.

---

\footnote{Surprisingly, the only choice of $z$ for which this can’t be done is $z = 0$. We will be able to define $0^w$ for relatively few choices of $w$.}
• **Complex numbers** are important for clarifying and unifying some of the disparate topics in exponentiation, giving easy geometric pictures of square roots, cube roots, etc. Also, the power series connections among the exponential and trigonometric functions require extending these functions to complex-valued functions of a complex variable.

• **Logarithmic functions** are already familiar to us, as inverse functions of exponential functions, and as integrals (recall that $\ln x = \int_1^x \frac{dt}{t}$).

In defining $z^w$ in various settings, we will be guided by three familiar properties of exponentiation which are valid (as you will check in Your Turn 2) when the exponents are whole numbers:

**Property 1:** $z^a \cdot z^b = z^{a+b}$

**Property 2:** $(z^a)^b = z^{ab}$

**Property 3:** $z^a u^a = (zu)^a$

As we’ll see in subsequent sections, each time we define $z^w$ for “more choices of $w$,” we try to make the definition so that some or all of these properties are valid. Simply by trying to make these properties hold, we will arrive at a definition of $z^w$ for any (nonzero) complex number $z$ and any real number $w$. Keeping track of which of the three properties remain valid will be of great interest to us.

As you can see from our “catalog of ideas related to exponentiation” above, exponentiation is a large topic. In this text, the ideas are developed over three chapters.

Chapter 9 deals with the cases that are of greatest relevance to high school mathematics: integer exponents (with any complex number as a base); and real exponents (with any positive real number as a base). In the context of the real numbers, we study concepts that are central to the secondary mathematics curriculum: power, root, logarithm, and the functions that rely on these concepts. Our treatment includes complex

---

3. In high school, we learn that “odd roots” behave differently from “even roots.” For example, there is no real twelfth root of a negative number, but there are two twelfth roots of a positive number. On the other hand, there is exactly one eleventh root of any real number (positive or negative). When one allows complex numbers as roots, this even/odd distinction disappears.

4. The situation where $w$ is complex requires more than these three properties. Specifically, we will use power series to motivate the definition of $z^w$ when $w$ is not real.
numbers only when they do not add any difficulties or complications, and the discussion is accomplished without any reliance on calculus.

In contrast, the ideas of calculus come to the fore in Chapter 10, where we see that key applications of exponential functions (such as compound interest) ultimately rely on derivative and limit facts about \( e^x \); even the historical impetus for logarithms (as an aid to computation) is connected to the definition of \( \ln x \) as \( \int_1^x \frac{dt}{t} \).

The more difficult constructions in exponentiation (particularly, the use of complex numbers as exponents) are examined in Chapter 11. Therein, the intimate connections among the exponential, trigonometric, and hyperbolic trigonometric functions emerge.

1. Whole Number Exponents

1.1. Shorthand for repeated multiplication. Let \( z \) be any complex number and let \( n \) be any positive integer. The expression \( z^n \) is defined to be the product \( z \cdot z \cdots z \) with \( n \) factors of \( z \). We refer to \( z \) as the base and \( n \) as the exponent.

Theorem 1. The three properties of exponents listed in the introduction are true when the exponents are positive integers. That is, if \( u, z \) are complex numbers and \( n, m \) are positive integers, then
\[
\begin{align*}
(z^n \cdot z^m) &= z^{n+m} \\
(z^n)^m &= z^{nm} \\
(zu)^n &= z^n u^n.
\end{align*}
\]

Your Turn 1. Suppose you ask a student to explain why \( 3^2 \cdot 3^5 = 3^7 \), and she responds, “Both sides are equal to 2187.” Do you consider this an explanation? What explanation would you prefer?

Your Turn 2. Devise simple but convincing explanations that would show a beginning algebra student the three properties of exponentiation are valid (or that might help them discover the properties). Compare your examples with those created by classmates.

The three Properties of Exponentiation listed in the introduction will be referenced many times in this chapter. We emphasize that at present, we have only asserted their validity when the bases are complex numbers and the exponents are positive integers.
1.2. The Functions $f(z) = z^n$. Now that we’ve defined $z^n$ for a complex number $z$ and a positive integer $n$, we can regard $z^n$ as a function of the variable $z$ (with a fixed choice of $n$). In high school algebra, one usually only allows $z$ to be a real variable, and we’ll recall this situation now. (We will address the more general situation—where $z$ can be any complex number—in Chapter 11, Section 1.)

Your Turn 3. For a fixed positive integer $n$, consider the function $f : \mathbb{R} \to \mathbb{R}, f(x) = x^n$. For $n$ between 1 and 6, the graphs are presented in a scrambled order in Figure 1. Label each of the graphs in Figure 1 with the correct function. Give a rationale for your answer.

Figure 1. Graphs of $y = x^n$
We observe that the graphs are qualitatively different depending on whether $n$ is even or odd, as is illustrated in the table below. In attempting to list the differences, we all eventually would arrive at the descriptions summarized in the table:

<table>
<thead>
<tr>
<th>$n$</th>
<th>$x^n$ odd</th>
<th>$x^n$ even</th>
<th>$x^n$ one-to-one</th>
<th>$x^n$ increasing</th>
<th>$x^n$ decreasing</th>
<th>range of $x^n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>even</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>$[0, \infty)$</td>
<td>$(-\infty, 0]$</td>
<td>$\mathbb{R}$</td>
</tr>
<tr>
<td>odd</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>$\mathbb{R}$</td>
<td>never</td>
<td>$[0, \infty)$</td>
</tr>
</tbody>
</table>

Exercises 11 through 14 explore aspects of this table.

### 1.3. Exercises.

1. Compute $(2 + 5i)^3$.

2. In this exercise we investigate notation for repeated exponentiation. Consider a complex number $z$. Suppose we raise $z$ to the $n$th power, raise that number to the $n$th power, then raise that number to the $n$th power, et cetera. Suppose that this exponentiation is done a total of $k$ times. What is the resulting number? (You should be able to express the answer with very concise notation.)

3. Let $a$, $b$, and $c$ be positive integers. Is it necessarily true that $a^{(b^c)} = (a^b)^c$? Give an explanation that a beginning algebra student could understand.

4. Sometimes we read $z \cdot z \cdot \cdots \cdot z$ as “the product of $z$ with itself $n$ times.” Unfortunately, students may misinterpret these words. For example, what might a student be thinking if he (incorrectly) interprets $5^4$ as 25 instead of 5? Or if he computes $3^4$ as $(3 \cdot 3) \cdot (3 \cdot 3) \cdot (3 \cdot 3)$?

5. Describe a real-world situation that makes use of repeated multiplication.

6. Give proofs of the three properties of exponentiation.

*Exercises 7 and 8 require the polar decomposition of complex numbers.*

7. If $z$ is a real number and $n$ is a positive integer, then clearly $z^n$ is a real number. However, it’s possible for $z^n$ to be real, even if $z$ is not. We explore this:
1. WHOLE NUMBER EXPONENTS

(a) Which complex numbers \( z \) have the property that \( z^2 \) is a positive real number? Which complex numbers \( z \) have the property that \( z^2 \) is a negative real number? Draw pictures to explain. (You should think about the “geometric rule” for multiplication of complex numbers: multiply lengths and add angles.)

(b) Same question, with \( z^3 \) instead of \( z^2 \).

(c) Same question, with \( z^4 \).

(d) Same question, with \( z^5 \).

(e) Give a general answer for \( z^n \), where \( n \) is any positive integer.

8. In this exercise, we consider the sequence \( \{z^n : n \in \mathbb{N}\} \) for a fixed real or complex number \( z \), and investigate properties of this sequence. A calculator or computer algebra system may be useful in this exercise.

(a) On the real line, plot \( \{1.5^1, 1.5^2, 1.5^3, 1.5^4, 1.5^5\} \). What is \( \lim_{n \to \infty} 1.5^n \)?

(b) On the real line, plot \( \{(-0.5)^1, (-0.5)^2, (-0.5)^3, (-0.5)^4\} \). What is \( \lim_{n \to \infty} (-0.5)^n \)?

(c) In the complex plane, plot \( \{(-1 + i)^1, (-1 + i)^2, (-1 + i)^3, (-1 + i)^4\} \). (You may find it easier to write \((-1 + i)\) in polar form first.) What is the behavior of \((-1 + i)^n\) as \( n \) increases?

(d) For each complex number in the set \( \{(-1 + i)^n\} \), we may compute the radian measure of the angle that the number makes with the positive real axis. Show that only finitely many different angle measures actually occur. We may also compute the absolute value of each complex number in \( \{(-1 + i)^n\} \). The absolute value will be a nonnegative real number. Do finitely many or infinitely many numbers occur?

(e) Find a non-real complex number \( z \) such that all numbers in the set \( \{z^n : n \in \mathbb{N}\} \) have the same modulus. Explain your answer. (Recall the the modulus of a complex number \( z = a + bi \) is \( |z| = \sqrt{a^2 + b^2} \).)

(f) Find a complex number \( z \) such that in the set \( \{z^n : n \in \mathbb{N}\} \), infinitely many different angles occur. Explain.
9. Consider the graphs of \( y = x^4 \) and \( y = x^6 \) (where \( x \) is a real variable). At what values of \( x \) do the graphs intersect? On what intervals is \( x^4 > x^6 \) and on what intervals is \( x^6 > x^4 \)? Give convincing explanations of your answers.

10. Generalize the last exercise to the graphs of \( y = x^n \) and \( y = x^m \), where \( m, n \) are positive integers.

11. Using the definitions of odd and even functions, prove that \( f(x) = x^n \) is odd when \( n \) is an odd positive integer, and that \( f(x) \) is even when \( n \) is an even positive integer.

12. Show that \( x^n \) is increasing on the interval \([0, \infty)\) (regardless of whether \( n \) is even or odd). In other words, you must show that if \( 0 \leq x_1 \leq x_2 \), then \( x_1^n \leq x_2^n \). (You might try induction on \( n \).)

13. We explore issues that arise in finding the range of \( x^n \). (Here the domain is the set of real numbers.)

(a) Show that \( \lim_{x \to \infty} x^n = \infty \).

(b) In finding the range of the function \( x^n \), one needs more than \( \lim_{x \to \infty} x^n = \infty \). For example, consider the function \( f(x) = x^2 \). We know that \( f(0) = 0 \) and \( f(10^3) = 10^6 \). How does this show that the range of \( f \) contains every number in the interval \([0, 10^6]\)? (The key is a theorem you learned in calculus.)

14. Give an example of a non-polynomial function \( g(x) \) that shares many of the characteristics of \( x^n \) when \( n \) is even. Specifically, find a non-polynomial function \( g(x) \) that is even, that is increasing on \([0, \infty)\), and whose range is \([0, \infty)\).

2. Integer Exponents

The main goal of this section is to discover that the three properties of exponents provide the motivation for using negative exponents to describe reciprocals. We also explore the fact that \( 0^n \) is not defined for \( n = 0, -1, -2, \ldots \).

2.1. Non-positive exponents. If \( z \) is a nonzero complex number and \( n \) is a positive integer, then

\[
(1) \quad z^0 = 1 \quad \text{and} \quad z^{-n} = \frac{1}{z^n}.
\]
Equation (1) extends our collection of exponents from the positive integers to the entire set of integers, but what motivates the notation in (1)? Discover the answer in the following Your Turn:

**Your Turn 4.** Consider the three properties of exponents listed in the chapter introduction.

(a) What motivation can you find for defining \( z^0 \) to be equal to 1 when \( z \) is a nonzero complex number? (Consider \( z^0 \cdot z^1 \), and take for granted that you want Property 1 to hold.)

(b) In your solution to (a), why was it important that \( z \) not be zero?

(c) Which property of exponentiation motivates the definition of \( z^{-1} \)? (Consider the product \( z^n \cdot z^{-n} \).)

---

### 2.2. Properties of exponentiation with integer exponents.

We have used the first property of exponentiation to motivate the definition of \( z^n \) where \( n \) is negative. But do the properties of exponentiation hold for integer exponents? The answer is yes:

**Theorem 2. Properties of Exponentiation with Integer Exponents**

1. For any nonzero complex number \( z \) and any integers \( n \) and \( m \), we have \( z^n \cdot z^m = z^{n+m} \).
2. For any nonzero complex number \( z \) and any integers \( n \) and \( m \), we have \( (z^n)^m = z^{nm} \).
3. For any nonzero complex numbers \( z \) and \( u \) and any integer \( n \), we have \( (zu)^n = z^n u^n \).

Theorem 2 is somewhat tedious to prove because it involves a lot of special cases depending on the signs of \( m \) and \( n \). We omit the proof and ask that you instead consider the specific examples given in Your Turn 5 below.

**Your Turn 5.** Prove each of the following special cases of Theorem 2. Don’t simply cite Theorem 2 itself. However, you can make use of the properties for positive integer exponents (see Theorem 1 from Section 1) and basic facts about fractions. Compare your proofs with those of classmates. Assume that \( z, u \) are nonzero complex numbers.

(a) \( z^{-2} \cdot z^{-3} = z^{-5} \)

(b) \( z^{-8} \cdot z^3 = z^{-5} \)

(c) \( z^{-4} \cdot z^7 = z^3 \)
(d) \((z^{-2})^3 = z^{-6}\)
(e) \((z^2)^{-3} = z^{-6}\)
(f) \((z^{-2})^{-3} = z^6\)
(g) \((zu)^{-3} = z^{-3}u^{-3}\)

2.3. Why not use zero as a base? Our definition of \(z^n\) assumed that \(z\) was not zero. What is special about the case where the base is zero? Let’s start with \(0^0\):

Your Turn 6. Consider \(0^0 \cdot 0^1 = 0^{0+1}\) and take for granted that Property 1 holds. Why doesn’t this yield any ideas about how to define \(0^0\)?

After further experimentation, one quickly discovers that none of the properties of exponentiation force any particular choice for \(0^0\). Thus \(0^0\) is not defined.

Likewise, zero raised to a negative integer power is not defined. We see this, for example, by considering the first property of exponentiation. If \(0^{-1}\) were defined, we would expect \(0^{-1} \cdot 0^1 = 0^0\) to be a valid equation. The left side would be zero, since it’s a number times 0. But this is a contradiction, since the right side is \(0^0\) which is not defined!

2.4. Exercises.

1. In the text, we explored the graphs of \(f(x) = x^n\) where \(n\) is a positive integer. We now recall the graphs for negative integer exponents.

(a) By hand, graph the functions \(x^{-1}\), \(x^{-2}\), \(x^{-3}\), and \(x^{-4}\). Use the nonzero real numbers as the domain.

(b) Check your answers using a graphing calculator or computer software.

(c) Consider the function \(f(x) = x^n\), where \(n\) is a negative integer. Make a conjecture on the values of \(\lim_{x \to \pm\infty} x^n\), \(\lim_{x \to -\infty} x^n\), \(\lim_{x \to 0^+} x^n\), and \(\lim_{x \to 0^-} x^n\).

(d) What is the range of \(f\)?

---

5All of this should spark a memory from calculus, where \(0^0\) is called an indeterminate form. The idea there was that given functions \(f\) and \(g\) with \(\lim_{x \to a} f(x) = 0\) and \(\lim_{x \to a} g(x) = 0\), then this is not enough information to conclude anything about \(\lim_{x \to a} (f(x)^{g(x)})\). For example, if \(f(x) = 0\) and \(g(x) = x\), then \(\lim_{x \to 0^+} f(x)^{g(x)} = 0\), whereas if \(f(x) = x\) and \(g(x) = 0\), then \(\lim_{x \to 0^+} f(x)^{g(x)} = 1\). So our ideas about indeterminate forms give further insight into why \(0^0\) is not defined.
2. In this exercise, we’ll consider mathematical and notational issues that arise with negative exponents.

(a) Write \((7 + \sqrt{2})^{-1}\) as a fraction. Then try to write it in the form \(a + b\sqrt{2}\), where \(a\) and \(b\) are rational.

(You may have learned this procedure under the name “rationalizing the denominator.”)

(b) Write \((7 + \sqrt{2})^{-2}\) in the form \(A + B\sqrt{2}\), where \(A\) and \(B\) are rational.

(c) Write \((4 + 13i)^{-1}\) in the form \(a + bi\), where \(a\) and \(b\) are real (actually, rational).

(d) Do the same for \((4 + 13i)^{-2}\).

(e) How are parts (b) and (d) similar? Explain as precisely as you can.

3. We’ve motivated the definition of \(z^n\) (for \(n\) a nonpositive integer) by the properties of exponentiation. Another good motivation is through geometric progressions. Let \(z\) be a nonzero complex number.

(a) Explain why \(z^1, z^2, z^3, \ldots\) is a geometric progression.

(b) Suppose that we want \(\ldots z^{-3}, z^{-2}, z^{-1}, z^0, z^1, z^2, z^3, \ldots\) also to be a geometric progression. Explain how this forces the values of \(\ldots z^{-3}, z^{-2}, z^{-1}, z^0\) to be fixed, and how this gives the same definition for \(z^n\) as in the text.

(c) Why does this strategy fail when \(z = 0\)?

3. Rational Exponents

The goal of this section is to give meaning to expressions such as \(\sqrt[3]{z}, z^{1/3}\), and \(z^{p/q}\) in certain limited settings. We will concentrate on roots that are real, with small detours into the complex numbers.\(^6\)

Let’s Go 2. Consider the following questions about roots.

(a) What property must a number have to be a “square root of 25”? Of what polynomial is it a root?

List all the roots of this polynomial (including nonreal roots in \(\mathbb{C}\)). Which of these roots should be called the square root of 25?

(b) Repeat part (a) for cube roots of 8.

---

\(^6\)The full story only emerges when one fully considers complex numbers and their roots. This will be discussed later, when we define and study \(z^{1/q}\) for any complex number \(z\) and for any positive integer \(q\) (see Chapter 11, Section 2).
Let’s Go 3. From high school algebra, we are accustomed to thinking of $z^{1/q}$ as representing one of the qth roots of z. Carefully explain how one of the Properties of Exponentiation makes it plausible for $z^{1/q}$ to be a qth root of z.

3.1. Roots and dueling notations: $r^{1/q}$ versus $\sqrt[q]{r}$. Let $r$ be a real number and $q$ be a positive integer. Let’s Go 2 illustrates the fact that $r$ possesses multiple q-th roots whenever $r \neq 0$ and $q \geq 2$. Humankind feels a need to pick a distinguished root from among these roots. Unfortunately, choosing a q-th root of $r$ is mainly driven by two sometimes conflicting desires:

(a) To pick a real root so that our distinguished root remains within the real numbers. In addition, if $r$ is positive we want our root to be positive.

(b) To pick a root with the smallest possible positive argument. (Recall that the argument of a complex number is essentially the angle it makes with the positive real axis.)

These desires coincide exactly when $r$ is greater than or equal to zero, and conflict when $r$ is negative. For example, when choosing a square root of 25, we wind up with 5 regardless of whether we follow item (a) or item (b). However, when choosing a cube root of $-8$ we obtain $-2$ when following item (a), but we wind up with $1 + i\sqrt{3}$ when following item (b) (see Figure 2).

To distinguish between these two types of roots we let $\sqrt[q]{r}$ denote the unique q-th root of $r$ chosen according to item (a) above, while we let $r^{1/q}$ denote the unique q-th root of $r$ chosen according to item (b) above. The meanings of the notations are equivalent when $r$ is positive and are different when $r$ is negative. We refer to $\sqrt[q]{r}$ as the real q-th root of $r$ (positive when $r$ is positive),
3. RATIONAL EXPONENTS

Figure 2. $\sqrt[3]{-8}$ and $(-8)^{1/3}$

while we call $r^{1/q}$ the principal $q$-th root of $r$. The uniqueness of $\sqrt[q]{r}$ can be illustrated by looking at the graph of $y = x^q$ (see Figure 3).

Figure 3. $\sqrt[q]{r}$ is unique

The following table gives summary information:

<table>
<thead>
<tr>
<th>$q$ even</th>
<th>$q$ odd</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r \geq 0$</td>
<td>$r^{1/q} = \sqrt[q]{r} \geq 0$</td>
</tr>
<tr>
<td>$r &lt; 0$</td>
<td>$r^{1/q}$ exists but $\sqrt[q]{r}$ does not</td>
</tr>
</tbody>
</table>

Much more on computing principal roots can be found in Chapter 11, Section 2.

3.2. Rational exponents: roots and powers of the base. Having figured out how to define expressions like $5^{1/4}$ in the last subsection, we will now see that it’s a very small step to define expressions such
as $5^{7/4}$. Our goal here is to define $z^w$ where $z$ is any positive real number and $w$ is any rational number.\footnote{To define $z^w$ when $z$ is negative (or more generally, when $z$ is any nonzero complex number that isn’t a positive real number) one needs the notion of principal $q$th root, as defined in the last subsection. We will have more to say about this in Chapter 11.}

We will see that when we choose a positive base $z$ (that is, when $z^{1/q} = \sqrt[q]{z}$) everything works beautifully.

**Your Turn 7.** Based on your secondary school experience, what are the values of $8^{2/3}$ and $81^{3/2}$? What definition are you using to perform these computations?

**Your Turn 8.** It would be problematic if $8^{2/3}$ did not equal $8^{4/6}$. What explanation can you give for their equality?

Let $r > 0$. For any fraction $\frac{p}{q}$ ($p$ being any integer and $q$ being a positive integer), we define:

**Definition 3.** $r^{p/q} = \left(r^{1/q}\right)^p$.

If $r > 0$, and $w$ is a rational number, can we define $r^w$? At first thought, it seems that we have already answered this question. After all, we can write $w$ as the fraction $w = \frac{p}{q}$ (with $q$ positive), and we have already defined $r^{\frac{p}{q}}$ to be $\left(r^{1/q}\right)^p$ in Definition 3. But how do we know that we get the same number if we write $w$ as some other fraction\footnote{Recall that a fraction $r/s$ is merely a pair of integers $r, s$ with $s \neq 0$, while a rational number is an equivalence class of fractions under the following equivalence relation: $r/s = m/n$ if and only if $rn = sm$.}, say $w = \frac{s}{t}$?

Let’s think about this in a specific example: To show that $7^{\frac{4}{10}} = 7^{\frac{6}{15}}$ our strategy is to show that both numbers are the (unique) positive 30th root of $7^{12}$. This forces the numbers to be equal. Since both $7^{\frac{4}{10}}$ and $7^{\frac{6}{15}}$ are positive (see Subsection 3.1), we only need to show that both numbers yield $7^{12}$ when raised to the 30-th power. Note

\begin{equation}
\left(7^{\frac{4}{10}}\right)^{30} = \left(\left(\frac{\sqrt[10]{7}}{1}\right)^4\right)^{30} = \left(\frac{\sqrt[10]{7}}{\sqrt[10]{7}}\right)^{120} = \left(\frac{\sqrt[10]{7}}{\sqrt[10]{7}}\right)^{12} = 7^{12}
\end{equation}

and likewise

\begin{equation}
\left(7^{\frac{6}{15}}\right)^{30} = \left(\left(\frac{\sqrt[15]{7}}{1}\right)^6\right)^{30} = \left(\frac{\sqrt[15]{7}}{\sqrt[15]{7}}\right)^{180} = \left(\frac{\sqrt[15]{7}}{\sqrt[15]{7}}\right)^{12} = 7^{12}.
\end{equation}

**Your Turn 9.**

(a) Justify each equality in Equation (2).
(b) Generalize the computation above: if \(p, q, s, t\) are integers with \(q > 0\) and \(t > 0\), if \(pt = qs\), and if \(r\) is any nonnegative real number, then \(r^{\frac{p}{q}} = r^{\frac{s}{t}}\).

Your Turn 9 ensures that the following definition makes sense:

**Definition 4.** Let \(r\) be a positive real number, let \(w\) be a rational number, and let us represent \(w\) by a fraction \(\frac{p}{q}\) where \(q > 0\). We define \(r^w = \left(r^{\frac{1}{q}}\right)^{pq} = (\sqrt[q]{r})^p\).

We’ll now consider the properties of exponentiation, and discover that all the expected properties are valid when the base is a nonnegative real number and the exponent is a rational number:

**Theorem 5.** Let \(r, s\) be positive real numbers, and let \(w_1, w_2, w\) be rational numbers. Then

\[
\begin{align*}
(1) \quad r^{w_1}r^{w_2} &= r^{w_1+w_2} \\
(2) \quad (r^{w_1})^{w_2} &= r^{w_1w_2} \\
(3) \quad (rs)^w &= r^ws^w
\end{align*}
\]

**Proof.** Our strategy is to write the rational exponents as fractions, and to exploit known facts about integer exponents.

Let’s start with property (3). We’ll write \(w\) as \(\frac{p}{q}\), where \(q > 0\). On the one hand, \((rs)^w = (\sqrt[q]{rs})^p\), and on the other hand, \(r^ws^w = (\sqrt[q]{r})^p (\sqrt[q]{s})^p\). Of course both of these numbers are positive. To finish, we’ll merely show that both are \(q\)th roots of \((rs)^p\). For the first expression of \(r^w\), we have

\[
\left((\sqrt[q]{rs})^p\right)^q = \left(\sqrt[q]{rs}\right)^{pq} = \left((\sqrt[q]{rs})^q\right)^p = (rs)^p
\]

so \((rs)^w\) indeed is the unique positive \(q\)th root of \((rs)^p\). (When we interchanged the order of \(p\) and \(q\) as exponents, we haven’t cheated at all—our calculation only used properties of exponentiation with integer exponents, which we established in Section 2.2.) The second expression is only slightly harder:

\[
\left((\sqrt[q]{r})^p (\sqrt[q]{s})^p\right)^q = \left((\sqrt[q]{r})^p\right)^q \left((\sqrt[q]{s})^p\right)^q = \left(\sqrt[q]{r}\right)^{pq} \left(\sqrt[q]{s}\right)^{pq} = \left((\sqrt[q]{r})^q\right)^p \left((\sqrt[q]{s})^q\right)^p = r^ps^p = (rs)^p.
\]

Exercise 12 asks for the proof of property (2).
Finally, we prove property (1). In order to make use of our definitions, we need to write \( w_1 \) and \( w_2 \) as fractions. Let’s try to make life as easy as possible: we can do this by finding a “common denominator” for the fractions. Hence, we can write \( w_1 = \frac{p_1}{q} \) and \( w_2 = \frac{p_2}{q} \), where \( q \) is a positive integer.

Having done this, it’s easy to prove (1):
\[
r^{w_1}r^{w_2} = r^{\frac{p_1}{q}}r^{\frac{p_2}{q}} = \left( r^{\frac{1}{q}} \right)^{p_1} \left( r^{\frac{1}{q}} \right)^{p_2} = r^{\frac{p_1+p_2}{q}} = r^{w_1+w_2}.
\]

\( \square \)

### 3.3. Exercises.

1.

(a) Find all solutions of the equation \( x^2 = 9 \) in \( \mathbb{W}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \) and \( \mathbb{C} \). Do the same for the equations \( x^2 = 10 \) and \( x^2 = -4 \). (What notation have you used to describe these solutions? Are you inclined to give exact values or numerical approximations?)

(b) Give exact values or numerical approximations of \( \sqrt{9}, \sqrt{10}, \sqrt{-4}, 9^{1/2}, 10^{1/2}, \) and \((-4)^{1/2}\). Are all of these defined?

2.  

(a) Find all the solutions of \( x^3 = \frac{8}{27} \) in \( \mathbb{W}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \) and \( \mathbb{C} \). Do the same for \( x^3 = -125 \). (What notation have you used to describe these solutions?)

(b) Give exact values of \( 3\sqrt[3]{\frac{8}{27}}, \sqrt[3]{-125}, (\frac{8}{27})^{\frac{1}{3}}, \) and \((-125)^{\frac{1}{3}}\). Are all of these defined?

3. Root functions can be viewed in the context of inverse functions, as you will now explore:

(a) Let \( f : \mathbb{R} \to \mathbb{R}, f(x) = x^3 \) and let \( g : \mathbb{R} \to \mathbb{R}, g(x) = \sqrt[3]{x} \). Prove (by computing \( f \circ g \) and \( g \circ f \)) that \( f \) and \( g \) are inverse functions. Finally, graph \( f \) and \( g \) on the same axes. What property of the graphs of inverse functions do you observe?

(b) Roughly the same is true for \( x^q \) and \( \sqrt[q]{x} \), except one must be more careful if \( q \) is even. Let \( f(x) = x^4 \) and \( g(x) = \sqrt[4]{x} \). Identify appropriate domain and range for \( f \) and \( g \) so that \( f \) and \( g \) are indeed inverse functions. Sketch the graphs of \( f \) and \( g \) (with the domains you specified).

4. If \( q \) is an odd positive integer and \( r \) is a real number, then how are \( \sqrt[q]{r} \) and \( \sqrt[q]{-r} \) related to each other? Explain.
5. Discuss why an elementary school student might believe that there is no fourth root of 7.

6. We quickly grow accustomed to the power of calculators, and may lose sight of the difficulty of some computations. For example, we know that 22.283 has a positive square root, and using the square root button on a calculator, we can effortlessly obtain a decimal approximation for $\sqrt{22.283}$.

(a) Try guessing $\sqrt{22.283}$ using only the $+ \times - \div$ buttons on your calculator. Allow yourself four guesses. Then use the square root button and check how close you are.

(b) Review Newton’s Method (check your calculus book) and use it to give an approximation of $\sqrt{22.283}$.

7. Consider $f(x) = x^2$ defined on $[1, 2] \cap \mathbb{Q}$ (i.e., defined for all rational numbers from 1 to 2). Show that $f(1) < 2 < f(2)$, but that there is no $x_0 \in [1, 2] \cap \mathbb{Q}$ with $f(x_0) = 2$. Does this contradict the Intermediate Value Theorem? Explain.

8. Carefully discuss the function $f : \mathbb{R} \to \mathbb{C}$, $f(x) = x^{\frac{1}{3}}$. (How does the function change in passing between negative and positive values? Contrast with the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \sqrt[3]{x}$.

9. In reviewing students’ work, you see that several students have made the assumption that $(x+y)^{1/2} = x^{1/2} + y^{1/2}$. Give several mathematical approaches that you could use to convince students that this equation is erroneous.

10. Your high-school student knows that $5^1 = 5$ and $5^2 = 25$, and concludes that $5^{3/2}$ is halfway between, at $\frac{5 + 25}{2} = 15$. Give several mathematically correct responses which will help him understand his error.

11.

(a) Which result in this section tells us that $\sqrt[3]{8} = \left(\sqrt[3]{8}\right)^2$?

(b) Which of the two expressions in (a) is more efficient for computing $8^{2/3}$? Why?

12. Prove property (2) in Theorem 5.

13. Suppose $r$ is a negative real number. Let $p, q, s, t$ be integers (with $q > 0$ and $t > 0$) such that $pt = qs$. Is it necessarily true that $(\sqrt[3]{r})^p = (\sqrt[3]{r})^q$? Discuss.
Let’s Go 4.

(a) Explain how you might introduce the graph of \( y = 2^x \) to high school students. With which values of \( x \) would you begin? What arguments do you make to fill in the graph?

(b) Describe how you would introduce the graph of \( y = \left(\frac{1}{2}\right)^x \) to high school students.

(c) Compare your ideas with those of your classmates. Were there substantive differences among you in your approaches to these problems?

4.1. The real exponential functions. We now are close to one of the main objectives of pre-calculus mathematics: making sense of the real exponential functions, such as \( 2^x \) and \( 10^x \) (where \( x \) is a real number). Drawing on our experience with rational exponents, we determine what is meant by \( \alpha^x \), where both \( \alpha \) is a positive real number and \( x \) is any real number, and we consider the properties of the associated exponential function \( x \mapsto \alpha^x \) with domain and codomain equal to \( \mathbb{R} \). The whole issue here is how to go from merely rational exponents to real exponents.

To accomplish our goal, we begin by trying to understand properties of the function \( x \mapsto \alpha^x \), just for \( x \) rational. None of the following facts should be surprising!

**Proposition 6.** Let \( \alpha > 0, \alpha \neq 1 \) be a real number and \( r, s \in \mathbb{Q} \). Then:

(i) \( \alpha^r \) is positive.

(ii) If \( \alpha > 1 \), then \( \alpha^r > 1 \) for \( r > 0 \).

(iii) If \( \alpha > 1 \), and \( r < s \), then \( \alpha^r < \alpha^s \). (That is, the exponential function \( \alpha^x \) is strictly increasing on \( \mathbb{Q} \) when \( \alpha > 1 \).)

**Proof.** We leave part (i) as an exercise; just try to recall the definitions!

For part (ii), since \( r > 0 \) we may write \( r = a/b \) where both \( a, b \) are positive integers. Thus \( \alpha^r = \alpha^{a/b} \).

Also, since \( a \) is a positive integer and \( \alpha > 1 \), we know \( \alpha^a > 1 \) (why?). Now, for a contradiction, suppose that \( \alpha^{a/b} < 1 \). Then \( \alpha^a = (\alpha^{a/b})^b < 1 \) (why?). But this contradicts the fact that \( \alpha^a > 1 \), so part (ii) holds.
4. REAL EXPONENTS

Finally for part (iii), since \( s - r > 0 \) we may apply part (i) together with properties of exponentiation (with rational exponents) to obtain

\[
1 < \alpha^{s-r} = \alpha^s \alpha^{-r} = \frac{\alpha^s}{\alpha^r}.
\]

Therefore \( \alpha^r < \alpha^s \).

**Your Turn 10.** Why must a strictly monotonic function (i.e., any function which is either strictly increasing or strictly decreasing) be one-to-one? After recalling why, deduce that the function \( \alpha^x \) is one-to-one on \( \mathbb{Q} \) whenever \( \alpha \) is a positive real number not equal to one.

We now move on to extending the definition of \( \alpha^x \) to the case where \( x \) might be irrational. We may think of a real number \( x \) as being the limit of some sequence \( \{r_n\} \) of rational numbers. From previous sections, we know that \( \alpha^r \) makes sense for any rational number \( r \) whenever \( \alpha > 0 \), and so it seems reasonable for us to attempt to define \( \alpha^x \) as \( \lim_{n \to \infty} \alpha^{r_n} \) (see Figure 4).

**Your Turn 11.** On your calculator, compute \( 3^{1.41} \) and \( 3^{\sqrt{2}} \). How close are the two values?

**Your Turn 12.** The development of integer and rational exponents was based on requiring the three basic properties of exponents to be true. Was this also the basis for the development of real exponents? Discuss.
In fact, this approach will work. By appealing to the graph of $\alpha^x$ when $x$ is rational, you could convince your high school students that this works. However, from a rigorous mathematical point of view, it is not so obvious that this approach will succeed. What might go wrong? Suppose that we have two distinct sequences of rational numbers $\{r_n\}$ and $\{s_n\}$ which both converge to $x \in \mathbb{R}$. In order for our notion of $\alpha^x$ to be well defined, we must prove that

$$\lim_{n \to \infty} \alpha^{r_n} = \lim_{n \to \infty} \alpha^{s_n} \text{(and they both exist!)}. \tag{3}$$

For the time being, we throw caution to the wind and take the results of Equation (3) for granted\(^{13}\). With this in hand, we have:

**Definition 7.** Let $\alpha$ and $x$ be real numbers with $\alpha > 0$. We put

$$\alpha^x = \lim_{n \to \infty} \alpha^{r_n},$$

where $\{r_n\}$ is any sequence of rational numbers converging to $x$. Further, we let $\alpha^x : \mathbb{R} \to \mathbb{R}$ be the function defined by $x \mapsto \alpha^x$.

Also, given our experience with rational exponentiation, we might expect real exponentiation (as characterized in Definition 7) to possess the same properties that rational exponentiation does, and additional ones as well.

**Theorem 8.** Let $\alpha > 0$ be a real number with $\alpha \neq 1$. Let $\alpha^x$ be as in Definition 7. We regard $x \mapsto \alpha^x$ as a function with domain and codomain $\mathbb{R}$.

(a) The range of $x \mapsto \alpha^x$ is the open interval $(0, \infty)$.

(b) The function $x \mapsto \alpha^x$ is strictly increasing if $\alpha > 1$, and strictly decreasing if $0 < \alpha < 1$.

(c) The function $x \mapsto \alpha^x$ is one to one.

(d) If $\gamma$ and $\sigma$ are real numbers, then

(i) $\alpha^\gamma \alpha^\sigma = \alpha^{\gamma + \sigma}$.

(ii) $(\alpha^\gamma)^\sigma = \alpha^{\gamma \sigma}$.

(e) The function $x \mapsto \alpha^x$ is continuous on $\mathbb{R}$.

\(^{13}\)The result of Equation 3 will be carefully dealt with in Section 6.
We omit the proof of Theorem 8 for the time being. Most parts of this theorem will be carefully addressed in Section 6.

Finally, let’s officially record what we’ve learned about the graph of \( x \mapsto \alpha^x \): the graph can have one of the following two basic shapes, depending on whether \( \alpha > 1 \) or \( 0 < \alpha < 1 \) (see Figure 5).

\[ y = \alpha^x \]

4.2. Exercises.

1. Why do we omit 1 from the possible values of \( \alpha \) in Theorem 8? Which parts of Theorem 8 are false if we allow \( \alpha = 1 \)?

2. One of the features of a linear function is that if \( x \) increases by 1 between two points on the graph, then \( y \) increases by a constant amount \( m \) (the slope of the graph), independent of the choice of \( x \).

   (a) Show that the function \( x \mapsto 2^x \) does not have this property.

   (b) Nevertheless, the function \( x \mapsto 2^x \) does have an important property. If \( x \) increases by 1 between two points on the graph of \( x \mapsto 2^x \), what happens to the \( y \)-values? (If this is murky, try skipping to the next question.)

3. Recall that a sequence of real numbers is an arithmetic sequence if the difference between consecutive terms is constant, and a sequence of (nonzero) real numbers is a geometric sequence if the ratio of consecutive terms is constant. Prove that if \( \{x_i\} \) is an arithmetic sequence, then \( \{\alpha^{x_i}\} \) is a geometric sequence (here \( \alpha \) is any positive real number).
4. Solve the following equations in the real numbers. Write your solutions in the simplest possible way. Do not use a calculator.

(a) $3^{2x} = 27$
(b) $3^x = 27$
(c) $(3^x)^2 = 27$
(d) $x^{3/2} = 27$
(e) $x^{2/3} = 27$
(f) $2^x = 16$
(g) $16^x = 2$

5. Using graphing transformations, sketch graphs of the following curves.

(a) $y = 2^{x-1} + 5$
(b) $y = 2^{-(x-1)} + 5$
(c) $y = |2^{x-1} - 4|$
(d) $y = 2^{|x-1|}$

6. Let $f(x) = \alpha^x$ for some unknown $\alpha$. Suppose that the value of $f$ is known for one value of $x$ (say, $f(c) = \lambda$). Show that this information completely determines $f$. (Solve for $\alpha$ in terms of $c$ and $\lambda$.)

7. Let $f(x) = k\alpha^x$, where $k$ and $\alpha$ are unknown. Suppose that the value of $f$ is known for two values of $x$. Show that this information completely determines $f$ (that is, one can deduce the original values of $k$ and $\alpha$).

8. State and prove analogs of items (ii) and (iii) of Proposition 6 in the case that $0 < \alpha < 1$.

9. Sketch the following graphs by hand on the interval $[-3,3]$:

(a) $f(x) = 2^x$
(b) $f(x) = 2^{-x} = 1/(2^x)$
(c) $f(x) = 2^{1/x}$ (perhaps best to graph on $[-3,0]$ and $[0,3]$ separately, with different scales on the $y$-axis)
(d) $f(x) = 2^{-1/x}$ (same suggestion)

10. Give $\lim_{x \to 0^+}$, $\lim_{x \to 0^-}$, $\lim_{x \to \infty}$, and $\lim_{x \to -\infty}$ for each of the functions in Exercise 9.

11. Consider the function $x \mapsto 2^{-1/x^2}$.

(a) Give a rough sketch by hand of its graph, paying particular attention to $\lim_{x \to 0^+}$, $\lim_{x \to 0^-}$, $\lim_{x \to \infty}$, and $\lim_{x \to -\infty}$.

(b) Using a calculator, make a table of values for $x = 0.1, 0.2, \ldots, 1$.

(c) Because of its bizarre growth properties, the function $x \mapsto 2^{-1/x^2}$ is tricky to graph using computer software. Try graphing the function with several different intervals for the domain. (Note: the graph is very flat near $x = 0$.)

12. Let $a > 1$. One of the most important features of the exponential function $a^x$ is that it increases much more rapidly than any polynomial function as $x \to \infty$. We illustrate this property with the exponential function $2^x$ and the polynomial function $x^5$.

(a) Create a table of values for $x^5$, $2^x$, and $2^x - x^5$, for integers $1 \leq k \leq 40$. You can do this with a spreadsheet, or you can use Mathematica with the following commands:

$$f[k_] := k^5; \quad g[k_] := 2^k;$$
$$t = \text{Table}\{\{f[k], g[k], g[k] - f[k]\}, \{k, 1, 40\}\};$$
$$\text{TableForm}[t, \text{TableHeadings} \to \{\text{Automatic}, \text{"k-5"}, \text{"2^k"}, \text{"2^k-k-5"}\}]$$

(b) Which value ($x^5$ or $2^x$) is greatest when $x = 40$?

(c) From the spreadsheet, it is clear that the graphs of $x \mapsto x^5$ and $x \mapsto 2^x$ cross at (at least) two places between $x = 1$ and $x = 40$. Use graphing technology to generate clear graphs of these crossing points.

(d) One might try to argue that $2^x$ is eventually greater than $x^5$ by graphing the two functions on the same axes and comparing the graphs. Try graphing the two functions on the same axes, for $1 \leq x \leq 40$. Does the graph provide better or worse information than the spreadsheet?
(e) Suppose that a student looked at the spreadsheet results and the graphs, but was not convinced that $2^x$ remains larger than $x^5$ for $x > 40$ (what’s to keep them from crossing one more time?). What mathematically convincing argument could you give to prove that $2^x > x^5$ when $x > 40$?

13. Since exponential functions increase quickly, composites of exponential functions increase very quickly.

(a) Plot the function $2^{2^x}$ on several intervals: $[0, 1]$, $[0, 5]$, $[5, 5.01]$, and $[0, 8]$.

(b) Use your plot on $[5, 5.01]$ to estimate the derivative of the function at $x = 5$.

(c) Write $2^{1024}$ in scientific notation (this is to help you realize that $2^{1024}$ is a very big number), and compute the value of $x$ such that $2^{2^x} = 2^{1024}$ (it will be quite a small number!).

5. The Real Logarithmic Functions

Leonhard Euler (1707-1783) in his *Introductio in Analysin Infinitorum* (essentially a precalculus tract), was the first to consider the logarithm as a function which is obtained as the inverse of an exponential function. In this section we establish basic facts about the logarithm following this approach. Further material on the logarithm, including both applications and history, may be found in Chapter 10.

5.1. The logarithm: definition and properties. Let $\alpha$ be a positive real number not equal to one. The function $\log_\alpha : (0, \infty) \to \mathbb{R}$, called the logarithm of base $\alpha$, is the inverse function corresponding to the exponential function $x \mapsto \alpha^x$ (with domain $\mathbb{R}$ and range $(0, \infty)$).

**Your Turn 13.** Suppose $\alpha$ is a positive real number not equal to one.

(a) *Why does the function $x \mapsto \alpha^x$ have an inverse?*

(b) *Why is $(0, \infty)$ the domain of $\log_\alpha x$? Why is $\mathbb{R}$ the range? Carefully explain your answers.*

Since the functions $f(x) = \alpha^x$ and $g(x) = \log_\alpha x$ are inverses of each other, we know that $g \circ f(x) = x$ for all $x \in \mathbb{R}$, and $f \circ g(x) = x$ for all $x > 0$. These equations read

\[
\log_\alpha(\alpha^x) = x \quad \text{for all } x \in \mathbb{R} \quad \text{and} \quad \alpha^{\log_\alpha x} = x \quad \text{for all } x > 0.
\]

Thus, the inverse relationship between exponential and logarithmic functions is conveniently expressed in the following form:
Proposition 9. (The Log-Exp Rule) Let $\alpha, \beta, \gamma$ be real numbers with $\alpha, \beta > 0$ and $\alpha$ not equal to one. We have

$$\log_\alpha \beta = \gamma \iff \alpha^\gamma = \beta.$$ 

Your Turn 14. Prove the Exp-Log Rule. (Use our observations about inverse functions.)

The Log-Exp rule makes it clear that the equation $\log_\alpha \beta = \gamma$ (a logarithmic equation) is equivalent to the equation $\alpha^\gamma = \beta$ (an exponential equation), and oftentimes the exponential equation is easy to solve. For example, the equation $\log_8(1/64) = \gamma$ is equivalent to the equation $8^\gamma = 1/64$, and based on our familiarity with powers of 8, we see that the solution is $\gamma = -2$.

When specifically attempting to compute $\log_\alpha \beta$, often we needn’t write down the exponential equation explicitly. Rather the Log-Exp rule suggests that we keep the following question in mind: To what power must $\alpha$ be raised in order to obtain $\beta$? By the Log-Exp rule, the answer to this question will be the logarithm we desire. For example, when computing $\log_8(1/64)$, we attempt to find the exponent 8 is raised to in order to achieve 1/64. Since this exponent is $-2$, we conclude that $\log_8(1/64) = -2$.

Your Turn 15. Let $\alpha$ be a positive real number not equal to one. Verify the following basic properties of the logarithm.

(a) $\log_\alpha (1) = 0$

(b) $\log_\alpha (\alpha) = 1$

The properties given in Your Turn 15 are not enough to justify close examination of logarithms. The real reason for considering logarithms is that they provide a means of converting tedious multiplication problems to easier addition problems (see part (a) of Theorem 10 below). As a result, logarithms were an indispensable tool in computation before the days of calculators\(^{14}\).

Theorem 10. Let $\alpha, \beta$ be a positive real numbers not equal to one, and let $r, s$ be positive real numbers. Then:

(a) $\log_\alpha (rs) = \log_\alpha r + \log_\alpha s$ (This is the important additive property of logarithms.)

\(^{14}\)We delve into this idea in Chapter 10.
(b) \( \log_\alpha (r/s) = \log_\alpha r - \log_\alpha s \)

(c) \( \log_\alpha (r^s) = s \log_\alpha r \)

(d) \( \log_\beta r = \frac{\log_\alpha r}{\log_\alpha \beta} \)

(e) The function \( \log_\alpha (x) \) is one to one, strictly increasing when \( \alpha > 1 \) and strictly decreasing otherwise, and is continuous on the interval \((0, \infty)\).

**Proof.** We begin with part (a). Let \( x \) and \( y \) denote \( \log_\alpha r \) and \( \log_\alpha s \), respectively. By the Log-Exp rule, we have \( \alpha^x = r \), \( \alpha^y = s \). Therefore, by the properties of real exponents together with the Log-Exp rule, we have

\[
\alpha^x \alpha^y = rs \Rightarrow \alpha^{x+y} = rs \Rightarrow x + y = \log_\alpha (rs) = \log_\alpha r + \log_\alpha s = \log_\alpha (rs).
\]

Parts (b) and (c) are similar to part (a), and are left as an exercise for the reader.

Now, let \( y = \log_\beta r \) and \( z = \log_\alpha \beta \). Then \( \alpha^z = \beta \) and \( \beta^y = r \). It follows that \( r = (\alpha^z)^y = \alpha^{zy} \), and so \( \log_\alpha r = zy = \log_\alpha \beta \log_\beta r \). Part (d) then follows by dividing through by \( \log_\alpha \beta \).

For part (e), we focus on the case \( \alpha > 1 \). Since \( \log_\alpha x \) possesses an inverse function, namely \( \alpha^x \), we know \( \log_\alpha x \) is one-to-one. Now, suppose \( r, s \) are positive real numbers with \( r < s \). Then \( s/r > 1 \), and by Lemma 15 and Exercise 8 we have \( s/r = \alpha^\gamma \) for some positive real number \( \gamma \). Therefore, by part (b) above,

\[
0 < \gamma = \log_\alpha (s/r) = \log_\alpha s - \log_\alpha r.
\]

We conclude that \( \log_\alpha x \) is strictly increasing.

Finally, we must show that \( \log_\alpha x \) is continuous on \((0, \infty)\). Let \( b \in (0, \infty) \). It suffices to show that

\[
\lim_{x \to b^+} \log_\alpha x = \log_\alpha (b) = \lim_{x \to b^-} \log_\alpha x.
\]

We focus our effort on the equation \( \lim_{x \to b^+} \log_\alpha x = \log_\alpha (b) \), and leave the other equation as an exercise. Since \( \log_\alpha x \) is strictly increasing and is bounded below by \( \log_\alpha b \) on the interval \((b, \infty)\), it follows from properties of real numbers that \( \lim_{x \to b^+} \log_\alpha x \), call it \( L \), exists with \( L \geq \log_\alpha b \). We wish to show that \( L = \log_\alpha b \). To do this, consider \( M > \log_\alpha b \). Since the range of \( \log_\alpha x \) is \( \mathbb{R} \) and \( \log_\alpha x \) is one-to-one, there exists a unique \( x_0 \in \mathbb{R} \) with \( \log_\alpha x_0 = M \). Further, since \( \log_\alpha x \) is strictly increasing, \( x_0 > b \). Thus, for all \( x \in (b, x_0) \), \( \log_\alpha x < M \), and hence \( L < M \) for any \( M > \log_\alpha b \). Therefore \( L = \log_\alpha b \).
5. THE REAL LOGARITHMIC FUNCTIONS

5.2. Exercises.

1. Here is a first attempt at producing the graph of $\log_2 x$, using properties of the function $2^x$.

   (a) Compute $2^x$ for $x = -3, -2, -1, 0, 1, 2, 3$.

   (b) Use your answer to (a) to produce seven points on the graph of the function $x \mapsto 2^x$. List these as ordered pairs.

   (c) Use your answer to (b) to produce seven points on the graph of the function $x \mapsto \log_2 x$.

   (d) Use (b) and (c) to produce rough graphs of the functions $x \mapsto 2^x$ and $x \mapsto \log_2 x$.

2. Let $\alpha$ be a positive real number not equal to one. Use properties of inverse functions, and your knowledge of the graph of $x \mapsto \alpha^x$, to sketch carefully the graph(s) of $x \mapsto \log_\alpha x$ from the graph(s) of $x \mapsto \alpha^x$. (There will be two qualitatively different cases to consider, depending on whether $\alpha > 1$ or $0 < \alpha < 1$.)

3. Compute the following logarithms without the aid of a calculator. In your work, interpret each of the logarithms by writing the related exponential equation.

   (a) $\log_{1/2} 8$

   (b) $\log_{1/2}(8^{10})$

   (c) $\log_{106} 1$

   (d) $\log_{64} 2$

4. Create several logarithm exercises like those in Exercise 3, using bases of 3, 9, 10, and $1/3$. Make certain that your exercises can be answered without use of a calculator.

5. Use the Log-Exp rule to convert exponential equations to logarithmic equations, and vice versa.

   (a) $5^{-3} = \frac{1}{125}$

   (b) $4^{t-1} = 22.381$

   (c) $\log_2 T^3 = 7$
6. Illustrate parts (a), (b), (c), and (d) of Theorem 10 with a good numerical example. In particular, make certain that all the logarithms you choose can be computed without a calculator.

7. Express \( \log_{9} x \) in terms of base-three logs. Find several justifications for your answer that would be convincing to a precalculus student.

8. Prove parts (b) and (c) of Theorem 10. (Hint for part (c): realize that by the Exp-Log Rule, your job is to show that \( a^{r \log_{a} r} = r^{s} \).

9. As in Exercise 4, solve the following equations. (What makes these equations more difficult to solve than those in Exercise 4?)

   (h) \( 2^{x} = 15 \)
   (i) \( (3/2)^{x} = 27 \)

10. Solve the following equations in the real numbers. If no solutions exist, explain why.

   (a) \( 5^{2x} - 2 \cdot 5^{x} - 3 = 0 \) (Use the quadratic formula appropriately.)
   (b) \( 2^{x} + 2^{-x} = 5 \)
   (c) \( \frac{5^{x} - 5^{-x}}{5^{x} + 5^{-x}} = 1/8 \)
   (d) \( \frac{5^{x} - 5^{-x}}{5^{x} + 5^{-x}} = 8 \)

11. A precalculus student reasons, “Since \( a^{x} \) and \( \log_{a} x \) are inverses, it must be true that \( \log_{a} x = \frac{1}{a^{x}} \).”

   (a) Explain the source of the student’s confusion.
   (b) Explain in several convincing ways that \( \log_{a} x \neq \frac{1}{a^{x}} \).

12. For each of the following expressions, first determine the values of the variables for which the expression is defined, and then write the expression in terms of a single logarithm.

   (a) \( \log_{2} \frac{2}{\sqrt{x}} - \log_{2} \sqrt{2x} \)
   (b) \( \log_{17} x - 3[\log_{17}(x - 5) + \log_{17}(x + 5)] \)
   (c) \( \frac{3}{2} \log_{3}(x^{3} - 8)^{6} - \frac{4}{5} \log_{3}(2x^{4} - 32)^{10} \)
   (d) \( \log_{9} x + \log_{3} y \)
(e) \( \log_3 z + \log_{\sqrt{3}} w \) (adopt 3 as the new base)

13. First, find the values of \( x \) for which the left sides of the following equations are defined. Second, solve the following equations, being careful to exclude extraneous solutions. Properties of both exponential and logarithmic functions may be needed.

(a) \( \log_2 (\log_2 x) = 2 \)
(b) \( (\log_3 x)^2 - \log_3 x^2 = 3 \)
(c) \( \log_4 (x + 3) + \log_4 (x - 3) = 2 \)
(d) \( \log_8 7 = \log_2 x \)
(e) \( \log_2 x = \log_9 7 \)

14. Find \( x + y + z \), given that \( \log_2 [\log_3 (\log_4 x)] = \log_3 [\log_2 (\log_4 y)] = \log_4 [\log_3 (\log_2 z)] = 0 \).

15. Let \( \alpha, \beta, x, y \) be positive real numbers, with \( \alpha, \beta, y \) not equal to one. Show that
   \[ \frac{\log_\alpha x}{\log_\alpha y} = \frac{\log_\beta x}{\log_\beta y} \]

16. Show that \( \log_{\alpha \beta} = \frac{1}{\log_\beta \alpha} \) for all \( \alpha, \beta \) that are positive real numbers not equal to 1.

17. Suppose that \( \alpha \) is a positive real number not equal to one, and that \( r, s \) are positive real numbers. Show that \( r_{\log_\alpha s} = s_{\log_\alpha r} \).

18. Let \( f(x) = \log_3 x \) and \( g(x) = \log_2 x \). In this exercise, you will show that the graphs of \( f \) and \( g \) are related to each other by a graphing transformation.

(a) Using properties of logarithms, show that \( g(x) = cf(x) \) for some constant \( c \) (find \( c \)). This shows that the graph of \( g \) is obtained from the graph of \( f \) by a vertical stretch.

(b) Graph \( f \) and \( g \) on the same axes using a graphing utility. How could you convince another person that \( g(x) = cf(x) \), using only the graph?

(c) Based on your answer to (a), how do you expect the derivatives of \( g \) and \( f \) to be related to each other?

(d) In general, how are the graphs of \( \log_{\alpha} x \) and \( \log_{\beta} x \) related to each other?
19. Let $\alpha$ be a positive real number not equal to 1. How are the graphs of $\alpha^x$ and $\left(\frac{1}{\alpha}\right)^x$ related to each other? How are the graphs of $\log_\alpha x$ and $\log_{1/\alpha} x$ related to each other? Illustrate your answer with hand-drawn sketches for the four graphs with $\alpha = 2$.

20. Let $\alpha$ and $\beta$ be positive real numbers, not equal to 1. Show that the graphs of the functions $\alpha^x$ and $\beta^x$ are related to each other by a graphing transformation (find the specific transformation).

21. Let $f(x) = 2^x$.

(a) Compute $f(x - 1)$, $f(x - 5)$, and $f(x - c)$.

(b) Show that any horizontal shift of the graph of $f$ can also be interpreted as a vertical stretch.

(c) Use graphs to give a convincing visual explanation of (b).

22. In this exercise, we consider why $\log_1 x$ is nonsensical.

(a) Describe the graph of the function $1^x$, and then explain why this function does not have an inverse function. (What would the graph of the inverse function be?)

(b) Plot the functions $\log_2 x$, $\log_{1.1} x$, $\log_{1.01} x$, and $\log_{1.001} x$ using a graphing utility. What seems to be happening to the graphs (concentrate on the slope around $x = 1$). Assuming that $\log_1 x = \lim_{\alpha \to 1} \log_\alpha x$, explain how this agrees with your answer to (a).

23. In high school, students learn several rules regarding products of expressions involving exponents. On the one hand, they learn that $a^rb^s = (ab)^r$ (a and $b$ are positive real and $r$ is real). In other words, “when the exponents are the same, the product can be simplified.” Likewise, they learn that $a^r \cdot a^s = a^{r+s}$ (“if the bases are the same, the product can be simplified”).

Consider $a^rb^s$, where $a$ and $b$ are positive real numbers and $r$ and $s$ are real numbers. How can you write $a^rb^s$ as base exponent? Can you do this in several different ways? (Expect to use logarithms somewhere.)

24. Just as exponential functions increase more rapidly than polynomial functions, it’s true that logarithmic functions increase less rapidly than polynomial functions—or even root functions. Illustrate this with a spreadsheet that compares the values of $\sqrt{x}$ and $\log_2 x$ for integers $1 \leq x \leq 64$. (See Exercise 12 for helpful Mathematica syntax.) Is the spreadsheet more helpful or less helpful than a graph?
25. Since logarithmic functions increase slowly, composites of logarithmic functions increase very slowly. Illustrate this in the following ways:

(a) Graph $\log_{10}(\log_{10} x)$ on the interval $[1000, 1001]$, and compute the slope of the secant line.
(b) Graph $\log_{10}(\log_{10} x)$ on the interval $[10, 10^{20}]$. What are the minimum and maximum value of the function on this interval?
(c) How big must $x$ be to have $\log_{10}(\log_{10} x) \geq 2$?

26. Prove that if $(x_i)$ is a geometric sequence, then $(\log_{a} x_i)$ is an arithmetic sequence.

27. Do you expect there to be a helpful identity for $\log_{a}(x + y)$? Why or why not?

6. A Rigorous Approach to Real Exponential Functions

In Section 4, we attempted to extend the function $x \mapsto \alpha^x$ from the rational numbers to the real numbers. We did this in the following way: define $\alpha^x$ to be $\lim_{n \to \infty} \alpha^{r_n}$, where $\{r_n\}$ is a sequence of rational numbers converging to $x$. We remarked that we were cheating: for this to make sense, the limit must exist, and must not depend on the sequence $\{r_n\}$. In this section, we verify that no problems exist in defining $\alpha^x$ in this way.

To understand why we are worrying about this at all, try the following Your Turn:

**Your Turn 16.** Given a rational number $x$, we know that we can write $x$ as a decimal that has a repeating block (interpret a terminating decimal as “0 repeating”). Define $f(x)$ to be the length of the repeating block.

(a) Find a sequence of rational numbers that converges to $\sqrt{2}$, such that $f(x) = 1$ for each of the rational numbers.
(b) Find a sequence of rational number that converges to $\sqrt{2}$, such that $f(x) = 2$ for each of the rational numbers.
(c) Explain why it’s not reasonable to try to define $f(\sqrt{2})$.

Note: in this section, Exercises are used to fill in details of proofs, and are interspersed in the text.

---

15 Math joke: the function $\log_{10}(\log_{10} x)$ is known to approach $\infty$ but has never been observed to do so.
6.1. Defining the exponential function on $\mathbb{R}$. The goal of this subsection is to verify that $\alpha^x$ is well-defined for any real number $x$ (see Theorem 14).

**Lemma 11.** If $\alpha > 1$, then $\displaystyle \lim_{n \to \infty} \alpha^{1/n} = 1$.

**Proof.** First, since $\alpha > 1$ and $\alpha^n - 1 = \alpha^{n-1} + \alpha^{n-2} + \cdots + \alpha + 1$,
we have

\[
\frac{\alpha^n - 1}{\alpha - 1} \geq \underbrace{1 + 1 + \cdots + 1}_{n \text{ times}} = n.
\]
Therefore $(\alpha^n - 1) \geq n(\alpha - 1)$. Since $\alpha^{1/n} > 1$ (part (ii) of Proposition 6), we may replace $\alpha$ by $\alpha^{1/n}$ in the previous inequality, yielding $(\alpha - 1) \geq n(\alpha^{1/n} - 1)$. From this one obtains

\[1 \leq \alpha^{1/n} \leq \frac{\alpha - 1 + n}{n},\]
for all natural numbers $n$, and thus, by the squeeze theorem, $\lim_{n \to \infty} \alpha^{1/n} = 1$. \hfill $\square$

**Corollary 12.** If $\alpha > 1$, then $\displaystyle \lim_{r \to 0} r \in \mathbb{Q} \alpha^r = 1$.

**Exercise 1.** Prove Lemma 11 in the case that $0 < \alpha < 1$. (Hint: If $0 < \alpha < 1$, then $1/\alpha > 1$.)

**Exercise 2.** Prove Corollary 12.

**Proposition 13.** Let $\gamma \in \mathbb{R}$, and let $\{r_n\}, \{s_n\}$ be sequences of rational numbers converging to $\gamma$. Both $\{\alpha^{r_n}\}$ and $\{\alpha^{s_n}\}$ converge to the same real number.

**Proof.** We first verify that both $\{\alpha^{r_n}\}$ and $\{\alpha^{s_n}\}$ converge. Without loss of generality, we consider $\{\alpha^{r_n}\}$. By the completeness property of the real numbers, it suffices to show that $\{\alpha^{r_n}\}$ is a Cauchy sequence. So, let $\epsilon > 0$ be given and pick $l \in \mathbb{Q}$ with $l > \gamma$. We may choose:

(a) $N_1 > 0$ such that if $n > N_1$ then $r_n < l$. (We may do this since $r_n$ converges to $\gamma$, which is less than $l$.)
(b) \(N_2 > 0\) such that if \(m, n > N_2\), then \(|\alpha^{(r_m-r_n)} - 1| < \epsilon/\alpha^l\). (We may do this since \(\alpha^r \to 1\) as \(r \to 0\) through rational numbers, and since \(\{r_n\}\) is a Cauchy sequence.)

Put \(N = \max\{N_1, N_2\}\). Then, for \(n, m > N\), we have

\[
|\alpha^{r_n} - \alpha^{r_m}| = |\alpha^{r_n}| \cdot |1 - \alpha^{(r_m-r_n)}| < \alpha^l \cdot \frac{\epsilon}{\alpha^l} = \epsilon.
\]

Therefore \(\{\alpha^{r_n}\}\) is a Cauchy sequence.

Now, suppose \(\{\alpha^{r_n}\}\) and \(\{\alpha^{s_n}\}\) converge to \(\beta_r \in \mathbb{R}\) and \(\beta_s \in \mathbb{R}\), respectively, and let \(l\) be as above.

Given \(\epsilon > 0\), we may choose:

(a) \(N_1 > 0\) such that if \(n > N_1\) then \(|\alpha^{r_n} - \beta_r| < \epsilon/3\). (We may do this because \(\{\alpha^{r_n}\}\) converges to \(\beta_r\).)

(b) \(N_2 > 0\) such that if \(n > N_2\) then \(|\alpha^{s_n} - \beta_s| < \epsilon/3\). (We may do this because \(\{\alpha^{s_n}\}\) converges to \(\beta_s\).)

(c) \(N_3 > 0\) such that if \(n > N_3\) then \(|1 - \alpha^{(r_n-s_n)}| < \epsilon/3\alpha^l\). (We may do this by Corollary 12, since \(\{r_n-s_n\}\) converges to 0 through rational numbers.)

Put \(N = \max\{N_1, N_2, N_3\}\). Then, for \(n > N\), the triangle inequality gives

\[
|\beta_r - \beta_s| \leq |\beta_r - \alpha^{r_n}| + |\alpha^{r_n} - \alpha^{s_n}| + |\alpha^{s_n} - \beta_s|
\]

\[
= |\beta_r - \alpha^{r_n}| + |\alpha^{s_n}| \cdot |1 - \alpha^{(r_n-s_n)}| + |\alpha^{s_n} - \beta_s|
\]

\[
< \epsilon/3 + \alpha^l \cdot \epsilon/3\alpha^l + \epsilon/3
\]

\[
= \epsilon.
\]

We have \(0 \leq |\beta_r - \beta_s| < \epsilon\) for any given positive \(\epsilon\), which implies \(\beta_r = \beta_s\).

\[\square\]

**Theorem 14.** Let \(\alpha, \gamma \in \mathbb{R}\) with \(\alpha > 1\). Let \(\{r_n\}\) be any sequence of rational numbers converging to \(\gamma\).

The function \(\alpha^x\) defined by \(\alpha^\gamma = \lim_{n \to \infty} \alpha^{r_n}\) is well defined on the real numbers.

**Proof.** By Proposition 13, the value of \(\alpha^\gamma\) does not depend on our choice of sequence \(\{r_n\}\) converging to \(\gamma\). Therefore \(\alpha^x\) is well defined on the real numbers.

\[\square\]
EXERCISE 3. Let $\alpha > 1$. Use the definition of $\alpha^x$ given in Theorem 14 to verify that $\alpha^{-\gamma} = 1/\alpha^\gamma$ for $\gamma \in \mathbb{R}$.

EXERCISE 4. Let $\alpha > 1$. For $\gamma, \sigma \in \mathbb{R}$, verify that $\alpha^\gamma \alpha^\sigma = \alpha^{\gamma+\sigma}$ and that $(\alpha^\gamma)^\sigma = \alpha^{\gamma\sigma}$.

6.2. The exponential function is positive, and is increasing when $\alpha > 1$. In this section we verify two important properties of $\alpha^x$: It takes on positive values, and it is increasing when $\alpha > 1$.

Lemma 15. Suppose $\alpha, \gamma \in \mathbb{R}$ with $\alpha > 1$. Then $\alpha^\gamma > 1$ if and only if $\gamma > 0$.

Proof. First suppose that $\gamma > 0$. Recall that $\alpha^r$ is an increasing function on $\mathbb{Q}$, and that $\alpha^r > 1$ for rational $r > 0$ (see Proposition 6). Since $\gamma$ is positive, we may choose a sequence $\{r_n\}$ of rational numbers which is increasing, positive, and converges to $\gamma$. By Theorem 14, $\alpha^\gamma = \lim_{n \to \infty} \alpha^{r_n}$. However, since $\alpha^r$ is an increasing function on $\mathbb{Q}$, and $\alpha^r > 1$ for rational $r > 0$, $\lim_{n \to \infty} \alpha^{r_n} \geq \alpha^{r_k} > 1$ for any choice of $k$. We conclude that $\alpha^\gamma > 1$.

For the reverse direction, we prove the contrapositive of the forward direction. Suppose $\gamma \leq 0$. If $\gamma = 0$, then $\alpha^\gamma = 1$. If $\gamma < 0$, then applying the forward direction gives $\alpha^\gamma = 1/(\alpha^{-\gamma}) \leq 1$. □

From this lemma one may conclude that the exponential function takes on positive values:

Corollary 16. Let $\alpha > 1$ be a real number. The function $\alpha^x$ takes on positive values.

Exercise 5. Prove Corollary 16.

Proposition 17. Let $\alpha > 1$ be a real number. The function $\alpha^x$ is increasing on $\mathbb{R}$.

Proof. Let $x, y \in \mathbb{R}$. Then, by Lemma 15, Exercise 4, and Corollary 16, we have

$$x > y \implies x - y > 0 \implies \alpha^{x-y} > 1 \implies \frac{\alpha^x}{\alpha^y} > 1 \implies \alpha^x > \alpha^y.$$ 

Therefore $\alpha^x$ is increasing on $\mathbb{R}$. □

Exercise 6. Lemma 15, Exercise 4, and Corollary 16 are referred to in the proof of Proposition 17. Match these results with individual steps in the chain of implications given in the proof.
6.3. The exponential function is continuous on \( \mathbb{R} \). We conclude by showing that the exponential function is continuous.

**Lemma 18.** If \( \alpha > 1 \), then \( \lim_{x \to 0^+} \alpha^x = 1 \).

**Proof.** We verify that the two one-sided limits exist and agree. For \( \lim_{x \to 0^+} \alpha^x \), let \( \varepsilon > 0 \) be given, and pick a positive integer \( n \) satisfying \( |\alpha^{1/n} - 1| < \varepsilon \). (Such an \( n \) may be chosen due to Lemma 11).

Put \( \delta = 1/n \). If \( 0 < x < \delta \), then, since \( \alpha^x \) is increasing and \( \alpha^x > 1 \) when \( x > 0 \), we have

\[
|\alpha^x - 1| = \alpha^x - 1 \leq \alpha^{1/n} - 1 < \varepsilon.
\]

We conclude that \( \lim_{x \to 0^+} \alpha^x = 1 \).

By taking reciprocals, one may use the previous result to show that \( \lim_{x \to 0^-} \alpha^x = 1 \). \( \square \)

**Exercise 7.** Complete the proof of Lemma 18 by showing that \( \lim_{x \to 0^-} \alpha^x = 1 \).

**Theorem 19.** Let \( \alpha > 1 \). The function \( \alpha^x \) is continuous on \( \mathbb{R} \).

**Proof.** Let \( \gamma \in \mathbb{R} \). We wish to show that \( \lim_{x \to \gamma} \alpha^x = \alpha^\gamma \), or rather that \( \lim_{x \to \gamma} \alpha^x - \alpha^\gamma = 0 \). By Lemma 18, we see

\[
\lim_{x \to \gamma} (\alpha^x - \alpha^\gamma) = \lim_{x \to \gamma} \alpha^\gamma (\alpha^{x-\gamma} - 1) = \alpha^\gamma \lim_{x \to \gamma} (\alpha^{x-\gamma} - 1) = \alpha^\gamma \cdot 0 = 0,
\]

concluding the proof. \( \square \)

**Exercise 8.** The goal of this exercise is to show that the range of \( \alpha^x \) is \((0, \infty)\) whenever \( \alpha > 1 \) is a real number. (The same result is true when \( 0 < \alpha < 1 \) by taking reciprocals.)

(i) Use the binomial theorem to show that \( (1 + a)^n > na \) whenever \( a > 0 \).

(ii) Show that \( \lim_{n \to \infty} a^n = +\infty \) whenever \( \alpha > 1 \). (Hint: Write \( \alpha = 1 + a \) for some \( a > 0 \), and apply part (i). You may assume that \( \lim_{n \to \infty} na = \infty \) if \( a > 0 \).)

(iii) Since \( \alpha^x \) is increasing, it follows from part (ii) that \( \lim_{x \in \mathbb{R}} \alpha^x = +\infty \). Using this fact, explain why \( \lim_{x \to -\infty} \alpha^x = 0 \).

(iv) Use part (iii) together with the Intermediate Value Theorem to show that the range of \( \alpha^x \) is \((0, \infty)\).
7. Exercises Involving Student Work

This section contains student solutions to various problems involving exponentiation. For each problem and corresponding set of solutions, complete the following tasks:

1. Solve the problem (without appealing to resources such as your text or technology).
2. Discuss each student’s solution strategy. Indicate ways in which the student was mathematically correct or incorrect. Identify any incorrect assumptions, misconceptions, and errors revealed in each student’s work. Finally, comment on whether the student’s approach was persuasive and effective.
3. Rate each solution on a scale from 1 to 5 with 5 as the highest rating and 1 as the lowest.
4. Compare your ratings in part 3 with another student (or group), and discuss your criteria for giving these ratings. Attempt to reach consensus with your partner (or group).
5. Carefully construct a solution to the original problem, taking care that it is mathematically correct, readable, and persuasive.
1. Consider the statement $a^b \cdot a^c = a^{b+c}$, where $a$, $b$, and $c$ are positive integers. Determine whether the statement is true or false. Then, give a convincing mathematical justification for your answer.

(1) True

$2^2 \cdot 2^3 = 2^{2+3} = 2^5$

$\underbrace{2 \cdot 2 \cdot 2 \cdot 2 \cdot 2}_{5 \text{ times}}$

(2) True: by using $a^{b-c} = \frac{a^b}{a^c}$ we have the opp of $a^b \cdot a^c = a^{b+c}$

If $b$ is a negative constant it will bring $a^c$ to the numerator making $a^b \cdot a^c$ true and on the LHS it would be $a^{b+c}$.

(3) True

$a^b = a \cdot a \cdot a \cdot a \cdots a$ (There are $b$ a’s)

$a^c = a \cdot a \cdot a \cdot a \cdots a$ (There are $c$ a’s)

$a^{b+c} = a \cdot a \cdot a \cdot a \cdots a$ (There are $b+c$ a’s)

So you get the same amount of a’s.

(4) True
\[ a^2 a^3 = a^{2+3} \]
\[ (aa)(aaa) = aaaaa = a^5 = a^{2+3} \]

(5) True
Here \( a^b \) is \( a \cdot a \) \( b \) times. Let us denote this as \( a_1 \cdot a_2 \cdots \cdot a_b \). The same can be said about \( a^c; \ a_1 \cdot a_2 \cdots \cdot a_c \).
When the two quantities are multiplied together, you will have \((a_1 \cdot a_2 \cdots \cdot a_b)(a_1 \cdot a_2 \cdots \cdot a_c)\). Since this is multiplication we may drop the parentheses and it is seen that the quantity will be \( a \cdot a \ b + c \) times. So therefore the statement is true.

(6) True
\[ x = a^b \quad b = \log_a x \]
\[ y = a^c \quad c = \log_a y \]
\[ x \cdot y = a^{b+c} \]
\[ \Leftrightarrow \log_a (x \cdot y) = \log_a (a^{b+c}) \]
\[ \Leftrightarrow \log_a a^b + \log_a a^c = (b + c) \log_a a = b + c \]
\[ \Leftrightarrow b \log_a a + c \log_a a = b + c \]

(9) True
Proof by example: \( a = 2, \ b = 2, \ c = 3 \)
\[ \text{RHS} \Rightarrow (2)^2 \cdot (2)^3 = (4)(8) = 32 \]
\[ \text{LHS} \Rightarrow (2)^{2+3} = (2)^6 = 32 \]
(10)

True

\[ a^b = \underbrace{a \cdot a \ldots a}_{b \text{ times}} \]
\[ a^c = \underbrace{a \cdot a \ldots a}_{c \text{ times}} \]
\[ a^b \cdot a^c = (a \cdot a \ldots a) \cdot (a \cdot a \ldots a) \]

If you just take away the parentheses, you see that there are “b” a’s being added to “c” a’s to form one long chain, a chain that contains \( b + c \) a’s.
2. Consider the statement \( \log_a(\log_a b) = b \), where \( a \) and \( b \) are positive integers. Determine whether the statement is true or false. Then, give a convincing mathematical justification for your answer.

(1) False

Let’s take for instance \( a = 3 \) and \( b = 9 \)

\[ \log_3(\log_3 9) \rightarrow \text{Here } \log_3 9 = 2 \]

\[ = \log_3 2. \text{ Here } 3 \text{ to the something will equal two. We see that it is not our original value of } b \text{ which is 9 so this proves the statement is false.} \]

(2) \( a^b = a^b \)

True

(3) True

\( \log_a b(\log_a b) = b \)

\( a^b = \log_a b \)

\( a^{a^b} = b \)

(4) True

\( a^b = \log_a b \)
(5)
\[ \log_a b = a^b \]
False
\[ \log_a 8 = 2^8 \]
\[ \Rightarrow 3 \neq 256 \]

(6)
\[ a^b = \log_a b \]
\[ a^{a^b} = b \]
\[ a^{ab} = b \]
False

(7)
This is false.
say \( a = 2 \) and \( b = 8 \).
Then this statement becomes \( \log_2(\log_2 8) \) on the left hand side. This equals \( \log_2 3 \).
\[ \log_2 3 \neq 8, \text{ because } 2^8 \text{ is far greater than } 3. \]

(8)
False, just another log
needs to be \( a^{\log_a b} = b \)
this says \( a^b = \log_a b \)
which says \( b = a^{ab} \)
(9)
No, false.
example \( \log_2(\log_2 16) \neq 16 \)
\( = \log_2(4) = 2 \)

(10)
True,
This would mean that \( a^b = \log_a b \)
and then \( b = a^{(a^b)} \), which could be written as \( \log_a(\log_a b) \)

(11)
False.
\( \log_a(\log_a b) = b \) implies \( a^b = \log_a b \) from the definition of log
but, exponentials and logs are inverse functions not equivalent functions.
To correct this, it should be \( a^{(\log_a b)} = b \)

(12)
This statement is true.
\( \log_a(\log_a b) = b \)
\( \Rightarrow \log_2(\log_2 4) = 4 \) \hspace{1cm} (where \( a = 2, b = 4 \))
\( \Rightarrow \log_2(x) = 4 \) \hspace{1cm} (\( x = \log_2 4 \))
\( = 2^4 = x \) so \( x = 16 \)
\( \Rightarrow \text{plug in} \rightarrow \log_2(16) = 4 \)
\( = 2^4 = 16 \) which is true
3. Consider the statement \((a^b)^c = a^{(b^c)}\), where \(a\), \(b\), and \(c\) are positive integers. Determine whether the statement is true or false. Then, give a convincing mathematical justification for your answer.

(1)
False, \((a^b)^c = a^{(b^c)}\)
When you take a power to a power, you multiply.

(2)
False.
\[
a^b = \underbrace{a \cdot a \cdot a \cdot \ldots}_{b \text{ times}}
\]
\[
(a^b)^c = \left[ \underbrace{a \cdot a \cdot a \cdot \ldots}_{b \text{ times}} \right]^c = \left[ \underbrace{a \cdot a \cdot a \cdot \ldots}_{b \text{ times}} \right]^c \cdot \left[ \underbrace{a \cdot a \cdot a \cdot \ldots}_{b \text{ times}} \right]^c \cdot \ldots
\]
which is really \(a^{(b^c)}\)
Ex. \((3^2)^{-1} = \frac{1}{3^2} = \frac{1}{9} \neq 3^{2^{-1}} = 3^{\frac{1}{2}} = \sqrt{3}\)

(3)
False
If we let \(a = 2\), \(b = 1\), and \(c = 3\)

The LHS= \((2^1)^3 = 2^3 = 8\)

The RHS= \(2^{(1^3)} = 2^1 = 2\)

\(2 \neq 8\) so this statement is false
(4)

False

Let \( b = 2, \quad c = 3 \)

\[
(a^2)^3 = \underbrace{a \cdot a \cdot a \cdot a \cdot a \cdot a}_{2} = a^6
\]

\[
a^{(2^3)} = a^8 = a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a = a^8
\]

Not Equal

(5)

False

Let \( a = 2, \quad b = 2, \quad c = 3 \)

\[
LHS : (2^3)^3 = 4^3 = 64
\]

\[
RHS : (2)^{(2^3)} = 2^8 = 1024
\]

NOT EQUAL!

Instead, \((a^b)^c = a^{(bc)}\)

(6)

False

\[
\log_\alpha (a^b)^c = c \log_\alpha (a^b) = bc \log_\alpha a
\]

\[
\log_\alpha a^{(b^c)} = b^c \log_\alpha a
\]

\[
b^c \log_\alpha a \neq bc \log_\alpha a
\]

(7)

False

According to exponential rules, \((a^b)^c = a^{b \cdot c}\)

and \(a^{(b^c)} \neq a^{(b \cdot c)}\) because in most cases \(b^c \neq b \cdot c\).
False

RHS $a \cdot a \cdot a \ldots a \cdot a \cdot a \ldots a \cdot a \cdot a \ldots$

$b$ times $b$ times $b$ times

LHS $b \cdot b \cdot b \ldots b \cdot b \cdot b \ldots$

$c$ times $c$
4. Solve for $x$: \( \log_b \log_a x = c \)

(1)
\[
\log_b \log_a x = c \\
\iff \log_a x = b^c \\
\iff x = a^{b^c}
\]

(2)
\[
\log_b (\log_a x) = c \\
\log_a x = b^c \\
x = a^{b^c}
\]

(3)
\[
b^c = a^x \\
c \ln b = x \ln a \\
x = \frac{c \ln b}{\ln a}
\]

(4)
\[
b^c = \log_a x \\
(a^b)^c = x
\]

(5)
\[
b^c = \log_a x \\
take \log_a \log_a (b^c) = x
\]
(6)

\[
\log_b(\log_y x) = c \\
\Rightarrow \log_b(y) = c \\
= b^c = y \\
\rightarrow \text{so } \log_a x = b^c \\
= a(^{b^c}) = x
\]

Hence \( x = a(^{b^c}) \)

(7)

\[
\log_a x = u \\
\log_b u = c \\
= b^c = u
\]

Therefore \( \log_a x = b^c \)

\( a(^{b^c}) = x \)
CHAPTER 10

Exponential and Logarithmic Functions: History, Computation, and Application

Typically, the strategy for developing logarithms in most algebra or precalculus classes is more or less the one we took in Chapter 9: Exponential functions are discussed first and then a logarithmic function is defined as the inverse of the exponential function with the same base, or simply by the rule “y = \log_\alpha x means \alpha^y = x.” However, the same students may see a very different strategy in their AP Calculus class.\textsuperscript{1} It’s common to define the natural logarithm function using calculus itself: ln \(x\) is defined as the integral \(\int_1^x \frac{dt}{t}\). Its inverse function is denoted exp \(x\), or somewhat more suggestively as \(e^x\). To justify the latter notation one needs to show that there is a number “\(e\)” such that exp \(x\) really equals the exponential function \(e^x\). Equally well, it means showing that ln \(x\) is a genuine logarithmic function—the one with base \(e\).

The second strategy illuminates two fundamental issues involving the exponential and logarithmic functions, which are concealed in the first strategy:

- When logarithms are introduced via the integral \(\int_1^x \frac{dt}{t}\), it becomes clear that the number \(e\) plays a very distinguished role, as an “ideal” base for exponential and logarithmic functions. The number \(e\) did not appear in the developments in the last chapter, since the special significance of \(e\) makes its appearance only when the ideas of calculus are applied to exponential and logarithmic functions.

- Much of the importance of exponential and logarithmic functions comes from “calculus-based reasons.” Again, we all recognize this fact—from our long experience integrating functions like \(\sec x\)—but calculus is involved in the very definition of the natural logarithm as an antiderivative of \(\frac{1}{t}\). Moreover, the reason that exponential functions (functions of the form \(f(t) = ce^{at}\)) arise in modeling the growth of bacteria cultures and investments is that they satisfy a very special derivative condition (namely, that \(f'(t)\) is a constant

\textsuperscript{1}Calculus texts vary. Nowadays a text with the subtitle Early Transcendentals may assume students’ familiarity with exponential and logarithmic functions. This results in a markedly different approach to deducing the derivative formulas for exponential and logarithmic functions.
multiple of \( f(t) \), as we shall see in Section 4). The relationship between periodically-compounded interest and continuously compounded interest, at its mathematical heart, is a limit formula that is conveniently proved using calculus.

Section 1 of this chapter emphasizes the computational origins of the study of logarithms, taking an historical point of view. The fact that \( x \mapsto \int_1^x \frac{dt}{t} \) really is a logarithmic function is discussed in detail. Section 3 revisits crucial ideas from calculus about exponential and logarithmic functions. Applications of the theory of exponential functions to exponential growth and compound interest are discussed in Section 4.

1. Logarithms and History: Logarithm Tables

The usefulness of logarithms\(^2\) in computation prompted Pierre-Simon de Laplace (1749-1827) to say, “Logarithms, by shortening the labors, doubled the life of the astronomer.” Indeed, from the 17th century until the advent of the hand-held calculator, virtually all calculations requiring multiplication were performed with the aid of logarithms, either via tables or slide rules\(^3\).

So what is it about a logarithm that makes it so powerful? *Logarithms allow us to convert multiplication problems into simple addition problems* via the familiar property

\[
\log_\alpha(xy) = \log_\alpha x + \log_\alpha y.
\]

For example, consider Laplace’s astronomers: In astronomy, one does a lot of trigonometry, so one needs to know the value of the trigonometric functions—not only at special angles such as \(\pi/4\) and \(\pi/3\), but at other angles as well. Toward this goal, we may employ trigonometric identities, including the basic sum formulas for sine and cosine:

\[
\cos(\phi + \theta) = \cos \phi \cos \theta - \sin \phi \sin \theta
\]

\[
\sin(\phi + \theta) = \cos \phi \sin \theta + \sin \phi \cos \theta.
\]

\(^2\)The writing of this portion of the chapter was significantly influenced by William Dunham’s book “Euler: The Master of Us All” (Mathematical Association of America, 1999).

\(^3\)The logarithm was first introduced by the Scottish Lord John Napier (1550-1617) in the late 16th century. Napier’s discovery was motivated by observations made by Michael Stifel (1486-1567): Stifel pointed out that a product of terms in a geometric sequence corresponds to a sum in the corresponding arithmetic sequence of exponents.
Observe that these identities routinely contain products, which become tedious upon repeated application. Logarithms allow us to convert those products to sums, thus lessening the difficulty of the computation.

Let’s overview our goals in this section. In Section 1.1, we explore the mechanics of using a table of logarithms, and “computing products by computing sums.” But how were the tables of logarithms created? (Remember, they existed for centuries before the invention of calculators.) We learn about an historical technique for computing base-ten logarithms in Section 1.2, which should convince us that computing base-ten logarithms is difficult.

1.1. Logarithm tables and interpolation. Before we can compute using logarithms, we must first learn to read the table (at back, taken from CRC Standard Math Tables). The table shows the values of common logarithms (i.e., $\log_{10} x$ or just $\log x$) for the numbers $1 \leq x < 10$, accurate to four decimal places. For example, to find $\log 2.93$ in the table, one reads down to 29 in the left hand column, and then over to the 3 column, finding the number 4669. We conclude that $\log 2.93 \approx 0.4669$.

All of this is well and good for numbers between one and ten, but what about all of the other positive real numbers? This is where the properties of logarithms work to our advantage. For example, suppose we wish to find $\log 0.00542$. We may use the additive property of the logarithm together with the table to conclude

$$\log 0.00542 = \log (5.42 \times 10^{-3}) = \log 5.42 + \log 10^{-3} \approx 0.7340 + (-3) = -2.266.$$ 

By expressing a given number in scientific notation, we can readily approximate its logarithm from the given table.

There is another problem to consider. Suppose we wish to find the logarithm of a number, say $\log 2.936$, whose value falls between values given in the table? A simple and mathematically compelling way to handle this situation is to apply linear interpolation. The idea behind linear interpolation is that a secant line joining two points close together on a curve provides a good approximation of the curve between these two points. For example, consider the secant line joining the points $(2.93, \log 2.93)$ and $(2.94, \log 2.94)$ on the curve $y = \log x$. (See Figure 1 below. Note that $\log 2.93 \approx 0.4669$ and $\log 2.94 \approx .4683$ may be found using the table.)

---

4Understandably, logarithms played a large role in constructing highly accurate modern trig tables.
From Figure 1, we see that the $y$-coordinate $y_0$ of the point on the secant line whose $x$-coordinate is 2.936 is very nearly equal to $\log 2.936$. The value $y_0$ is the approximation we seek, and it can be found by computing the slope $m_{sec}$ of the secant line in two different ways:

$$m_{sec} = \frac{0.4683 - 0.4669}{2.94 - 2.93} = \frac{y_0 - 0.4669}{2.936 - 2.93}.$$ 

To finish the computation, one solves the second equation for $y_0$, giving $\log 2.936 \approx y_0 = 0.4677$.

You may have noticed that linear interpolation can be tedious, especially when it comes to the cross multiplication of decimals. We wish to avoid this sort of multiplication if we are going to use the tables to simplify calculations. To help with this, the right-hand portion of the table includes a list of proportional parts which provides additional information intended to take the place of interpolation. To see how it works, suppose we wish to find 2.936. We find from the table that $\log 2.93 = 0.4669$. Finally, to account for the missing 0.006, we read over to the 6-th column of the proportional parts table and read off the number 9, which is intended to mean 0.0009. We then compute $0.4669 + 0.0009$ and obtain 0.4678 as our approximation to $\log 2.936$.

**Your Turn 1.** Compare the value of $\log 2.936$ obtained by proportional parts and by linear interpolation. Are they equal? Do the same for $\log 1.211$. 
Since the logarithm function is one-to-one, one may also use the table to find (or rather approximate) the unique number whose logarithm is some given value. For example, suppose we want to use the table to approximate the number \( \beta \) whose common logarithm is 2.8779 (that is, we want to approximate \( 10^{2.8779} \)).

First, we express 2.8779 as a sum of an integer and a number between 0 and 1, namely 2.8779 = 2 + 0.8779.

Now, in the body of the table we locate 8779, and we see that 0.8779 \( \approx \) \( \log 7.55 \). Thus

\[
\log \beta = 2.8779 \approx \log(7.55) + 2 = \log 7.55 + \log 10^2 = \log(7.55 \cdot 10^2) = \log(755).
\]

We conclude that \( \beta \approx 755 \).

Now that we can read the table, we can use it together with the properties of the logarithm to convert multiplication problems to addition problems, as well as to convert exponentiation to multiplication.\(^5\) We consider two examples.

First, suppose we wish to compute 293.6 \times 0.00048 using logarithms. We begin by computing \( \log(293.6 \times 0.00048) \) using the table (using proportional parts if necessary), which we express as the sum of an integer and a number between zero and one:

\[
(1) \quad \log(293.6 \times 0.00048) = \log(2.936 \cdot 4.8 \cdot 10^{-2}) = \log 2.936 + \log 4.8 - 2 \approx 0.4678 + 0.6812 - 2 = 0.1489 - 1.
\]

From the table (using proportional parts) we see that 0.1489 \( \approx \) \( \log 1.409 \), and we know that 1 = \( \log 10 \). Thus Equation (1) becomes

\[
\log(293.6 \times 0.00048) \approx \log 1.409 - \log 10 = \log \frac{1.409}{10}.
\]

Finally, since \( \log(x) \) is one-to-one, we have 293.6 \times 0.00048 \( \approx \) 0.1409. The key point here is that we never had to multiply decimals to compute this product! We merely added them.

Next, suppose we wish to compute \((31.5)^{2.3}\). As before, we use the properties of logarithms together with the table to express \( \log(31.5)^{2.3} \) as the sum of an integer together with a number between 0 and 1:

\[
\log(31.5)^{2.3} = 2.3 \log 31.5 \approx 2.3 \cdot 1.4983 \approx 3.4461 = 3 + 0.4461.
\]

\(^5\)One should keep in mind that from the 17th-century until the advent of the hand-held calculator in the late 20-th century, computations were routinely made with the aid of logarithm tables.
(Note that the product $2.3 \times 1.4983$ could be computed using logarithms if one wished to avoid multiplication.)

Now, from the table we find $0.4461 \approx \log 2.793$, and so

$$\log(31.5)^{2.3} \approx 3 + 0.4461 \approx \log 1000 + \log 2.793 = \log 2793.$$ 

From this we conclude that $(31.5)^{2.3} \approx 2793$.

**Your Turn 2.** What does your calculator give for $293.6 \times 0.00048$ and for $31.5^{2.3}$?

### 1.2. Tables and Tedium: An Historical Approach to Generating Log Tables

In the previous section we showed that an accurate logarithm table can be used effectively to carry out multiplication and exponentiation without doing anything more difficult than adding or subtracting. So then, how do we obtain the table in the first place?

The first common logarithm tables were painstakingly constructed by Henry Briggs (1561-1631), requiring years of labor. To see why Briggs needed years to complete a reasonable table, let’s stand in Briggs’ shoes and approximate $\log 5$ as he would have.\(^6\) To begin, we note that $\log 10 = 1$, and so

$$\log \sqrt{10} = \log 10^{1/2} = \frac{1}{2} \log 10 = 0.5.$$ 

Upon extraction\(^7\) of $\sqrt{10}$, we find $\sqrt{10} \approx 3.1622777$, and so we conclude that $\log 3.1622777 \approx 0.5$. By extracting another root and using the properties of the logarithm, we find that

$$0.25 = \frac{1}{2} \cdot 0.5 = \frac{1}{2} \log 10^{1/2} = \log(10^{1/2})^{1/2} \approx \log(3.1622777^{1/2}) \approx \log 1.7782794.$$ 

By repeating this process, one may create the following table.

<table>
<thead>
<tr>
<th>Number</th>
<th>Logarithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^1$</td>
<td>$1.0$</td>
</tr>
<tr>
<td>$3.1622777 = 10^{1/2}$</td>
<td>$0.5$</td>
</tr>
<tr>
<td>$1.7782794 = 10^{1/4}$</td>
<td>$0.25$</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>$1.00011249 = 10^{1/2048}$</td>
<td>$0.00048828$</td>
</tr>
<tr>
<td>$1.0005623 = 10^{1/4096}$</td>
<td>$0.00024414$</td>
</tr>
<tr>
<td>$1.0002811 = 10^{1/8192}$</td>
<td>$0.00012207$</td>
</tr>
</tbody>
</table>

What has this to do with $\log 5$? By repeatedly extracting square roots from 5, one finds that $5^{1/4096} \approx 1.0003930$, which falls between the last two roots of 10 in the left column of our table. Thus, from the

---

\(^6\)This calculation is taken from William Dunham’s book “Euler: The Master of Us All”, Mathematical Association of America, 1999.

\(^7\)Extraction of square roots by hand is a relatively simple procedure, but involves tedious, time-consuming calculation.
right-hand column of the table, we learn that

\[
0.00012207 \leq \log 5^{1/4096} \leq 0.00024414.
\]

To obtain a more accurate approximation of \( \log 5^{1/4096} \), Briggs would have applied linear interpolation, assuming for approximation purposes that the point \( (1.0003930, \log 5^{1/4096}) \) lies on the secant line for \( y = \log x \) joining the points \( (1.0002811, 0.00012207) \) and \( (1.0005623, 0.00024414) \) (this is illustrated below in Figure 2).

Figure 2. Interpolation for \( \log 5^{1/4096} \).

Performing the interpolation gives

\[
\frac{0.00024414 - 0.00012207}{1.0005623 - 1.0002811} \approx \frac{\log 5^{1/4096} - 0.00012207}{1.0003930 - 1.0002811},
\]

and so \( \log 5^{1/4096} \approx 0.000170646 \). By properties of logarithms, we have \( \frac{\log 5}{4096} \approx 0.000170646 \). Finally, we multiply both sides of this last equation by 4096 to obtain \( \log 5 \approx 0.698966 \).

Briggs clearly made the best use of the mathematical tools at his disposal, but this seems like a lot of tedious calculation to compute just one logarithm, and in fact one would have to do most of it all over again to obtain logarithms of numbers other than five. In the next section we will see how the advent of calculus greatly simplified the approximation of logarithms.

1.3. Exercises.
1. Use the table of logarithms to find the following common (base ten) logarithms. Be sure to use linear interpolation when necessary.

(a) \( \log 0.0877 \)

(b) \( \log 22 \)

(c) \( \log 45.68 \)

(d) \( \log 45680 \)

2. Use the table to find the number whose common logarithm is:

(a) \( 2.4843 \)

(b) \( -3.3458 \) (Note \(-3.3458 = 0.6542 + (-4)\).)

(c) \( 0.3860 \) (You will need to do some interpolation here. To do this, assume that \( 0.3860 \) is the \( y \)-coordinate of a point on the secant line joining nearby points on \( y = \log x \) determined from the table. Your goal is to find the \( x \)-coordinate of this point on the secant line.)

3. Use proportional parts to assist with approximating the following common logarithms.

(a) \( \log 35.47 \)

(b) \( \log 0.08972 \)

4. Use proportional parts to assist with approximating the number whose common logarithm is:

(a) \( 0.4775 \)

(b) \( 2.873 \)

5. Perform the following computations using the table of common logarithms, without resorting to multiplication or exponentiation. (If a product is required within an exponentiation problem, compute it via logarithms.)

(a) \( 2.762 \times 603 \)

(b) \( 7.6 \times 0.00004765 \)

(c) \( 22.83^{25} \)

(d) \( 0.87^{15} \)
In the last section we see that the computation of logarithms was initially quite difficult. The natural (base $e$) logarithm we all know as $\int_1^x \frac{dt}{t}$, was invented precisely to make the calculation of logarithms easier (Sections 2.1 and 2.2). Once we have a way of estimating values of $\ln x$, it’s interesting to see how we can estimate values of $e^x$; we give an historical argument for the famous Taylor series for $e^x$ that predates the general theory of Taylor series.

2.1. The natural logarithm: If it looks like a logarithm then it probably is one. As Briggs’ life was drawing to a close, a new crop of mathematicians were beginning to make discoveries in what would eventually become known as calculus. Two of these mathematicians, Gregory of St. Vincent (1584-1667) and Alfonso de Sarasa (1618-1667), made an observation about hyperbolas which was to have a profound effect on the process of approximating logarithms, giving rise to what we know as the natural logarithm.

In modern terms, the observation is as follows. For $x \in (0, \infty)$, define

$$A(x) = \int_1^x \frac{1}{t} dt.$$  

Geometrically, one may think of $A(x)$ as measuring the signed area under the hyperbola $y = 1/t$ from $t = 1$ to $t = x$ (see Figure 3). St. Vincent and Sarasa noticed that $A(xy) = A(x) + A(y)$ and $A(1) = 0$, properties which are reminiscent of the logarithm:

---

*These early pioneers include the likes of Pierre Fermat, Blaise Pascal, John Wallis, Gilles Roberval, and James Gregory. Isaac Newton and Gottfried Liebniz, who are often given joint credit for “inventing” calculus, built upon the existing work of these and other mathematicians. As Newton said, “If I have seen further it is by standing on the shoulders of giants.”*
Your Turn 3. Let $A(x) = \int_1^x \frac{1}{t} \, dt$. Show that $A(xy) = A(x) + A(y)$. (Hint: 
$$A(xy) = \int_1^{xy} \frac{1}{t} \, dt = \int_1^x \frac{1}{t} \, dt + \int_x^{xy} \frac{1}{t} \, dt.$$ 
Make the substitution $u = t/x$ in the right-hand-most integral.)

The area function $A(x)$ looks like a logarithm, but is it really a logarithm? If so, what is its base? To address these questions, we present the following theorem.

**Theorem 1.** Let $\alpha$ be a positive real number not equal to one. If $f : (0, \infty) \to \mathbb{R}$ is a continuous function satisfying $f(1) = 0$, $f(\alpha) = 1$, and $f(xy) = f(x) + f(y)$ for all $x, y \in (0, \infty)$, then $f(x) = \log_\alpha x$ for all $x \in (0, \infty)$.

**Proof.** It suffices to show that $f(\alpha^\gamma) = \gamma$ for each $\gamma \in \mathbb{R}$ (see Exercise 7). If $m$ is a positive integer, then by hypothesis

$$f(\alpha^m) = f(\alpha) + \cdots + f(\alpha) = 1 + \cdots + 1 = m.$$ 

Further, if $r \in \mathbb{Q}$ is a positive rational number represented by the fraction $m/n$ where both $m$ and $n$ are positive, we have

$$m = f(\alpha^m) = f((\alpha^r)^n) = n \cdot f(\alpha^r),$$

which implies $f(\alpha^r) = m/n = r$. These arguments together with the hypotheses may be used to give similar results for negative integers and rational numbers, respectively (see Exercise 8). We conclude that the desired result holds for all rational exponents.

Now, let $\{r_n\}$ be a sequence of rational numbers converging to $\gamma \in \mathbb{R}$. Since $f$ and the exponential function $\alpha^x$ are continuous, we see that

$$f(\alpha^\gamma) = f(\lim_{n \to \infty} \alpha^{r_n}) = f(\lim_{n \to \infty} \alpha^{r_n}) = \lim_{n \to \infty} f(\alpha^{r_n}) = \lim_{n \to \infty} r_n = \gamma,$$

concluding the proof. 

---

9A technical note: the proposition is still true if the continuity hypothesis is replaced by a number of other (interchangeable) conditions, including “continuous at a point”, or “strictly monotonic.” However, the proposition is false if the continuity hypothesis is removed entirely without any replacement, as the remaining hypotheses do not uniquely determine a function. To see this, observe that by composing with an exponential function we may transfer the question about the uniqueness of $f$ to a question about the uniqueness of a function $g : \mathbb{R} \to \mathbb{R}$ satisfying $g(0) = 0$, $g(1) = 1$, and $g(x + y) = g(x) + g(y)$. Since $\mathbb{R}$ is a vector space over $\mathbb{Q}$, by appealing to the Axiom of Choice (or rather Zorn’s Lemma), $\mathbb{R}$ must have a basis over $\mathbb{Q}$ containing the number 1. Now, one can define scads of distinct $\mathbb{Q}$-linear functions $g : \mathbb{R} \to \mathbb{R}$ satisfying the hypotheses by defining $g$ as we please on this basis, with the exception of requiring $g(1) = 1$. 
2. LOGARITHMS AND HISTORY: THE NATURAL LOGARITHM

We may apply Theorem 1 to the area function $A(x) = \int_1^x (1/t)dt$: We already know that $A(xy) = A(x) + A(y)$ and $A(1) = 0$. By appealing to the Fundamental Theorem of Calculus, $A(x)$ is differentiable on its domain, and hence is continuous. Finally, area estimates together with the intermediate value theorem show that $A(e) = 1$ for a number $e \approx 2.718$. We conclude that $A(x)$ satisfies the hypotheses of Theorem 1, and so $A(x)$ is a logarithm function, called the natural logarithm (denoted by $\ln x$), whose base is the number $e$.

2.2. Estimating natural logarithms. We have just seen that we can produce a logarithm, called the natural logarithm, by considering areas of regions under the hyperbola $y = 1/t$. How does this help us compute approximations to logarithms more easily than Henry Briggs did? The answer, as first discovered by Nicolaus Mercator (1620-1687), involves infinite series. First, observe that

$$\frac{1}{1+t} = 1 - t + t^2 - t^3 + \cdots$$

for $|t| < 1$ due to geometric series considerations. Also, by performing a substitution in the integral defining $\ln x$, we have

$$\ln(1 + x) = \int_0^x \frac{1}{1+t} dt.$$

Putting Equations (3) and (4) together and integrating term by term gives a series expansion\(^{10}\)

$$\ln(1 + x) = \int_0^x \left(1 - t + t^2 - t^3 + \cdots\right) dt = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$$

for $\ln(1 + x)$ valid for $-1 < x \leq 1$. This formula indicates that approximations of the natural logarithm may be made by taking partial sums of the series in the right-hand side of Equation (5). We can improve our situation: Observe, as Euler did, that

$$\ln \frac{1+x}{1-x} = \ln(1 + x) - \ln(1 - x) = \left(x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots\right) - \left(-x + \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} + \cdots\right) = 2 \left[\frac{x^3}{3} + \frac{x^5}{5} + \cdots\right].$$

This latter series is better suited for approximating a wider range of natural logarithms, and requires much less effort than Briggs’ method.

---

\(^{10}\) Note that we have essentially obtained the Taylor series of $\ln(1 + x)$, without computing all the higher derivatives of the function.
2.3. Euler’s method of estimating exponential functions. Previously, we found a power series expansion for ln(1 + x), but what about a power series expansion for $e^x$? If we believe that $\frac{d}{dx}e^x = e^x$ (see Equation (15)), then Taylor’s Theorem assures us that

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{j=0}^{\infty} \frac{x^j}{j!}.$$ 

Historically, however, this series expansion was known and understood by Euler before the advent of Taylor’s Theorem. In this section, we present Euler’s development of power series for exponential functions, and we see once again why $e$ is “natural”.\(^{11}\)

Before we get to Euler’s argument, a short sermon about the importance of power series is in order. You may recall from calculus that taking derivatives and integrals of polynomial functions is fairly easy: one simply uses the power rule. For non-polynomial functions, things can be quite a bit trickier (for example, try computing $\int_0^1 \sin(x^2) \, dx$ by hand). The beauty of power series is that they bridge the gap between polynomials and non-polynomials. More specifically, if one has a power series expression for a given function $f(x)$, say

$$(7) \quad f(x) = a_0 + a_1(x - c) + a_2(x - c)^2 + \cdots,$$

for $x$ in some interval, then it can be shown that one can integrate and differentiate $f(x)$ at will by treating the righthand side of (7) as a large polynomial, which can be integrated and differentiated term-by-term using the power rule.

Of course, there is a catch. Finding a power series expansion for a given function can be a tricky matter. One historically important way to obtain power series for certain non-polynomial functions is to use Newton’s binomial theorem:

**Theorem 2.** (Newton’s Binomial Theorem) Let $r$ be any real number. Then

$$(1 + x)^r = 1 + rx + \frac{r(r - 1)}{2!}x^2 + \frac{r(r - 1)(r - 2)}{3!}x^3 + \cdots$$

if $x \in (-1, 1)$.

\(^{11}\)By the standards of modern mathematics, Euler’s argument takes an unacceptably casual approach to limits and convergence, which should not surprise us, since the formal definitions of limit and convergence we use today were not developed until long after Euler’s death. Euler’s argument is a thing of beauty and genius, a product of one of the finest mathematical minds the world has known.
Newton never published this result and never proved it in general. He did, however, check to see that it worked for some specific examples.

**Your Turn 4.** *Show that Newton’s Binomial Theorem holds in the cases* \( r = 0, \ r = 1, \ r = 2, \) *and* \( r = -1. \)

Finally we embark on Euler’s series expansion of the exponential functions. To begin, let \( \alpha \) be a positive real number not equal to one, let \( \omega \) be an *extremely* small positive number, and suppose that

\[
\log_{\alpha} x = \omega.
\]

Our first goal is to determine a value for \( x \). From the Golden Rule of Logarithms (part (d) of Theorem 10 from Chapter 9), we know that \( \omega = \ln x / \ln \alpha \), and hence

\[
\omega \ln \alpha = \ln x.
\]

Recall that \( \ln x \) is the area under the hyperbola \( y = 1/t \) for \( 0 < t < x \) (see Figure 4). Since \( \omega \) (and hence \( \omega \ln \alpha \)) is *very small*, the rectangle of height 1 pictured in Figure 4 has essentially the same area as the corresponding region under \( y = 1/t \). We conclude that

\[
\omega \ln \alpha = \int_{1}^{x} \frac{1}{t} \, dt = \text{Area of rectangle} = 1 \cdot (x - 1).
\]

Therefore \( x = 1 + \omega \ln \alpha \), so by substituting this value of \( x \) into (8) and exponentiating, we obtain

\[
\alpha^\omega = 1 + \omega \ln \alpha.
\]
Next, let \( k = \ln \alpha \), let \( y \) be any real number, and pick \( j \in \mathbb{R} \) with \( y = \omega j \) (note that \(|j| \) will generally be large since \( \omega \) is small). Using (9) we have

\[
\alpha^y = \alpha^{\omega j} = (1 + k\omega)^j = \left(1 + \frac{ky}{j}\right)^j.
\]

Applying Newton’s Binomial Theorem to the righthand side of (10) gives

\[
\alpha^y = 1 + \frac{j}{1!} \left(\frac{ky}{j}\right) + \frac{j(j-1)}{2!} \left(\frac{ky}{j}\right)^2 + \cdots.
\]

Since \( j \) is to be regarded as very large, Euler puts

\[
1 = \frac{j-1}{j} = \frac{j-2}{j} = \frac{j-3}{j} = \cdots,
\]

allowing (11) to simplify to

\[
\alpha^y = 1 + ky + k^2 \frac{y^2}{2!} + k^3 \frac{y^3}{3!} + \cdots.
\]

Note that in the case \( k = 1 \) (that is, in the case \( \alpha = e \)) the series simplifies considerably to

\[
e^y = 1 + y + \frac{y^2}{2!} + \frac{y^3}{3!} + \cdots,
\]

which is the series for \( e^y \) which you obtained in your calculus course.

**Your Turn** 5. List as many aspects of Euler’s argument as you can which are not quite mathematically correct. Explain your reasoning.

### 2.4. Exercises.

1. In this exercise we compute \( \log 5 \) by way of series expansion of the natural logarithm.
   
   (a) Use \( x = 1/3 \) and the first four terms of the series given in (6) to approximate \( \ln 2 \).
   
   (b) Use \( x = 1/9 \) and the first four terms of the series given in (6) to approximate \( \ln 5/4 \).
   
   (c) Use the results of parts (a) and (b) together with properties of the logarithm to approximate both \( \ln 5 \) and \( \ln 10 \).
   
   (d) Finally, approximate \( \log 5 \) by using the part (d) of Theorem 10 from Chapter 9. (Perhaps this seems like a long process, but it is far shorter than the path Briggs took to compute \( \log 5 \).)

2. Here we approximate \( \log 24 \).
2. LOGARITHMS AND HISTORY: THE NATURAL LOGARITHM

(a) Use \( x = 5/7 \) and the first four terms of the series given in (6) to approximate \( \ln 6 \).

(b) Use part (a) together with part (a) of the previous exercise to approximate \( \ln 24 \). (The properties of the logarithm will be helpful here.)

(c) Use part (b) together with part (c) of the previous exercise to approximate \( \log 24 \). (Part (d) of Theorem 10 from Chapter 9 will be helpful here.)

3.

(a) Remind yourself of the reason that \( \int_1^e \frac{dt}{t} = 1 \). (You will need Theorem 1.)

(b) Part (a), together with a computer algebra system, gives a way to introduce students to a decimal approximation of \( e \). The goal is to find the number \( e \) such that \( \int_1^e \frac{dt}{t} = 1 \). Using a computer, one can see that \( \int_1^3 \frac{dt}{t} > 1 \) and \( \int_1^2 \frac{dt}{t} < 1 \), so \( 2 < e < 3 \). By trial and error, and without peeking at the known numerical value of \( e \), use a computer to estimate \( e \) to two decimal places.

(c) Similarly, using a computer, find the value \( x \) such that \( \int_1^x \frac{dt}{t} = 2 \) and the value \( y \) such that \( \int_1^y \frac{dt}{t} = 3 \).

(d) How are your answers in (c) related to \( e \)?

4. Use a computer algebra system to compute approximate values of \( e \) to 10,000 decimal places. (In Mathematica, the syntax is \( \text{N}[\text{E}, n] \), where \( n \) specifies the number of digits.) Did this take long? About how long does it take Mathematica to compute \( e \) to 100,000 decimal places?

5. In this exercise, we show that Equation 6 is effective for computing the natural logarithm of any positive real number.

(a) Show that the series \( 2 \left[ x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots \right] \) converges for all \( x \) with \( |x| < 1 \). (Hint: recall the notion of “radius of convergence” from calculus.)

(b) Show that every real number \( y > 0 \) can be written as \( \frac{1+x}{1+x} \) for some \( x \) with \( |x| < 1 \).

6. Assuming that one may extract roots to obtain \( 24^{1/8192} = 1.00038802 \), use the log table for roots of 10 to approximate \( \log 24 \) as Briggs would have. How does your answer compare to the one given by your calculator?
7. Let \( f(x) \) be as in Theorem 1. In order to show that \( f(x) = \log_\alpha x \), explain why it suffices to show that \( f(\alpha^\gamma) = \gamma \) for each \( \gamma \in \mathbb{R} \).

8. Let \( f(x) \) be as in Theorem 1, and suppose \( f(\alpha^r) = r \) for all positive real numbers \( r \). Show that the same result holds for negative values of \( r \). (Hint: \( \alpha^{-r} \cdot \alpha^r = 1 \).)

9. In (2), where is the continuity of \( \alpha^x \) being used? Where is the continuity of \( f(x) \) being used?

10. Suppose \( n \) is a positive integer.
   
   (a) Show that Newton’s Binomial Theorem (Theorem 2) implies
   
   \[
   (1 + x)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + \cdots + \binom{n}{n-1}x^{n-1} + \binom{n}{n}x^n,
   \]
   
   where \( \binom{n}{k} = \frac{n!}{(n-k)!k!} \). (Remember that \( 0! = 1 \).
   
   (b) Use induction to prove Newton’s Binomial Theorem (in the case when the exponent is a positive integer).

11. It is well known that \( \int_{-\infty}^{\infty} e^{-x^2} \, dx = \sqrt{\pi} \). Use this fact to show that \( \int_0^1 \sqrt{-\ln x} \, dx = \sqrt{\pi}/2 \). (Think about inverse functions.)

12. Using Taylor’s Theorem, how many terms must be taken in the power series expansion of \( e^x \) to obtain an approximation for \( e \) that is correct within \( 5 \times 10^{-10} \)?

13. In 1873 Charles Hermite (1822-1910) showed that \( e \) is transcendental. Proving this fact is beyond the scope of the text, but using a series expansion of \( e^x \) we can show that \( e \) is irrational.
   
   (a) Use a series expansion for \( e \) to show that \( e < 3 \). (Compare \( 1 + 1/1! + 1/2! + \cdots \) to a geometric series for which you can calculate the exact sum.)
   
   (b) Use the estimate \( e < 3 \) to show that
   
   \[
   0 < e - \left( 1 + \frac{1}{1!} + \cdots + \frac{1}{n!} \right) < \frac{3}{(n+1)!}.
   \]
   
   (c) For a contradiction, assume that \( e \) is rational, say \( e = a/b \) where \( a, b \) are positive integers. Pick \( n > b \) and also \( n > 3 \). Substitute \( e = a/b \) into the inequality given in part (b) and multiply the
inequality by \( n! \). Prove that this implies the existence of an integer \( N \) with \( 0 < N < 3/4 \). Conclude that \( e \) is irrational.

3. Compelling Properties of the Natural Logarithm and Natural Exponential Function

In Chapter 9 we defined the exponential function \( \alpha^x \), where \( \alpha \) is any positive real number not equal to one. From it, we obtain the logarithmic function \( \log_\alpha x \). On the other hand, we saw in Section 2.1 that the function \( \ln x \), defined to be \( \int_1^x \frac{dt}{t} \), really is a logarithm function relative to a particular base (namely, the real number \( e \)). This singles out a particular exponential function \( (e^x) \) and its inverse function \( (\ln x) \).

Indeed, in calculus and elsewhere, the most commonly used exponential function is the “natural exponential function” \( e^x \). With a name involving the word “natural” we expect \( e^x \) and its inverse \( \ln x \) (shorthand for \( \log_{e} x \)) to have some special characteristics which make them especially appealing among exponential and logarithmic functions, respectively. Unfortunately, these characteristics elude a first glance; in fact, in a world where we seem to prefer rational numbers to irrational ones, our first impressions of \( e \) may be decidedly uncomfortable, as not only is \( e \) irrational, it is in fact transcendental (see Section 1, Exercise 13). In this section, our goal is to ferret out compelling reasons for referring to \( e^x \) and \( \ln x \) as “natural” functions. We will verify that \( \ln x \) and \( e^x \) have properties which make them easier to use than other logarithmic and exponential functions. The most compelling reason is that the derivatives of the natural exp and log functions are simpler and more beautiful than for other bases. Moreover, we collect some intriguing facts about the number \( e \).

3.1. Logarithms with base \( \alpha \), revisited. Once the natural logarithm is firmly in our grasp, we may use it to express any logarithm in terms of an integral. Using the Golden Rule of Logarithms (part (d) of Theorem 10 from Chapter 9), we may write

\[
\log_\alpha x = \frac{\ln x}{\ln \alpha} = \frac{1}{\ln \alpha} \int_1^x \frac{1}{t} dt.
\]

Applying the Fundamental Theorem of Calculus in (13) gives

---

\(^{12}\)The notation “\( e \)” is due to Euler. The irrationality of \( e \) was first demonstrated by Euler in 1737, while Charles Hermite (1822-1910) verified that \( e \) is transcendental.
Of course from Equation (14) we see that in the case \( \alpha = e \) we have the simple formula \( \frac{d}{dx} \ln x = 1/x \), which supports our assertion that \( e \) is ideal as a logarithm base.

3.2. Inverse functions and their derivatives; the derivative of exponential functions. A (bijective) function and its inverse are closely linked: the behavior of one determines in large measure the behavior of the other. Therefore, we expect pairs of inverse functions, such as \( x^3 \) & \( x^{1/3} \), \( \log_2 x \) & \( 2^x \), and \( 1/x \) & \( 1/x \), to exhibit similar mathematical traits. For example, if an invertible function \( f \) is differentiable, we might reasonably expect its inverse \( f^{-1} \) to be differentiable also. In fact, this turns out to be the case in many instances.

**Theorem 3.** Suppose \( f \) is a one-to-one continuous function defined on an open interval \( I \). If \( f \) is differentiable at \( x_0 \in I \) and if \( f'(x_0) \neq 0 \), then \( f^{-1} \) is differentiable at \( y_0 = f(x_0) \) and

\[
(f^{-1})'(y_0) = \frac{1}{f'(x_0)}.
\]

A rigorous proof of this theorem will perhaps take us too far afield into the land of advanced calculus, so we omit it here. However, if we assume that \( (f^{-1})'(y_0) \) exists, then we can quickly check the validity of the derivative formula in Theorem 3. Also, if one strays from the hypotheses of Theorem 3, then \( f^{-1} \) can be poorly behaved (see Exercises 7 and 2).

We are now in a position to find the derivative of the function \( \alpha^x \). Thanks to (14), we apply Theorem 3 with \( \log_\alpha x \) in place of \( f \) in the theorem to obtain

\[
(15) \quad \frac{d}{dx}\alpha^x = \alpha^x \ln \alpha,
\]

and for \( \alpha = e \) this further simplifies to

\[
\frac{d}{dx}e^x = e^x.
\]

(You will be asked for a proof of (15) in Exercise 1.)
3.3. More on the number $e$. In previous sections, we saw that the number $e$ is the base of the natural logarithm, and is defined so that the area of the region bounded by $x = 1, x = e, y = 1/x$, and $y = 0$ is equal to 1 (see Figure 5). Using Mathematica, we learn that $e \approx 2.718281828459045$. But what else do we know about $e$ and $e^x$? In this section we investigate $e$ a bit more deeply.

**Theorem 4.** \( \lim_{h \to 0} (1 + h)^{1/h} = e. \)

**Proof.** Since $\ln x = \int_1^x 1/t \, dt$, by the Fundamental Theorem of Calculus we know $\ln'(x) = 1/x$, and so $\ln'(1) = 1$. On the other hand, using the definition of the derivative and properties of logarithms, we have

$$\ln'(1) = \lim_{h \to 0} \frac{\ln(1 + h) - \ln(1)}{h} = \lim_{h \to 0} \frac{1}{h} \ln(1 + h) = \lim_{h \to 0} \ln \left( (1 + h)^{1/h} \right).$$

Therefore $1 = \lim_{h \to 0} \ln \left( (1 + h)^{1/h} \right)$, and applying the natural exponential function on both sides of the previous equation yields

\[ e = e^{\lim_{h \to 0} \ln((1+h)^{1/h})}. \]

Finally, since the natural exponential function is continuous, in Equation (16) we may extract the limit from the argument of the exponential function to obtain

\[ e = e^{\lim_{h \to 0} \ln((1+h)^{1/h})} = \lim_{h \to 0} e^{\ln((1+h)^{1/h})} = \lim_{h \to 0} (1 + h)^{1/h}, \]

concluding the proof. \qed

**Corollary 5.**
(a) \( e = \lim_{n \to \infty} (1 + 1/n)^n. \)

(b) \( e^x = \lim_{n \to \infty} (1 + x/n)^n \) for all real numbers \( x. \)

You will be asked for a proof of Corollary 5 in Exercise 3.

These formulae have interesting practical interpretations that are discussed in Section 4.2.

3.4. Summary. In summary, we have discussed several different ways to view the number \( e. \) They are:

- The number \( e \) determines a region of area 1 under the hyperbola \( y = 1/t. \) (See Figure 5.)
- The number \( e \) is the sum of an infinite series. Specifically,
  \[
  e = 1/0! + 1/1! + 1/2! + 1/3! + 1/4! + \cdots .
  \]
- The number \( e \) is expressible as a limit, namely
  \[
  e = \lim_{h \to 0} (1 + h)^{1/h}.
  \]
- Given a base \( \alpha \) (\( \alpha > 0 \) and \( \alpha \neq 1 \)), the derivative of \( \alpha^x \) is a constant times \( \alpha^x \), and the constant is 1 exactly when \( \alpha = e. \)
- Given a base \( \alpha, \) the derivative of \( \log_\alpha x \) is a constant times \( 1/x, \) and the constant is 1 exactly when \( \alpha = e. \)

3.5. Exercises.

1. Use Theorem 3 to prove the formula given in Equation (15).

2. Consider \( f(x) = x^3 \) (defined on \( \mathbb{R} \)) and the branch function \( g(x) \) given by

\[
g(x) = \begin{cases} 
  x & \text{if } x \in (0, 1]; \\
  x - 1 & \text{if } x \in (2, 3). 
\end{cases}
\]

(a) Find inverses for both \( f \) and \( g. \) Be sure to specify the domain of each inverse function.

(b) Is \( f^{-1} \) differentiable on its domain? If not, does this contradict Theorem 3? Why?

(c) Is \( g^{-1} \) differentiable on its domain? If not, does this contradict Theorem 3? Why? (In order for a function \( h \) to be differentiable at a point \( x_0, \) \( h \) must first be defined in an open interval containing \( x_0). \)
3. Prove Corollary 5. (Part (a) follows from Theorem 4 by a simple substitution. Part (b) then follows from part (a) by another substitution.)

4. Use Corollary 5(b) to find a formula that expresses \( \ln y \) as a limit.

5. In Section 3.1 above we found the derivative of \( \log_\alpha x \) (Fundamental Theorem of Calculus) and then used Theorem 3 to find the derivative of \( \alpha^x \). In this exercise, we start with the definition of \( \alpha^x \), find its derivative, and then use the theorem about inverse functions to find the derivative of \( \log_\alpha x \).

   Let \( f(x) = \alpha^x \).

   (a) Using the definition of the derivative, show that \( f'(0) = \lim_{h \to 0} \frac{\alpha^h - 1}{h} \).

   (b) Using the definition of the derivative and (a), show that \( f'(x) = f'(0) \cdot f(x) \). (You should use the definition of the derivative, not Equation (15).)

   (c) Using Theorem 3 and (b), show that the derivative of \( \log_\alpha x \) is \( \frac{1}{x} \cdot f'(0) \).

   (d) Comparing (c) with Equation (14), what do you conclude about \( \lim_{h \to 0} \frac{\alpha^h - 1}{h} \)?

6. We have seen that there are several useful expressions for \( e \) as a limit, namely, \( e = \lim_{h \to 0} (1 + h)^{1/h} \) and \( e = \sum_{k=0}^{\infty} \frac{1}{k!} \). In this exercise, we investigate the usefulness of these expansions for purposes of computation.

   You will need to use the value of \( e \) produced by your calculator or computer.

   (a) Using a calculator, compute \( (1 + h)^{1/h} \) for \( h = 1, 1/2, 1/10, 1/100, 1/1000 \). When \( h = 1000 \), what percent error is made approximating \( e \)?

   (b) Find a value of \( h \) so that the approximation \( (1 + h)^{1/h} \) is correct to the fifth decimal place (that is, the approximation is 2.71828xxxxxx...).

   (c) Compute \( \sum_{k=0}^{n} \frac{1}{k!} \) for \( n = 0, 1, 2, 3, 4, 5 \). What percent error is made in approximating \( e \) when \( n = 5 \)?

   (d) Find the smallest \( n \) such that \( \sum_{k=0}^{n} \frac{1}{k!} = 2.71828xxxxxx... \).

7. Suppose \( f \) is as in Theorem 3, and further assume that \( (f^{-1})'(y_0) \) exists. Show that

\[
(f^{-1})'(y_0) = \frac{1}{f'(x_0)}.
\]

(Consider the fact that \( f^{-1}(f(x)) = x \) for \( x \in I \). Now differentiate both sides at \( x = x_0 \).)
4. Applications of Exponential Functions

4.1. Exponential growth. Exponential functions model several very common situations in nature, namely, those in which the rate of growth of a quantity is proportional to the quantity. The following theorem explains why exponential functions occur in this setting:

**Theorem 6.** Let $f$ be a differentiable real-valued function, defined on an open interval, and let $c \in \mathbb{R}$. Then $f'(x) = cf(x)$ if and only if $f$ is a constant multiple of $e^{cx}$.

**Proof.** On the one hand, if $f(x) = ke^{cx}$, then $f'(x) = kce^{cx} = cf(x)$. On the other hand, suppose that $f'(x) = cf(x)$. We consider the function $f(x)/e^{cx}$; its derivative is $\frac{e^{cx}f'(x) - f(x)e^{cx}}{(e^{cx})^2} = 0$. Hence $f(x)/e^{cx}$ is a constant. \qed

In most applications, the independent variable is time, and it’s more traditional to denote the variable by $t$. The theorem becomes: if $f'(t) = cf(t)$, then $f(t) = ke^{ct}$ for some constant $k$.

In terms of applications, we see that the theorem applies in any situation where the rate of change of $f(t)$ is a multiple of $f(t)$. Here are some classic examples:

**Continuously compounded interest.** Let $P(t)$ be the value of the principal (value of the bank account at time $t$). Note $P'(t) = rP(t)$ for some constant $r$, so by Theorem 6, $P(t) = ke^{rt}$ for some constant $k$. Setting $t = 0$, we see that $k = P(0)$, so finally, we obtain the formula $P(t) = P(0)e^{rt}$. We haven’t said why we call this “continuous compounding.” We will address this in Section 4.2.

**Exponential growth of populations.** Suppose $f(t)$ is the population of some species. Under certain circumstances, the rate of change of the population is proportional to the population. This happens for bacteria reproducing in some medium by division (at least if the growth is not constrained by crowding or lack of nutrients). By the same argument as above, we have that $f(t) = f(0)e^{rt}$.

**Radioactive decay.** Let $f(t)$ be the number of radioactive atoms in some sample. The radioactive atoms decay at some rate that is proportional to the number of radioactive atoms, that is, to $f(t)$ itself. It follows as above that $f(t) = f(0)e^{rt}$ for some $r$; the difference in this case is that $r$ is negative (it’s the instantaneous rate of change of $f(t)$, which is negative since $f$ is a decreasing function).
4.2. **Compound interest.** The example of “continuously compounded interest,” mentioned above, is mathematically rich. Let’s start by thinking about this example from a practical point of view. We probably all know what is meant by “compound interest,” but what could we possibly mean by “continuous compounding”? First, let’s realize that even if your bank “compounds continuously,” there isn’t a number dial at the bank that shows your account value going up like the number dial on a gas pump. Realistically, your bank pays you interest at certain regular intervals (perhaps once per month). So what does it mean to “compound continuously”?

We begin by recalling how to compute compound interest that is computed at discrete intervals. The formula for compound interest is well-known. As you may recall, either from Algebra II or some business math class, the pertinent formula is

\[
P = P_0(1 + r/n)^{nt},
\]

where \( P_0 \) is the initial investment, \( r \) is the annual interest rate, \( n \) is the number of compoundings per year, and \( P \) is the value of the investment after \( t \) years. To illustrate how to use this formula, suppose Bart Simpson extorts $4000 from Krusty the Clown, and puts it in an account earning 5% annual interest compounded quarterly (that is, compounded four times per year). The value of his investment after 6 years is

\[
P = 4000(1 + 0.05/4)^{4 \cdot 6} = 5389.40.
\]

Much more than being able to use the formula, we need to understand why it works.\(^{13}\) We now ask you to provide an explanation:

**Your Turn** 6. *How could you explain the validity of Equation 17 to someone unfamiliar with this equation? (Perhaps start with a specific case like \( r = 12\% \), \( n = 4 \), \( t = 2 \), and \( P_0 = \$1000 \), or the example with Bart and Krusty.)*

According to Exercise 3 below, our investment returns increase as the number of compoundings per year increases. Operating under the principle that there is never too much of a good thing, we should achieve

\[^{13}\text{Thirty years later, one of the authors can still recall the mathematical nadir of his K-8 school experience: spending days upon days in eighth grade, plugging numbers into the formula to compute interest, without any hint of why this could be interesting.}\]
the ideal optimal return on our investment if we let the number of compoundings per year tend to infinity 
(which is what banks routinely do nowadays). This is the precise meaning of *continuous compounding*.

Continuous compounding is closely related with the number $e$. From Corollary 5, we deduce that under 
continuous compounding, an initial investment of $P_0$ at annual rate $r$ invested for $t$ years will have a final 
value of $P$ dollars, where

$$P = P_0 e^{rt}. \tag{18}$$

**Your Turn 7. Fill in the details in the derivation of Equation 18.**

Therefore, we may think of $e$ as the amount of money an initial investment of one dollar will earn in one 
year, given that the annual rate is 100% (i.e., $r = 1$) and interest is compounded continuously.

4.3. Exercises.

1. It is well-known that “money grows faster when interest is compounded more often,” since one earns 
“interest on the interest.” Consider the $4000 which Bart has extorted from Krusty. Suppose the annual 
rate is 5%.

   (a) What is the value of Bart’s investment after 6 years if interest is compounded monthly?
   
   (b) What is the value of Bart’s investment after 6 years if interest is compounded weekly?

   (c) What is the value of Bart’s investment after 6 years if interest is compounded daily?

   (d) If everything except the number of compoundings per year is held constant, what seems to be 
happening to the final value of the investment as the number of compoundings increases?

2. Consider the $4000 which Bart has extorted from Krusty. Suppose the annual rate is 5%. What is 
the value of Bart’s investment after 6 years if interest is compounded continuously? Compare with your 
answers to Exercise 1.

3. From Exercise 1 we see that the return on our investment seems to increase as the number $n$ of 
compoundings per year increases. Let’s use our calculus skills to verify this, by showing that the function

$$f(x) = (1 + r/x)^x,$$
where $r$ is a fixed positive real number, is increasing on the interval $(0, \infty)$. (Take a derivative, but be careful: you’ll need to use logarithmic differentiation.)

4. Suppose you intend to deposit $1000 into an account each year on January 1, for five years. Your bank pays 4%, compounded continuously.

(a) How much money will you have at the end of five years?

(b) What fraction of your money (after five years) came from the first $1000 you deposited and the interest due to it?

5. A bank advertises a checking account with an “interest rate of 2% compounded continuously, with an effective annual yield of 2.02013%.” Explain carefully what this means in the context of the formula $P(t) = P(0)e^{rt}$.

6. We’ve said that banks don’t pay you interest continuously, even if they use continuous compounding in their calculations. Suppose you deposit $1000 in a savings account on January 1, and the bank offers an interest rate of 3%, compounded continuously but only paid four times per year. What do you expect your account balance to be after 3 months, 6 months, 9 months, and one year?

7. Suppose that you start with $1000 and your goal is to earn (at least) $1000 in interest.

(a) If the interest rate is 10% and interest is compounded yearly, how many years must you wait? What if interest is compounded daily?

(b) Suppose you want to reach your goal in 5 years. If interest is compounded yearly, what must the interest rate be? What if the interest is compounded daily?

(c) If the interest rate is 10%, how many times per year would interest have to be compounded for you to reach your goal in 7 years? (You may not be able to solve this problem algebraically. Try using technology.)

(d) If the interest rate is 10%, show that you can’t reach your goal in 6 years, no matter how often interest is compounded.
8. Here is another way of looking at how compounding actually contributes to wealth. Consider the function \( f(n) = (1 + \frac{1}{n})^n \).

(a) Interpret \( f(n) \) in financial terms. (What is the interest rate? What is the length of time that the investment is made?)

(b) Using technology, plot the function \((1 + 1/n)^n\) for \( n \geq 1 \). Recall that \( \lim_{n \to \infty} = e \).

(c) \( f(1) \) is what fraction of the maximum value of \( f \)?

(d) \( f(12) \) is what fraction of the maximum value of \( f \)?

(e) \( f(365) \) is what fraction of the maximum value of \( f \)?

9. Suppose that a certain sample is composed of 50% radioactive atoms at time \( t = 0 \). It is declared “safe” when only 0.001% of its atoms are radioactive. At time \( t = 1 \), suppose that 25% of its atoms are radioactive. At what time will it be “safe”?

10. The formula \( P(t) = P(0)e^{rt} \) for a biological population is clearly incorrect since \( \lim_{t \to \infty} P(t) = \infty \), which is not realistic. (It is easy to come up with word problems where after some period of time, the number of bacteria allegedly in a test tube is greater than the number of atoms in the known universe). So, instead of modeling the population with the differential equation \( P'(t) = kP(t) \), it’s customary to use the equation \( P'(t) = kP(t)(L - P(t)) \), with \( k > 0 \) and \( L > 0 \). Here \( L \) is regarded as a sort of bound on the size of the population. If \( P(t) \) is close to zero, then \( P'(t) \) is small and essentially proportional to \( P(t) \) as with exponential growth (since \( P'(t) \approx k \cdot P(t) \cdot L \)). Likewise, when \( P(t) \) is close to \( L \), then \( P'(t) \) is small and essentially proportional to \( L - P(t) \), since \( P'(t) = k \cdot L \cdot (L - P(t)) \). This “puts the brakes” on \( P(t) \) when \( P(t) \) is close to \( L \), which keeps \( P(t) \) from going to infinity.

(a) Show that \( P(t) = \frac{L e^{kt} \cdot L - P(t)}{e^{kt} + 1} \) is a solution to the differential equation \( P'(t) = kP(t)(L - P(t)) \).

(b) Assuming that \( P(t) \) is never equal either to 0 or \( L \), show that every solution of \( P'(t) = kP(t)(L - P(t)) \) is of the form in (a). (Hint: start by writing \( \frac{P'(t)}{P(t)(L - P(t))} = k \) and use a partial fractions decomposition.)

(c) For \( P(t) \) as in (a), prove that \( \lim_{t \to \infty} P(t) = L \).

(d) Using Mathematica, compare the graphs of \( e^t \) and \( \frac{2e^{2t}}{e^{2t} + 1} \) on the intervals \([0, 0.1] \), \([0, 0.2] \), and \([0, 2] \).
Transcendental Functions and Complex Numbers

In algebra, precalculus, and calculus, one studies many functions: polynomials, exponential and logarithmic functions, (circular) trigonometric functions like \( \sin x \) and \( \cos x \), and perhaps the hyperbolic trigonometric functions \( \sinh x \) and \( \cosh x \). At first, these functions may seem to be simply a collection of unrelated examples. However, in a student’s mathematical education, certain connections among the functions emerge:

- Exponential and logarithmic functions are inverses of each other.
- The hyperbolic trigonometric functions are related to exponential functions by the rules \( \cosh x = \frac{e^x + e^{-x}}{2} \) and \( \sinh x = \frac{e^x - e^{-x}}{2} \).
- The exponential / trigonometric / hyperbolic trigonometric functions are related to polynomials. In fact, these transcendental functions are not polynomials, but they are “the next best thing”: they can be given by power series, which are infinite analogs of polynomials. For example, in calculus, we learn that for all real values of \( x \),

\[
\begin{align*}
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \frac{x^7}{7!} + \ldots \\
\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots \\
\sinh x &= x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \ldots
\end{align*}
\]

Let’s Go 1. Can you give a quick reason why \( \sin x \) can’t possibly be a polynomial function of \( x \)?

The purpose of this chapter is to explore the connection between the exponential and hyperbolic trigonometric functions on the one hand, and the circular trigonometric functions on the other hand. The strategy for linking these functions is to extend all of them to the complex numbers.
Why should there be a connection among these functions? The following Let’s Go activities will set the stage.

**Let’s Go 2.** Replace \( x \) with \( ix \) in the power series for \( \sinh x \). Simplify the expression and compare it to the power series for \( \sin x \).

**Let’s Go 3.** Consider the function \( f : \mathbb{R}^2 \to \mathbb{R}^2 \) defined by the rule

\[
f(x, y) = (\sin x \cosh y, \cos x \sinh y).
\]

(a) Several curves appear in Figure 1. Match each of them with the following functions of a single variable (the curve is the range of the function, as the variable runs through all real numbers).

1. \( f(x, 0) \)
2. \( f(x, \ln 2) \)
3. \( f(x, \ln 3) \)
4. \( f(x, \ln 4) \)
5. \( f(0, y) \)
6. \( f(\pi/6, y) \)
7. \( f(\pi/4, y) \)
8. \( f(\pi/3, y) \)

(b) Find the equation of each of the curves in (a), using \( u \) and \( v \) as coordinates in the plane, and give a geometric description of each curve.

In this chapter we also return to one of our goals from Chapter 9—defining \( z^w \) for complex values of \( z \) and \( w \). The course of this investigation naturally leads us to the complex exponential and trigonometric functions, and the connections among them.

1. **The Functions** \( f(z) = z^n \)

---

1. If we identify \( \mathbb{R}^2 \) with \( \mathbb{C} \), then \( f \) is the (complex) sine function, given by the power series \( z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \ldots \) (See Section 4, Exercise 18.)
1. THE FUNCTIONS $f(z) = z^n$

1.1. The complex case. In Chapter 9 we defined the number $z^n$ for any complex number $z$ and any positive integer $n$. Consider the function $f: \mathbb{C} \rightarrow \mathbb{C}$, $f(z) = z^n$, where $n$ is a fixed positive integer.

If $n = 1$, then this function is simply the identity function $f(z) = z$. Our goal now is to understand the very interesting behavior of $f(z) = z^n$ when $n \geq 2$. We can’t “graph” such functions in the usual way, since that would require four real dimensions (why?). Nevertheless, the functions can be understood with the aid of some pictures.

To begin, the functions $f(z) = z^n$ is easiest to understand if write $z$ in polar form, so let us write $z = r(\cos \theta + i \sin \theta)$. De Moivre’s Theorem (Chapter 4, Equation (9)) tells us that $z^n = r^n(\cos(n\theta) + i \sin(n\theta))$. That is, to obtain the $n$th power of $z$, we raise the length of $z$ to the $n$th power to get the length of $z^n$, and we multiply the angle of $z$ by $n$ to get the angle of $z^n$. It is important to get a sense of what is going on geometrically:

**Your Turn 1.** Sketch the image of each region in Figure 2 under the function $f(z) = z^2$. Then do the same for $f(z) = z^3$.

Using the experience gained in Your Turn 1, the following theorem emerges:
THEOREM 1. Let \( n \) be a positive integer. Let \( D_n = \{ r(\cos \theta + i \sin \theta) : r \geq 0 \text{ and } -\pi/n < \theta \leq \pi/n \} \).

Let \( f \) be the function \( f(z) = z^n \). Then

1. The restriction of \( f_n \) to \( D_n \) is one-to-one.
2. \( f_n \) maps \( D_n \) onto \( \mathbb{C} \).

1.2. Exercises.

1. Sketch \( D_1, D_2, D_3, D_4, \) and \( D_5 \), and explain why Theorem 1 appears to be correct for \( n = 1, 2, 3, 4, 5 \).

2. Consider the function \( f(z) = z^3 \) (where \( z \in \mathbb{C} \)). Show that \( f \) is not one-to-one as a function with domain \( \mathbb{C} \).

3. Prove Theorem 1. (Hint: Use De Moivre’s Theorem.)

2. Roots of Complex Numbers

The goal of this section is to define and understand expressions like \( z^{1/2}, z^{1/3} \), and more generally, \( z^{1/q} \) where \( z \) is a complex number and where \( q \) is a positive integer.
2. Roots of Complex Numbers

2.1. Examples of roots. Let’s begin by trying to find the fourth roots of 1. By examining where the graph of \( y = x^4 \) intersects the line with equation \( y = 1 \), we see that \( \pm 1 \) are the only real fourth roots of 1.

We know further that \( i^4 = 1 \) (why?) and likewise \((-i)^4 = 1\). Thus 1, \( i \), \(-1 \), and \(-i \) are fourth roots of 1.

Why might we believe that \( \pm 1, \pm i \) are the only fourth roots of 1? One reason is that a fourth root of 1 has to be a root of the polynomial \( x^4 - 1 \), and since this polynomial has degree four, it has at most four distinct roots in \( \mathbb{C} \). Even more to the point, it is easy to factor \( x^4 - 1 \):

\[
x^4 - 1 = (x^2 - 1)(x^2 + 1) = (x - 1)(x + 1)(x - i)(x + i),
\]

showing that the four 4th roots of 1 are \( 1, -1, i, \) and \(-i \).

We will not be so lucky in general, if we just hope to find roots by factoring. For example, if our goal were to find the 5th roots of 1, what could we do? We’d be looking for the roots of the polynomial \( x^5 - 1 \), so we would know that there are at most five such roots in \( \mathbb{C} \). Since 1 is itself a 5th root of 1, we could could factor \( x^5 - 1 \) using \( x - 1 \) as a factor, obtaining \( x^5 - 1 = (x - 1)(x^4 + x^3 + x^2 + x + 1) \). It would take some inspiration to factor \( x^4 + x^3 + x^2 + x + 1 \) as a product of four linear factors (more inspiration than most of us could summon up quickly).

So how can we find the 5th roots of 1 quickly—or indeed the \( q \)th roots of any complex number \( z \)? The trick is to use the polar form of complex numbers. Let’s return to the example of the fourth roots of 1. Say that \( \omega \) is some complex fourth root of 1. We can write

\[
\omega = r(\cos \theta + i \sin \theta)
\]

where \( r > 0 \). Then \( \omega^4 = r^4(\cos(4\theta) + i \sin(4\theta)) \), by DeMoivre’s Theorem. By assumption, \( \omega^4 = 1 \), and the polar form of 1 is \( 1(\cos 0 + i \sin 0) \). Thus we have

\[
r^4(\cos(4\theta) + i \sin(4\theta)) = 1(\cos 0 + i \sin 0).
\]

This is equivalent to

\[
r^4 = 1 \quad \text{and} \quad 4\theta = 2\pi n \text{ for some } n \in \mathbb{Z}.
\]

\(^2\)Here we are appealing to the Factor Theorem: A number \( \alpha \) is a root of a polynomial if and only if \( x - \alpha \) is a divisor of that polynomial.
(We couldn’t conclude that $4\theta = 0$, but only that $4\theta$ is a multiple of $2\pi$—why?) Now since $r$ is a positive real number, if $r^4 = 1$ then $r = 1$ (why?), so we finally see that $\omega$ is a fourth root of 1 if and only if
\begin{align*}
    r = 1 \quad \text{and} \quad \theta = \frac{\pi}{2}n \quad \text{for some } n \in \mathbb{Z}.
\end{align*}
There are exactly four such points in $\mathbb{C}$; they are on the unit circle (since $r = 1$) at angles that are multiples of $\frac{\pi}{2}$. Of course, these numbers are just $1$, $i$, $-1$, $-i$, as we had found previously.

For similar reasons, we can find the 5th roots of 1 easily: they are the five points on the unit circle that make angles of $\frac{2\pi}{5}n$ with the positive real axis (for $n = 0, 1, 2, 3, 4$). In Figure 3, we see the $q$th roots of 1, for $q = 1, 2, \ldots, 10$. Note that 1 is always a $q$th root of 1; there are exactly $q$ $q$th roots of 1; and these roots are “evenly spaced” on the unit circle, separated by angles of radian measure $\frac{2\pi}{q}$.

**Your Turn 2.** Imitate the discussion above to find all the fourth roots of 81. (You should find four of them.) Write your answers in both polar form and $a + bi$ form.

As a very similar example, we can find the fourth roots of $-1$. Writing $-1$ in polar form, we have $-1 = 1(\cos \pi + i \sin \pi)$. Assuming that $\nu = r(\cos \theta + i \sin \theta)$ is a fourth root of $-1 = 1(\cos \pi + i \sin \pi)$, we have
\begin{align*}
    \nu^4 &= r^4(\cos(4\theta) + i \sin(4\theta)) = 1(\cos \pi + i \sin \pi),
\end{align*}
so
\begin{align*}
    r^4 &= 1 \quad \text{and} \quad 4\theta = \pi + 2\pi n \quad \text{for some } n \in \mathbb{Z},
\end{align*}

hence
\begin{align*}
    r &= 1 \quad \text{and} \quad \theta = \frac{\pi}{4} + \frac{\pi}{2}n \quad \text{for some } n \in \mathbb{Z}.
\end{align*}
Thus the fourth roots of 1 lie on the unit circle in $\mathbb{C}$, at angles $\frac{\pi}{4}$, $\frac{3\pi}{4}$, $\frac{5\pi}{4}$, and $\frac{7\pi}{4}$ (see Figure 4).

This looks very much like the picture of the fourth roots of 1, only rotated by $45^\circ$. This is no coincidence. If $\nu$ is a fourth root of $-1$ and $\omega$ is a fourth root of 1, then $\nu \omega$ is again a fourth root of $-1$, since $(\nu \omega)^4 = \nu^4 \omega^4 = (-1) \cdot 1 = -1$. (The converse is also true: if $\nu_1$ and $\nu_2$ are fourth roots of $-1$, then $\nu_2 = \nu_1 \omega$ where $\omega$ is a fourth root of 1—why?) Now $\nu_1 = \cos(\pi/4) + i \sin(\pi/4)$ is a fourth root of $-1$, so we get all the fourth roots of $-1$ by multiplying $\nu_1$ by the fourth roots of 1, which are 1, $i$, $-1$, and $-i$. This says that the fourth roots of $-1$ are obtained by rotating $\nu_1$ by 0, $\pi/2$, $\pi$, and $3\pi/2$, just as we calculated above.
2. ROOTS OF COMPLEX NUMBERS

Your Turn 3. Let $z$ be a complex number on the circle of radius 32 centered at the origin, and suppose $z$ makes an angle of 20° with the positive real axis. Find the five 5th roots of $z$ (in polar form). (What is the magnitude of each? What are the five possible angles that a 5th root of $z$ might make with the positive real axis?)

2.2. Main results. Let’s formalize and collect the facts that we’ve hinted at above. First, let’s describe the $q$th roots of 1:

Figure 3. $q$th roots of 1
Theorem 2. Let $q$ be a positive integer. Define $\omega_q = \cos \left( \frac{2\pi}{q} \right) + i \sin \left( \frac{2\pi}{q} \right)$. There are exactly $q$ complex $q$th roots of 1, namely the powers of $\omega_q$:

\[
\omega_q = \cos \left( \frac{2\pi}{q} \right) + i \sin \left( \frac{2\pi}{q} \right)
\]
\[
\omega_q^2 = \cos \left( 2 \cdot \frac{2\pi}{q} \right) + i \sin \left( 2 \cdot \frac{2\pi}{q} \right)
\]
\[
\omega_q^3 = \cos \left( 3 \cdot \frac{2\pi}{q} \right) + i \sin \left( 3 \cdot \frac{2\pi}{q} \right)
\]
\[
\vdots
\]
\[
\omega_q^{q-1} = \cos \left( (q-1) \cdot \frac{2\pi}{q} \right) + i \sin \left( (q-1) \cdot \frac{2\pi}{q} \right)
\]
\[
1 = \omega_q^q = \cos \left( q \cdot \frac{2\pi}{q} \right) + i \sin \left( q \cdot \frac{2\pi}{q} \right) = \cos 0 + i \sin 0
\]

Proof. Suppose $\cos \phi + i \sin \phi$ is a $q$th root of 1. This means $1 = (\cos \phi + i \sin \phi)^q = \cos(q\phi) + i \sin(q\phi)$, so $q\phi$ is a multiple of $2\pi$, or in other words, $\phi$ is a multiple of $\frac{2\pi}{q}$. But these are exactly the angles occurring in the complex numbers listed in the theorem. (We only need to consider the multiples $k \cdot \frac{2\pi}{q}$ for $0 \leq k \leq q - 1$; why?)

Finally, the equations in the statement of the theorem are a consequence of DeMoivre’s Theorem. □
Next, we can describe all the $q$th roots of any complex number. It’s easy to see that the only $q$th root of 0 is 0 (why?). For nonzero complex numbers, we have the following all-important definition and theorem:

**Definition 3.** Let $q$ be a positive integer, and let $z$ be a nonzero complex number. Write $z$ in polar form as $z = r(\cos \theta + i \sin \theta)$, where $r > 0$ and $-\pi < \theta \leq \pi$. Define $z^{1/q} = \sqrt[q]{r} \left(\cos \left(\frac{\theta}{q}\right) + i \sin \left(\frac{\theta}{q}\right)\right)$.

**Theorem 4.** Let $q$ be a positive integer, and let $z$ be a nonzero complex number.

(a) The complex number $z^{1/q}$ is a $q$th root of $z$, called the principal $q$th root of $z$.

(b) The complex number $z$ has exactly $q$ different complex $q$th roots. These are obtained by multiplying $z^{1/q}$ by each of the $q$th roots of 1. Using the polar form of $z$ and the $q$th roots of 1, the collection of all $q$th roots of $z$ can be described explicitly as

\[
\begin{align*}
z^{1/q} \cdot 1 &= \sqrt[q]{r} \left(\cos \left(\frac{\theta}{q}\right) + i \sin \left(\frac{\theta}{q}\right)\right) \\
z^{1/q} \cdot \omega_q &= \sqrt[q]{r} \left(\cos \left(\frac{\theta + 2\pi}{q}\right) + i \sin \left(\frac{\theta + 2\pi}{q}\right)\right) \\
z^{1/q} \cdot \omega_q^2 &= \sqrt[q]{r} \left(\cos \left(\frac{\theta + 2 \cdot 2\pi}{q}\right) + i \sin \left(\frac{\theta + 2 \cdot 2\pi}{q}\right)\right) \\
z^{1/q} \cdot \omega_q^3 &= \sqrt[q]{r} \left(\cos \left(\frac{\theta + 3 \cdot 2\pi}{q}\right) + i \sin \left(\frac{\theta + 3 \cdot 2\pi}{q}\right)\right) \\
z^{1/q} \cdot \omega_q^4 &= \sqrt[q]{r} \left(\cos \left(\frac{\theta + 4 \cdot 2\pi}{q}\right) + i \sin \left(\frac{\theta + 4 \cdot 2\pi}{q}\right)\right) \\
&\quad \vdots \\
z^{1/q} \cdot \omega_q^{q-1} &= \sqrt[q]{r} \left(\cos \left(\frac{\theta + (q-1) \cdot 2\pi}{q}\right) + i \sin \left(\frac{\theta + (q-1) \cdot 2\pi}{q}\right)\right)
\end{align*}
\]

**Proof.** First, let’s check all the equations in the statement of the theorem. They are all of the form $z^{1/q} \cdot \omega_q^k = \sqrt[q]{r} \left(\cos \left(\frac{\theta + k \cdot 2\pi}{q}\right) + i \sin \left(\frac{\theta + k \cdot 2\pi}{q}\right)\right)$. These are easy to verify, since

\[
\begin{align*}
z^{1/q} \cdot \omega_q^k &= \sqrt[q]{r} \left(\cos \left(\frac{\theta}{q}\right) + i \sin \left(\frac{\theta}{q}\right)\right) \cdot \left(\cos \left(\frac{k \cdot 2\pi}{q}\right) + i \sin \left(\frac{k \cdot 2\pi}{q}\right)\right) \\
&= \sqrt[q]{r} \left(\cos \left(\frac{\theta + k \cdot 2\pi}{q}\right) + i \sin \left(\frac{\theta + k \cdot 2\pi}{q}\right)\right)
\end{align*}
\]
This proves the equations. We still must verify that these q numbers are exactly the qth roots of z. First, we can show that
\[
\left( z^{1/q} \right)^q = \left( \sqrt[q]{r} \left( \cos \left( \frac{\theta}{q} \right) + i \sin \left( \frac{\theta}{q} \right) \right) \right)^q
\]
\[
= \left( \sqrt[q]{r} \cos \left( \frac{\theta}{q} \right) + i \sin \left( \frac{\theta}{q} \right) \right)^q
\]
\[
= r \left( \cos \left( q \cdot \frac{\theta}{q} \right) + i \sin \left( q \cdot \frac{\theta}{q} \right) \right)
\]
\[
= r \cdot (\cos \theta + i \sin \theta) = z
\]
Likewise, \( z^{1/q} \cdot \omega_q^k \) is a qth root of z, since
\[
\left( z^{1/q} \cdot \omega_q^k \right)^q = \left( z^{1/q} \right)^q \cdot (\omega_q^k)^q = z \cdot 1 = z.
\]
Thus, all of the numbers in the list in the theorem are qth roots of z. Conversely, suppose \( x \) is a qth root of z. Then
\[
\left( \frac{x}{z^{1/q}} \right)^q = \frac{x^q}{(z^{1/q})^q} = \frac{z}{z} = 1,
\]
so \( \frac{x}{z^{1/q}} \) is a qth root of 1. By Theorem 2, \( \frac{x}{z^{1/q}} = \omega_q^k \) for some k, so \( x = z^{1/q} \cdot \omega_q^k \). Thus every qth root of z occurs in the list in the theorem.

2.3. Principal roots. Something important happened in Definition 3 and Theorem 4: we singled out one of the q different qth roots of a, and used it to produce all the qth roots of a (by multiplying it by the various qth roots of 1).

Picking a “principal” root is simply a way of selecting one of the qth roots of a, for the sake of producing something definite. We’ve already made an attempt at this in Chapter 9 (Section 3) where we defined \( \sqrt[q]{a} \) and \( a^{1/q} \) in certain circumstances. Recall that these definitions are sometimes at odds with each other. For example, as we showed back in Chapter 9 (Section 3), \( (-1)^{\frac{1}{2}} \neq \sqrt[2]{-1} \). On the other hand, \( (1)^{\frac{1}{2}} = \sqrt[2]{1} \). Now, with the results of this section, we are able to make quick work of this project:

**Proposition 5.** If \( a > 0 \) and q is a positive integer, then \( \sqrt[q]{a} = a^{1/q} \).

**Proof.** Writing a in polar form as \( a(\cos 0 + i \sin 0) \), then by definition, \( a^{1/q} = \sqrt[q]{a} \left( \cos \frac{0}{q} + i \sin \frac{0}{q} \right) = \sqrt[q]{a} \).
Your Turn 4. In the proof, where did we use the assumption that \( a \) is positive?

**Proposition 6.** If \( a < 0 \) and \( q \) is an odd positive integer, then \( \sqrt[q]{a} \neq a^{1/q} \) unless \( q = 1 \).

**Proof.** Begin by writing \( a \) in polar form: we have \( a = (-a)(\cos \pi + i \sin \pi) \) (note that \( -a > 0 \)). By definition, \( a^{1/q} = \sqrt[q]{-a} \left( \cos \frac{\pi}{q} + i \sin \frac{\pi}{q} \right) \). If \( q = 1 \), this is equal to \( \sqrt[q]{a}(-1) \), which is \( \sqrt[q]{a} \) (why?). On the other hand, if \( q > 1 \), then \( a^{1/q} \) is not real but \( \sqrt[q]{a} \) is real, so \( \sqrt[q]{a} \neq a^{1/q} \). \( \square \)

**2.4. Exercises.**

1. Explain why there are no real fourth roots of \(-81\). Then, find all the complex fourth roots of \(-81\). Write them in polar form and in \( a + bi \) form. Graph them in the complex plane.

2. Find the complex sixth roots of 1. Write them in polar form and in \( a + bi \) form, and graph them in the complex plane.

3. Find the complex fourth roots of \(-1 + \sqrt{3}i\). Write them in polar form and in \( a + bi \) form, and graph them in the complex plane. (You should start by writing \(-1 + \sqrt{3}i\) in polar form. The angles in this problem are “nice.”)

4. What is the length of each complex fifth root of \(3 - 7i\)?

5. If \( z = \cos(3\pi/7) + i \sin(3\pi/7) \), find the fifth roots of \( z \) (in polar form).

6. Compute each of the following complex numbers, or explain why the number does not exist. The exercise is critical in making certain that you have understood the definitions of \( \sqrt[q]{r} \) and \( z^{1/q} \). Be careful to use the definitions from the text. Do not use a calculator.

   (a) \( \sqrt[4]{81} \)
   (b) \( 81^{1/4} \)
   (c) \( \sqrt[3]{5} \)
   (d) \( 8^{1/3} \)
   (e) \( \sqrt[5]{-1} \)
(f) \((-1)^{1/4}\)
(g) \(\sqrt[3]{-8}\)
(h) \((-8)^{1/3}\)

7. Let \(z = 1 + \sqrt{3}i\). Compute \(z^{1/2}\) (in polar form and \(a + bi\) form). Then, find all square roots of \(z\).

8. Let \(z = -1 - \sqrt{3}i\). Compute \(z^{1/4}\) (in polar form and \(a + bi\) form). Then, find all fourth roots of \(z\).

9. Draw (on separate copies of the unit circle in \(\mathbb{C}\)) all the \(q\)th roots of 1, for \(q = 6, 8,\) and 12.

10. Factor \(x^4 - 7\), writing it as the product of four degree-one polynomials.

11. Show that if \(z_1\) and \(z_2\) are the two square roots of a complex number \(w\), then \(z_2 = -z_1\). Give several explanations if you can.

12. Suppose a point \(z\) is marked in the complex plane. In words, explain how to mark all of the 5th roots of \(z\).

13. Let \(q\) be a positive integer. Define \(f : D_q \rightarrow \mathbb{C}, f(z) = z^q\), and let \(g : \mathbb{C} \rightarrow D_q, g(z) = z^{1/q}\). Explain why both \(f\) and \(g\) are one-to-one and onto, and by composing functions, show that \(f\) and \(g\) are inverse functions.

14. We consider the polynomial \(x^3 - 1\).

(a) Write down the three third roots of 1 in \(a + bi\) form. (You will use your knowledge of 30° - 60° - 90° triangles.)

(b) Using the third roots of 1, factor \(x^3 - 1\) into the product of three linear factors.

(c) We will obtain the factorization in a different way. First, realizing that \(x - 1\) must be a factor of \(x^3 - 1\) (why?), use long division to write \(x^3 - 1\) as \(x - 1\) times a degree-two polynomial. Finally, use the quadratic formula to find the roots of the degree-two polynomial.

15. We consider the polynomial \(x^3 + 1\), in a similar fashion as the last exercise.
(a) On the unit circle, mark the three complex numbers that are the roots of this polynomial. (They are the three third roots of −1.) Write the three numbers in $a + bi$ form. Finally, use these numbers to give the factorization of $x^3 + 1$ into degree-one factors.

(b) As an alternate strategy for factoring $x^3 + 1$: find an (obvious) real root of $x^3 + 1$. Use this root to write $x^3 + 1$ as the product of a degree-one factor and a degree-two factor. Finally, use the quadratic formula to find the roots of the degree-two factor.

16. Using the results of the previous two exercises, give the factorization of $x^6 − 1$ into degree-one factors.

17. By a method of your choice, give the factorization of $x^{12} − 1$ into degree-one factors. (The methods in the previous three problems will help.)

18. Factor $x^2 − (1 + \sqrt{3}i)$ into the product of two degree-one factors. (You will need to find the roots. Make certain that you write them in $a + bi$ form, not simply as $±\sqrt{1 + \sqrt{3}i}$.

19. Explain why every third root of 1 is an 18th root of 1. Then, generalize this statement.

20. Show that every 8th root of 1 is either a 4th root of 1 or a 4th root of −1.

21. If $z$ is a 6th root of 64, then $z$ is a square root of what possible numbers? Also, $z$ is a cube root of what possible numbers? Explain.

22. Draw each of the 12th roots of 1 on the unit circle in $\mathbb{C}$. Obviously the twelfth power of each of these numbers is 1, but in some cases, a lesser power is already 1. Label each 12th root of 1 (call it $z$) with the smallest positive integer $d$ such that $z^d = 1$.

23. Suppose the complex number $\omega$ is an $q$th root of 1. We say that $\omega$ is a primitive $q$th root of 1 if $\omega^q = 1$ but $\omega^m \neq 1$ if $1 \leq m < q$.

For example, −1 is a 2nd root of 1, a 4th root of 1, a 6th root of 1... and actually is an $n$th root of 1 for any even $n$. But −1 is a primitive 2nd root of 1 since 2 is the smallest power to which −1 is raised to give 1. Likewise, among the 5th roots of 1, all of them are primitive 5th roots of 1 except for 1 itself, which is a primitive 1st root of 1.
(a) Among the 6th roots of 1, which ones are primitive 6th roots of 1?
(b) Among the 12th roots of 1, which ones are primitive 12th roots of 1?
(c) Among the 30th roots of 1, which ones are primitive 30th roots of 1?
(d) Among the 31st roots of 1, which ones are primitive 31st roots of 1?

24. Recall that a monic polynomial is a polynomial for which the coefficient of the highest power of the variable is one. For example, \( x^3 - \pi x^2 + 0.9x - 13 \) is monic but \( 2x^3 - 8x^2 + \frac{1}{2}x - 9 \) is not.

Let \( p_q(x) \) be the monic polynomial whose roots are exactly the primitive \( q \)th roots of 1. For example, \( p_1(x) = x - 1 \), since 1 is the only primitive 1st root of 1. Likewise \( p_5(x) = x^4 + x^3 + x^2 + x^1 + 1 \), since \( x^5 - 1 \) is the monic polynomial whose roots are all the 5th roots of 1, but we must divide by \( x - 1 \) to account for the fact that 1 is not a primitive 5th root of 1. That gives \( p_5(x) = \frac{x^5 - 1}{x - 1} = x^4 + x^3 + x^2 + x^1 + 1 \), by long division.

As a final example, let’s see why \( p_4(x) = x^2 + 1 \). The four 4th roots of 1 are 1, \( i \), \(-1\), \(-i\), but only \( i \) and \(-i\) are primitive 4th roots of 1, so \( p_4(x) = (x - i)(x + i) = x^2 + 1 \). Alternatively, we get the primitive 4th roots of 1 starting with all the 4th roots and kicking out the nonprimitive ones, so \( p_4(x) = \frac{x^4 - 1}{(x - 1)(x + 1)} = x^2 + 1 \).

(a) Find \( p_8(x) \) and \( p_{16}(x) \).
(b) Find \( p_3(x) \), \( p_9(x) \), and \( p_{27}(x) \).
(c) Find \( p_6(x) \), \( p_{12}(x) \), and \( p_{18}(x) \).
(d) Find \( p_{30}(x) \) and \( p_{60}(x) \).

A computer algebra system can help with the long division.

3. Rational Exponents: Roots and Powers of the Base

In this section our goal is to define \( z^w \) where \( z \) is any (nonzero) complex number and \( w \) is any real number. We saw in Chapter 9, Section 3.2 that when the base \( z \) is positive (like 2, 10, or \( e \)), everything works beautifully: the expected properties of exponentiation hold. We will see however that the situation is trickier if the base is negative or non-real.

Before we begin, here is a reminder of the three common properties of exponents that hold for positive bases and real exponents. These were discussed extensively in Chapter 9:
3. RATIONAL EXPONENTS: ROOTS AND POWERS OF THE BASE

Property 1. \(a^{w_1}a^{w_2} = a^{w_1+w_2}\)

Property 2. \((a^{w_1})^{w_2} = a^{w_1w_2}\)

Property 3. \(a^w b^w = (ab)^w\)

3.1. Problems arise: negative and complex bases. Let \(z\) be a complex number. Write \(z\) in polar form as \(z = r(\cos \theta + i \sin \theta)\), where \(r > 0\) and \(-\pi < \theta \leq \pi\). Given any rational number \(w\), we define \(z^w\) by giving its polar form:

\[
z^w = r^w (\cos(w\theta) + i \sin(w\theta)).
\]

Equation (1) makes perfectly good sense, since previously we’ve defined the expression \(r^w\), and likewise the expression \(w\theta\) is defined (it’s the product of a rational number and a real number), so it’s a real number, hence we may apply the cosine and sine functions to it.

Your Turn 5. Use Equation (1) to compute \((-8 + 8i)^{5/3}\).

As always, our definition of exponentiation is motivated by the goal of making some or all of the properties of exponentiation come true. We’ll check that Property 1 is valid:

**PROPOSITION 7.** If \(z\) is a nonzero complex number and if \(w_1\) and \(w_2\) are rational numbers, then \(z^{w_1} z^{w_2} = z^{w_1 + w_2}\).

**Proof.** Writing \(z = r(\cos \theta + i \sin \theta)\), we have

\[
z^{w_1} z^{w_2} = r^{w_1} (\cos(w_1\theta) + i \sin(w_1\theta)) r^{w_2} (\cos(w_2\theta) + i \sin(w_2\theta)).
\]

The product of \(r^{w_1}\) and \(r^{w_2}\) is \(r^{w_1+w_2}\), by Theorem 5 from Chapter 9, Section 3.2. The product of \((\cos(w_1\theta) + i \sin(w_1\theta))\) and \((\cos(w_2\theta) + i \sin(w_2\theta))\) is \(\cos(w_1\theta + w_2\theta) + i \sin(w_1\theta + w_2\theta) = \cos((w_1 + w_2)\theta) + i \sin((w_1 + w_2)\theta)\), via trig identities. Hence \(z^{w_1} z^{w_2} = r^{w_1+w_2} (\cos((w_1 + w_2)\theta) + i \sin((w_1 + w_2)\theta)) = z^{w_1+w_2}\).

\[\square\]

Unfortunately, the good news stops there: Properties 2 and 3 both can fail for complex (or negative real) bases. We can show that Property 2 fails by comparing \((-1)^{3/2}\) and \((-1)^{3 \cdot 1/2}\). We can compute \((-1)^{3/2} = (-1)^{3 \cdot 1/2} = (-1)^{3/2} = i\), whereas by (1)

\[
(-1)^{3 \cdot 1/2} = (-1)^{3/2} = 1^{3/2} (\cos(3\pi/2) + i \sin(3\pi/2)) = -i.
\]
Thus Property 2 does not hold in general, if we allow complex numbers as the base and rational numbers as exponents.

**Your Turn 6.** A similar failure occurs for Property 3. Check this by comparing \((-1)^{\frac{1}{2}}(-1)^{\frac{1}{2}}\) and \(((−1)(−1))^\frac{1}{2}\).

### 3.2. Exercises.

1. On the basis of this section, explain why in high school, we never try to graph \((-2)^x\) even though we do graph \(2^x\).

2. Given a complex number \(z\), is it always true that the principal 14th root of \(z\) equals the principal 28th root of \(z^2\)? Justify your answer.

3. Find a sequence of points \(z_1, z_2, \ldots\) in \(\mathbb{C}\) that converges to a number \(z\), but such that the sequence \(z_1^{1/2}, z_2^{1/2}, \ldots\) does not converge to \(z^{1/2}\). (Hint: try picking \(z\) where the choice of angle \(\theta\) has a “break”.)

### 4. The Complex Exponential Function \(e^z\)

In this section we discuss the function \(e^z\), where \(z\) is a complex number. Its algebraic and geometric properties are beautiful, and the function has far-reaching generalizations in higher mathematics. The exponential function is very closely related to the sine and cosine functions (and to their hyperbolic cousins as well). To discover this connection, one only needs to add complex numbers into the mix. We will see now exactly how this happens.

#### 4.1. Definition of \(e^z\).

Recall the goal stated in the introduction to Chapter 9—we would like to be able to define \(z^w\) for any complex numbers \(z\) and \(w\). In this section, we make the important step of defining \(e^z\) for any complex number \(z\). But how should we make this definition?

At this point, trying to define \(e^z\) only using properties of exponentiation doesn’t work. Instead, our strategy is to define \(e^z\) using the same power series that works when \(z\) is a real number:

**Definition 8.** If \(z\) is any complex number, we define \(e^z = \sum_{j=0}^{\infty} \frac{z^j}{j!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \frac{z^5}{5!} + \ldots\)
4. THE COMPLEX EXPONENTIAL FUNCTION $e^z$

Your Turn 7. Use a truncated power series for $e^z$ (through the $z^5$ term) to give an approximate value for $e^i$. Compare it to the value for $e^i$ that your calculator gives.

Now $e^z$ is given by an infinite sum (of complex numbers!), so for $e^z$ to exist, we need to be assured that the series actually converges. This is quite easy to accomplish, using the fact that the series converges when the variable is a real number. To show that the series converges when $z$ is any complex number, we need to show that

$$
\lim_{N \to \infty} \sum_{j=N}^{\infty} \frac{z^j}{j!} = 0
$$

(why?)

However, we know that

$$
\left| \sum_{j=N}^{\infty} \frac{z^j}{j!} \right| \leq \sum_{j=N}^{\infty} \left| \frac{z^j}{j!} \right|
$$

and the latter expression approaches 0 as $N$ approaches infinity, since the series $\sum_{j=0}^{\infty} \frac{|z|^j}{j!}$ converges (note that $|z|$ is a real number).

4.2. Additive property of exponents for $e^z$. Here we discuss the additive property of exponentiation for $e^z$:

**Theorem 9.** $e^z \cdot e^w = e^{z+w}$ for all complex numbers $z$ and $w$.

Let’s do some preliminary calculations, based on truncating the power series. We have that

$$
e^z = 1 + z + \frac{1}{2} z^2 + \text{higher order terms}
$$

$$
e^w = 1 + w + \frac{1}{2} w^2 + \text{higher order terms}
$$

so $e^z \cdot e^w = (1 + z + \frac{1}{2} z^2 + \text{higher order terms}) \cdot (1 + w + \frac{1}{2} w^2 + \text{higher order terms})$. If we multiply this out, we find that $e^z \cdot e^w = 1 + (z + w) + \frac{1}{2} z^2 + zw + \frac{1}{2} w^2 + \text{higher order terms}$. On the other hand, $e^{z+w} = 1 + \frac{(z+w)^1}{1!} + \frac{(z+w)^2}{2!} + \text{higher order terms} = 1 + (z + w) + \left( \frac{1}{2} z^2 + zw + \frac{1}{2} w^2 \right)$. Thus the expressions for $e^z e^w$ and $e^{z+w}$ “agree with each other through order-two terms.”

Your Turn 8. Show that $e^z e^w$ and $e^{z+w}$ agree with each other through order-three terms.
A compelling argument in favor of Theorem 9 can be made by showing that the degree \( n \) terms of \( e^z e^w \) and \( e^{z+w} \) are the same. In \( e^z e^w \), we obtain a degree \( n \) term as the product of the degree \( k \) term in \( e^z \) (which is \( \frac{z^k}{k!} \)) with the degree \( n-k \) term in \( e^w \) (which is \( \frac{w^{n-k}}{(n-k)!} \)). Thus the degree \( n \) part of \( e^z e^w \) is \( \sum_{k=0}^n \frac{z^k w^{n-k}}{k!(n-k)!} \).

On the other hand, the degree \( n \) part of \( e^{z+w} \) is \( \left( \frac{z+w}{n} \right)^n \), which by the binomial theorem is

\[
\sum_{k=0}^n \binom{n}{k} \frac{z^k w^{n-k}}{n!} = \sum_{k=0}^n \frac{n!}{k!(n-k)!} \frac{z^k w^{n-k}}{n!} = \sum_{k=0}^n \frac{z^k w^{n-k}}{k!(n-k)!}.
\]

4.3. Euler’s Formula: writing \( e^z \) in terms of familiar functions. In this section, we explore a truly remarkable formula, due to Euler:

**Theorem 10.** For any complex number \( z = x + iy \) (with \( x, y \in \mathbb{R} \)), we have \( e^z = e^x (\cos y + i \sin y) \). In particular, taking \( x = 0 \), we have \( e^{iy} = \cos y + i \sin y \).

What does this accomplish? Note that \( e^x \) is a positive real number, and \( \cos y + i \sin y \) is a point on the unit circle in \( \mathbb{C} \). Thus, one interpretation of Euler’s Formula is that it gives \( e^x \) in polar form. Note also that \( e^x \), \( \cos y \), and \( \sin y \) are all familiar real-valued functions that are understandable by students in a precalculus class.

The proof of Euler’s Formula is quite easy. As a special case of Theorem 9, we have that \( e^x e^{iy} = e^{x+iy} = e^z \). Thus Euler’s Formula is proved after we show that \( e^{iy} = \cos y + i \sin y \). To do this, we use the definition of \( e^{iy} \), namely, \( e^{iy} = \sum_{j=0}^{\infty} \frac{(iy)^j}{j!} = \sum_{j=0}^{\infty} \frac{i^j y^j}{j!} \). We now separate the nonnegative integers \( j \) into even and odd integers; the even nonnegative integers can be written as \( 2k \) (for \( k \in \{0, 1, 2, 3 \ldots \} \)) and the odd nonnegative can be written as \( 2l+1 \) (for \( l \in \{0, 1, 2, 3 \ldots \} \)). Thus \( e^{iy} = \sum_{k=0}^{\infty} \frac{i^{2k} y^{2k}}{(2k)!} + \sum_{l=0}^{\infty} \frac{i^{2l+1} y^{2l+1}}{(2l+1)!} \). We wrap things up by examining the powers of \( i \) that occur. Note that \( i^{2k} = (i^2)^k = (-1)^k \), and likewise \( i^{2l+1} = i^{2l+1} = (-1)^l i^l \).

Finally, we have

\[
e^{iy} = \sum_{k=0}^{\infty} \frac{(-1)^k y^{2k}}{(2k)!} + \sum_{l=0}^{\infty} \frac{i(-1)^{2l+1} y^{2l+1}}{(2l+1)!} = \cos y + i \sin y,
\]

using the power series expansions of \( \cos y \) and \( \sin y \).

A final note: just as we extended the usual exponential function to a function from \( \mathbb{C} \) to \( \mathbb{C} \), we may define the following functions from \( \mathbb{C} \) to \( \mathbb{C} \) using power series:

\[\text{\footnote{A complete proof requires attention to convergence, but we will skip this discussion.}}\]
4. THE COMPLEX EXPONENTIAL FUNCTION $e^z$

$$e^z = \sum_{j=0}^{\infty} \frac{z^j}{j!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \frac{z^5}{5!} + \ldots$$

$$\cosh z = \sum_{k=0}^{\infty} \frac{z^{2k}}{(2k)!} = 1 + \frac{z^2}{2!} + \frac{z^4}{4!} + \frac{z^6}{6!} + \ldots$$

$$\sinh z = \sum_{k=0}^{\infty} \frac{z^{2k+1}}{(2k+1)!} = z + \frac{z^3}{3!} + \frac{z^5}{5!} + \frac{z^7}{7!} + \ldots$$

$$\cos z = \sum_{k=0}^{\infty} (-1)^k \frac{z^{2k}}{(2k)!} = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \ldots$$

$$\sin z = \sum_{k=0}^{\infty} (-1)^k \frac{z^{2k+1}}{(2k+1)!} = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \ldots$$

Of course, when $z$ is a real number, these definitions coincide with the usual definitions of the hyperbolic cosine, hyperbolic sine, cosine, and sine functions. These functions will be explored in the exercises.

4.4. The complex logarithm. We now want to define the complex logarithm to be the inverse function of the complex exponential map. However, we’ll show that the complex exponential map is neither one-to-one nor onto, which means that some judicious choices must be made in the definition of the complex logarithm.

**Theorem 11.**

(a) The range of the complex exponential map is $\mathbb{C} \setminus \{0\}$.

(b) The complex exponential map is not one-to-one. In particular, given two complex numbers $z$ and $w$, then $e^z = e^w$ if and only if $z - w$ is an integral multiple of $2\pi i$.

**Proof.** Why is 0 the only number missing from the range? Let $z = x + iy$. We have shown that $e^z = e^x (\cos y + i \sin y)$. We may write any point on the unit circle as $(\cos y + i \sin y)$ for some real number $y$, and we may write any positive real number $r$ as $e^x$ for some $x$ (take $x = \ln r$). Thus the range of the complex exponential map consists of all points in the complex plane that can be written as a positive real number times a point on the unit circle. All points in $\mathbb{C}$ except 0 can be written in this way.

The proof of (b) is a straightforward calculation: write $z = x + iy$ and $w = a + bi$. We have that $e^z = e^x (\cos y + i \sin y)$ and $e^w = e^a (\cos b + i \sin b)$. Since this gives the polar decompositions of the complex
numbers $e^z$ and $e^w$, we see that if $e^z = e^w$, then $x = a$ and $y - b$ is a multiple of $2\pi$. This amounts to saying that $z - w$ is a multiple of $2\pi i$, since $z - w = (x - a) + (y - b)i$. \qed

This discussion gives us guidance about how to define the complex logarithm. First, we now know that the domain should be $\mathbb{C} \setminus \{0\}$. Secondly, we know that the range cannot include two complex numbers that differ by a multiple of $2\pi i$. Note that $S = \{a + bi : -\pi < b \leq \pi\}$ has this property. This inspires the following definition of the complex logarithm:

**Definition 12.** Let $w$ be a nonzero complex number. Write $w$ (uniquely) as $w = a(\cos b + i \sin b)$, where $a > 0$ and $-\pi < b \leq \pi$. We define $\log w = \ln a + bi$.

From the definition, the domain of the complex logarithm function is $\mathbb{C} \setminus \{0\}$, and the range is the strip $S$ in the complex plane, as defined above.

Needless to say, the complex logarithm function is defined to make the following proposition true:

**Theorem 13.** For any nonzero complex number $w$, we have $e^{\log w} = w$.

**Proof.** Writing $w$ as $a(\cos b + i \sin b)$ (where $a + bi \in S$), we have $\log w = \ln a + bi$, and then $e^{\log w} = e^{\ln a}(\cos b + i \sin b) = a(\cos b + i \sin b) = w$. \qed

4.5. Geometry of the complex exponential map. Let’s try to describe the complex exponential map geometrically. Perhaps the easiest way to do this is by finding the image under exponentiation of horizontal and vertical lines in $\mathbb{C}$. For example, if we exponentiate points along the vertical line $x = 2$, we find

$$e^z = e^{2+iy} = e^2(\cos y + i \sin y)$$

and thus we obtain points on the circle centered at the origin of radius $e^2$. Likewise, if we exponentiate points along the horizontal line $y = \pi/3$, we compute that

$$e^z = e^{x+i\pi/3} = e^x(\cos(\pi/3) + i \sin(\pi/3))$$

which consists of the ray starting at (but not including) the origin and going through the point on the unit circle at angle $\pi/3$. Similarly, the image of any horizontal line is a circle centered at the origin, and the image of any vertical line is a ray emanating from the origin.
4.6. Exercises.

1. Use a truncated power series for \(e^z\) (through the \(z^5\) term) to give approximate values for \(e^{\pi i/6}\) and for \(e^{1+i}\). Compare your answers to the values for \(e^{\pi i/6}\) and for \(e^{1+i}\) that your calculator gives.

2. Using Theorem 10, show that \(e^{\pi i} = -1\). (This equations combines four of the most important numbers.)

3. Using Theorem 10, compute each of the following: \(e^{3+i\pi/4}\), \(e^{3\pi i/2}\), and \(e^{\ln 5+2\pi i/3}\).

4. Find all complex numbers \(z\) such that
   
   (a) \(e^z = e^5(\cos(\pi/3) + i\sin(\pi/3))\)

   (b) \(e^z = 7(\cos(5\pi/4) + i\sin(5\pi/4))\)

   (c) \(e^z = 1 + i\)

5. Using the power-series definitions, show that \(\cosh z\) and \(\cos z\) are even functions and \(\sinh z\) and \(\sin z\) are odd functions.

6. Use Definition 12 to compute each of the following, writing the answer in the form \(a + bi\). (The latter parts of the exercise are more difficult since one must find the polar form of the argument.)

   (a) \(\log(e^{7.2}(\cos 4 + i\sin 4))\)

   (b) \(\log(3(\cos(5\pi/4) + i\sin(5\pi/4)))\)

   (c) \(\log(5 - 5i)\)

   (d) \(\log(-\sqrt{3} - i)\)

   (e) \(\log(-1)\)

   (f) \(\log(-e)\)

   (g) \(\log(-5)\)

   (h) \(\log(4 + 7i)\)

7. Plot the image (under the exponential map) of each horizontal and vertical line and the diagonal line illustrated in Figure 5.
8. In Figure 6, plot $e^z$ for each shaded point $z$. (Use Euler's Formula.)

9. In Figure 7, plot $\log w$ for each shaded point $w$.

10. Give a simple description, in geometric terms, of how to obtain the point $\log(e^z) \in \mathbb{C}$ from the point $z \in \mathbb{C}$. 
11. Using the power-series definitions, show that

(a) $\cosh z = \cos(iz)$ and $\cos z = \cosh(iz)$
(b) $\sinh z = -i \sin(iz)$ and $\sin z = i \sinh(iz)$
(c) $e^z = \cosh z + \sinh z$ and $e^z = \cos(iz) - i \sin(iz)$

12. Using the results of Exercise 11, show that $\cosh z = \frac{e^z + e^{-z}}{2}$, and $\sinh z = \frac{e^z - e^{-z}}{2}$. Then, combine this with Exercise 11 to write $\cos z$ and $\sin z$ in terms of exponentials.

13. Using the preceding exercises, compute $\cosh(i\pi)$, $\sinh(3 + i\pi/4)$, $\sin(i)$, and $\cosh(1 + i)$. Check your answers with a calculator or computer algebra system.

14. Show that in general, $\log(e^z) \neq z$. What must be true about $z$ in order for $\log(e^z) \neq z$ to hold?

15. Describe the image of the line $y = 2x$ under the complex exponential map.

16. We have shown that $e^z e^w = e^{z+w}$ for all $z, w \in \mathbb{C}$. According to Exercise 11 and Exercise 12, we may write $\cos z, \cosh z, \sin z$, and $\sinh z$ in terms of exponentials.
Combining these two facts, it’s possible to prove identities involving the trigonometric and hyperbolic trigonometric functions. Your job: show how to recover the identities for \( \cosh(z+w) \), \( \cos(z+w) \), \( \sinh(z+w) \), and \( \sin(z+w) \). Do the same for the identities \( \cos^2 z + \sin^2 z = 1 \) and \( \cosh^2 z - \sinh^2 z = 1 \).

17. Investigate and report on the hyperbolic cosine function \( \cosh : \mathbb{C} \to \mathbb{C} \). You should consider the following questions:

(a) What is the power series for \( \cosh z \)?
(b) How can \( \cosh z \) be expressed in terms of \( e^z \) and \( e^{-z} \)?
(c) How can \( \cosh z \) be written in terms of (real-valued) functions of \( x \) and \( y \), where \( z = x + iy \)?
(d) How can \( \cosh z \) be represented graphically?
(e) What is the range of the hyperbolic cosine function?
(f) For what values of \( z \) does one have \( \cosh z = 0 \), \( \cosh z = 1 \), or \( \cosh z = -1 \)?

18. In this exercise we revisit the function \( f(x, y) = (\sin x \cosh y, \cos x \sinh y) \) that was discussed in Let’s Go 3.

(a) Recall that \( e^z \) and \( \sin z \) have power series representations valid for all complex numbers \( z \):

\[
e^z = 1 + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots \quad \text{and} \quad \sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots .
\]

Show that \( \sin z = \frac{e^{iz} - e^{-iz}}{2i} \). (Hint: Replace \( z \) by \( iz \) and \( -iz \) in the power series representation for the exponential function.)

(b) Let \( z = x + iy \) with \( x, y \in \mathbb{R} \) and recall that \( e^z = e^x(\cos y + i \sin y) \). Use this fact together with the result of part (a) to show that \( \sin(x + iy) = \sin x \cosh y + i \cos x \sinh y \). (It is helpful to recall that \( \cosh y = \frac{e^y + e^{-y}}{2} \) and \( \sinh y = \frac{e^y - e^{-y}}{2} \).)

5. Complex Bases and Complex Exponents

5.1. What is \( z^w \)? We end this section by considering the question we first posed in Chapter 9: How should we define \( z^w \), if \( z \) and \( w \) are complex numbers?

**Definition 14.** Let \( z \) and \( w \) be complex numbers, with \( z \neq 0 \). We define \( z^w = e^{w \log z} \).
Your Turn 9. Regarding Definition 14,

(a) Why have we required \( z \neq 0 \)?

(b) Explain how Property 2 of exponentiation (that is, \((a^b)^c = a^{bc}\)) could be used to motivate the definition.

On the basis of Your Turn 9, we should be convinced that Definition 14 is reasonable. However, we need to check for consistency with a previous definition. In Equation (1) in Section 3.1, we defined \( z^w \) if \( w \) is a rational number and \( z \) is a nonzero complex number. Say \( w \) is rational and suppose \( z = r(\cos \theta + i \sin \theta) \) with \(-\pi < \theta \leq \pi\). From Equation (1), we have that \( z^w = r^w(\cos(w\theta) + i \sin(w\theta)) \). On the other hand, according to Definition 14, we have \( z^w = e^{w \log z} = e^{w \ln r + i w \theta} = e^{w \ln r} e^{i w \theta} = r^w(\cos(w\theta) + i \sin(w\theta)) \).

Thus, we discover (with relief) that Definition 14 is consistent with Equation (1).

In Section 3.1, we saw that Properties 2 and 3 can fail if the base is not a positive real number. On the other hand, we should be gratified that the additive property of exponents holds even in our most general setting:

**Proposition 15.** Let \( z \) be a nonzero complex number, and let \( w_1 \) and \( w_2 \) be complex numbers. We have that \( z^{w_1} \cdot z^{w_2} = z^{w_1+w_2} \).

**Proof.** \( z^{w_1} \cdot z^{w_2} = e^{w_1 \log z} e^{w_2 \log z} = e^{w_1 \log z + w_2 \log z} = e^{(w_1+w_2) \log z} = z^{w_1+w_2} \). \qed

5.2. Exercises.

1. Use Definition 14 to compute \((1 + i)^{3+4i}\), \((i)^{3+4i}\), and \((-2)^{3+4i}\).

2. Revisit items (q)–(z) in Chapter 9, Let’s Go 1. Compute each, using the definitions in this chapter.

3. Consider \((1 + i)^z\).

   (a) Compute \((1 + i)^z\) (where \( z = x + iy \)).

   (b) For what values of \( z \) is \(|(1 + i)^z| = 1|?\)

   (c) For what values of \( z \) is \((1 + i)^z\) a positive real number?

   (d) For what values of \( z \) is \((1 + i)^z\) a pure imaginary number?
4. Consider $z^{3+4i}$.

(a) Compute $z^{3+4i}$ (where $z = r(\cos \theta + i \sin \theta)$).

(b) For what values of $z$ is $|z^{3+4i}| = 1$?

(c) For what values of $z$ is $z^{3+4i}$ a positive real number?

(d) For what values of $z$ is $z^{3+4i}$ a pure imaginary number?

5. Let $z = r(\cos \theta + i \sin \theta)$, with $r > 0$ and $-\pi < \theta \leq \pi$.

(a) Compute each of the following: $z^3$, $z^{-3}$, $z^{2i}$, $z^{3+2i}$, and $z^{-3+2i}$.

(b) Which of these approaches a limit as $r \to 0$?

6. Justify each equality in the proof of Proposition 15.

7. Is it always true that $\log(z^w) = w \log z$? Give a proof or a counterexample.
CHAPTER 12

Beyond Quadratics: Higher Degree Polynomials

We are not intimidated by any linear or quadratic polynomial. For instance, if someone demands that we solve $3x^2 - 16x + 24 = 0$, we are ready to do so, by completing the square or by using the quadratic formula. We are also prepared to factor $3x^2 - 16x + 24$, and by translating and stretching the graph of $y = x^2$, we can produce the graph of $y = 3x^2 - 16x + 24$. However, things are murkier for higher degree polynomials, and the following questions come into play:

- Can we easily graph higher degree polynomials, and is there any hope of classifying the various shapes manifested by polynomials of a given degree? We know that any graph of a quadratic polynomial is a translation of the graph of $y = ax^2$. Even though graphs of higher degree polynomials often have more ‘twists and turns’, could they satisfy a similar property?

- What methods are available for finding exact solutions of higher degree polynomial equations? For instance, is there some analog of the quadratic formula that can be used to solve the cubic equation $x^3 - 15x + 10 = 0$? To solve $x^7 + 4x^4 - x^2 + 2 = 0$?

- To what extent can a higher degree polynomial be factored? Thanks to the quadratic formula, we can factor any quadratic polynomial. But is it theoretically possible to factor, say, $x^8 + 2x^7 + x^4 - x + 29$ into a product of linear factors, or at least into a product of two polynomials of lower degree?

Many of these are not easy questions—some have occupied the minds of mathematicians for thousands of years. In this chapter we investigate these questions, largely in the context of cubic polynomials. We will see that methods that worked for quadratic polynomials (e.g., completing the square) are also useful for higher degree polynomials, but that the analogy between quadratics and higher degree polynomials is imperfect. Also, we will find that understanding connections between the coefficients of a polynomial and its roots is crucial for unlocking the secrets of higher degree polynomials.
The topics of this chapter are motivated by the questions posed above. First, we investigate the graphical features of cubic polynomials, leading naturally to a discussion of the relationship between the roots of a polynomial and its coefficients. We then discuss Cardano’s method for solving cubic polynomials, the possibility (or lack thereof) of extending such a method to all higher degree polynomials, and Euler’s attempt at proving the Fundamental Theorem of Algebra. We finish by investigating a numerical technique for approximating roots of functions (Newton’s method), and we find that our knowledge of polynomials can be used to improve the efficiency of the technique.

To ease our transition to higher degree polynomials, for most of this chapter we will assume the Fundamental Theorem of Algebra:

**Theorem 1.** (Fundamental Theorem of Algebra) If $n$ is a positive integer and $f(x)$ is a polynomial of degree $n$ with complex coefficients, then there is a unique complex number $c$ and a unique collection $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ of complex numbers (not necessarily distinct) satisfying

$$f(x) = c(x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n).$$

The Fundamental Theorem of Algebra is the ultimate factorization theorem for polynomials. It asserts that, at least theoretically, any nonconstant polynomial with complex coefficients can be factored over the complex numbers, thus providing a positive answer to our third question above. (Further discussion of the Fundamental Theorem of Algebra will occur in Section 4 below.)

1. **Graphical Features of Cubic Polynomials**

We enter the world of higher degree polynomials (particularly cubics) by examining their graphs. We will find that the coefficients of a cubic polynomial tell us a lot about its graph.

**Let’s Go 1.** Using a graphing utility, graph the cubic polynomials $f(x) = x^3$, $g(x) = x^3 + 10x + 1$, and $h(x) = -x^3 - 4x^2 - x + 3$. Based on these graphs, brainstorm ways in which the graphs of cubic polynomials are different from the graphs of quadratic polynomials. (To generate ideas, it may help you to focus on typical aspects of graphs, including intercepts, local extrema (i.e., local maximums and minimums), and concavity.)
1.1. **Shape investigation.** In Let’s Go 1 you may have noticed that graphs of cubic polynomials come in more ‘flavors’ than those of quadratic polynomials. While quadratic polynomials come in one basic shape, Figure 1 indicates that there are graphs of cubic polynomials which look drastically different from each other.¹

![Figure 1](image-url) The graphs of \( y = x^3 + x + 1 \) and \( y = x^3 - x + 1 \) appear quite different.

It is natural to wonder to what extent graphs of cubic polynomials can vary in shape. We begin by investigating the shape of a cubic polynomial \( f(x) = x^3 + bx^2 + cx + d \), called a *monic* polynomial because the leading coefficient (the number in front of the highest power of \( x \)) is one. Observe that

\[
\begin{align*}
    f'(x) &= 3x^2 + 2bx + c  \\
    f''(x) &= 6x + 2b.
\end{align*}
\]

Since \( f''(x) \) changes from negative to positive at \( x = -b/3 \) (the only zero of \( f''(x) \)), we learn that every monic cubic polynomial has exactly one change in concavity (from negative to positive, at \( x = -b/3 \)). Meanwhile, any local extrema for \( f(x) \) must occur at zeros of \( f'(x) \), that is, at solutions of \( 3x^2 + 2bx + c = 0 \). This is a quadratic equation with solutions

\[
x = \frac{-b \pm \sqrt{b^2 - 3c}}{3}.
\]

So, using Equation (2) together with either the First or Second Derivatives Tests, we see that the nature of the extrema for \( f(x) \) is determined by the sign of \( b^2 - 3c \):

<table>
<thead>
<tr>
<th>Sign</th>
<th>Extrema for ( f(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b^2 - 3c &gt; 0 )</td>
<td>two local extrema (a max and a min)</td>
</tr>
<tr>
<td>( b^2 - 3c = 0 )</td>
<td>no local extrema and one horizontal tangent at ( x = -b/3 )</td>
</tr>
<tr>
<td>( b^2 - 3c &lt; 0 )</td>
<td>no local extrema and no horizontal tangents</td>
</tr>
</tbody>
</table>

¹Unlike graphs of quadratic polynomials, the graphs in Figure 1 cannot be obtained from each other by means of simple graphing transformations (translation, reflection, and stretching).
Using the table above together with the facts that \( \lim_{x \to \infty} f(x) = \infty \) and \( \lim_{x \to -\infty} f(x) = -\infty \), we conclude that there are three possible ‘shapes’ for the graph of \( f(x) = x^3 + bx^2 + cx + d \), depending on the sign of \( b^2 - 3c \). These are shown in Figure 2 below.

**Figure 2.** Shapes for monic cubic polynomials

**Your Turn 1.**

(a) Use the First or Second Derivative Test to show how the sign of \( b^2 - 3c \) translates to the information given in the table above.

(b) In determining the shape of the graph of \( f(x) = x^3 + bx^2 + cx + d \) from Equations (1) and (2), why is it convenient that \( f' \) is a quadratic polynomial, as opposed to, say, a polynomial of eleventh degree?

**Your Turn 2.** When asked to draw the graph of a cubic polynomial, our first impulse may be to graph \( y = x^3 \). Based on the classification of shapes given above, is the shape of the graph of \( y = x^3 \) typical or atypical of that for ‘most’ cubic polynomials? Why?

**1.2. Interlude: Depressed Cubics.** Polynomials would be much easier to deal with if they didn’t have all of those troublesome middle terms. For example, finding the shape and intercepts for the graph of \( f(x) = x^3 + 5 \) (just a vertical translation of \( y = x^3 \)) is much easier than doing the same for \( g(x) = x^3 + 6x^2 - 2x - 15 \). So it seems reasonable that to understand a cubic like \( g(x) \) fully, we might first try to
remove all, or at least some, of its middle terms sensibly. For example, let \( h(x) = g(x - 2) \). Notice that

\[
h(x) = (x - 2)^3 + 6(x - 2)^2 - 2(x - 2) - 15
\]

\[
= (x^3 - 6x^2 + 12x - 8) + 6(x^2 - 4x + 4) - 2(x - 2) - 15 = x^3 - 14x + 5.
\]

Through a simple substitution \((x - 2)\) in place of \(x\), we produced a cubic polynomial \( h(x) \) that has no \(x^2\)-term. Cubics like \( h(x) \) that have the form \( x^3 + px + q \) are called *depressed cubics*.

In general, the depressed form of a monic cubic polynomial \( x^3 + bx^2 + cx + d \) is obtained by replacing \( x \) by \( x - \frac{b}{3} \) (see Exercise 4 below).

In addition to the fact that depressed cubics lack an \( x^2\)-term, there are a number of important geometric observations about depressed cubics. These are explored in exercises and/or later sections of this chapter:

- The graph of a monic cubic and its depressed form are horizontal translates of each other. Therefore, the graphical features of the depressed form completely determine the graphical features of the original cubic, and vice versa (see Your Turn 3).
- The inflection point on the graph of a depressed cubic lies on the \(y\)-axis (see Your Turn 3 and Exercise 8).
- If a depressed cubic has three \(x\)-intercepts, then the sum of those intercepts is zero (see Section 2.4).

**Your Turn 3.** Let \( g(x) = x^3 + 6x^2 - 2x - 15 \) and \( h(x) = x^3 - 14x + 5 \) be as above, and recall that \( h(x) = g(x - 2) \).

(a) Using a graphing device, graph both \( g(x) \) and \( h(x) \). In terms of graphing transformations, how does one obtain the graph of \( h(x) \) from the graph of \( g(x) \)?

(b) Given the \(x\)-intercepts of \( g(x) \), how could you find the \(x\)-intercepts for the graph of \( h(x) \)?

(c) What is the inflection point for the graph of \( h(x) \)?

(d) Estimate the sum of the \(x\)-intercepts for the graph of \( h(x) \).

---

2Ideally we would like to remove all the middle terms from a cubic using a ‘simple’ substitution, but this is not an easy matter. In the 1680’s, Ehrenfried Walther von Tschirnhaus (1651–1708) showed that all the middle terms of a cubic polynomial can be removed by putting \( x^2 + b_1 x + b_2 \) in place of \( x \) for a suitable choice of \( b_1, b_2 \). A similar method can be used to remove all the middle terms of a quartic polynomial, but the method fails in general for polynomials of degree 5 and higher.
Your Turn 4. How is the process and outcome of producing a depressed cubic from a given cubic like completing the square for quadratic polynomials?

1.3. Intercepts Investigation. In this section we see what the coefficients of a cubic polynomial tell us about the number of its $x$-intercepts. Since Part (b) of Your Turn 3 indicates that any monic cubic polynomial has the same number of $x$-intercepts as its depressed form, we simplify the situation by only considering depressed cubics $f(x) = x^3 + px + q$.

Earlier we discovered that graphs of cubic polynomials $x^3 + bx^2 + cx + d$ come in three basic shapes depending on whether $b^2 - 3c$ is positive, negative, or zero. In the case of our depressed cubic, $b^3 - 3c$ becomes $-3p$, so the shape of the corresponding graph depends on the sign of $p$, as shown in Figure 3. Whenever $p \geq 0$, it is evident from Figure 3 that $f(x) = x^3 + px + q$ will have exactly one $x$-intercept. On the other hand, when $p < 0$ the graph of $f(x)$ has two local extrema, and a bit more work is required to determine the number of $x$-intercepts.

Your Turn 5. Show that a cubic polynomial whose graph looks like the right-hand graph from Figure 3 may have one, two, or three $x$-intercepts. (Vary the position of an $x$-axis superimposed on the graph.)

Your Turn 6. On Geometer’s Sketchpad, set parameters $p$ and $q$ and graph $f(x) = x^3 + px + q$.

---

3Here, we count intercepts without ‘multiplicity.’ Counting roots with ‘multiplicity’ means that each root $r$ is repeated in the list of roots according to the number of times $x - r$ occurs as a factor of the polynomial. For example, the set of positive roots of $(x + 6)(x + 2)^2(x - 0.5)^3(x - 5)(x - 5)^4$ listed with multiplicity is \{0.5, 0.5, 0.5, 2, 5, 5, 5, 5\}, making a total of 8 positive roots counted with multiplicity.
(a) Observe the behavior of the graph of $f(x)$ as $p$ changes from positive to negative.

(b) By varying the values of $p$ and $q$, illustrate that the graph of $f(x)$ can have one, two, or three $x$-intercepts.

(c) The graph of $f(x)$ has three $x$-intercepts when $q = 0$ and $p < 0$. Why?

In Your Turn 6 we saw that $f(x)$ has three $x$-intercepts when $p < 0$ and $q = 0$, but what about for other values of $q$? Again, calculus comes to the rescue. To begin, observe that $f'(x) = 3x^2 + p$, with sign chart given in Figure 4. (We recall that the sign chart summarizes specific information about where $f(x)$ is increasing or decreasing, as well as about the nature of the critical points $\pm \sqrt{-\frac{p}{3}}$.)

![Figure 4. Sign chart for $f'(x) = 3x^2 + p$, when $p < 0$.](image)

Now, let’s assume that both $p$ and $q$ are negative in $f(x) = x^3 + px + q$. Since $q = f(0)$, and $f(x)$ decreases from $x = 0$ to $x = \sqrt{-\frac{p}{3}}$, we know $f(\sqrt{-\frac{p}{3}}) < 0$. However, since $f(x)$ increases to infinity on $(\sqrt{-\frac{p}{3}}, \infty)$, the Intermediate Value Theorem tells us that $f(x)$ has precisely one $x$-intercept on the interval $[0, \infty)$ (see Figure 5). Meanwhile, on the negative real numbers $f(x)$ achieves its maximum value at $x = -\sqrt{-\frac{p}{3}}$. Taking this into consideration along with the facts that $f(0) < 0$ and $\lim_{x \to -\infty} f(x) = -\infty$, we see that the graph

![Figure 5. The $x$-intercepts of $y = x^3 + px + q$ for $x \geq 0$. (Both $p$ and $q$ are negative.)](image)
of \( f(x) \) has either zero, one, or two \( x \)-intercepts on \((−\infty,0)\) depending on whether \( f(−\sqrt{−\frac{p}{3}}) \) is negative, zero, or positive, respectively (see Figure 6). All together, we conclude that \( f(x) \) has either one, two, or three \( x \)-intercepts depending on the sign of \( f(−\sqrt{−\frac{p}{3}}) \). Finally, to express the sign of \( f(−\sqrt{−\frac{p}{3}}) \) in terms of \( p \) and \( q \), we observe

\[
f \left( −\sqrt{−\frac{p}{3}} \right) = \frac{2(−p)^{\frac{3}{2}}}{3^2} + q,
\]

which (upon manipulation and squaring) implies that the sign of \( f(−\sqrt{−\frac{p}{3}}) \) is the opposite of that for \( \frac{4p^3}{27} + q^2 \).

For example,

\[
f \left( −\sqrt{−\frac{p}{3}} \right) < 0 \iff \frac{2(−p)^{\frac{3}{2}}}{3^2} < −q \iff \frac{4}{27}(−p)^3 < q^2 \iff \frac{4p^3}{27} + q^2 > 0.
\]

Combining all of this with a similar computation for \( p < 0 \) and \( q > 0 \) (see Exercise 13), we have:

<table>
<thead>
<tr>
<th>Conditions on ( p, q )</th>
<th>Number of ( x )-intercepts for ( y = x^3 + px + q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p \geq 0 )</td>
<td>1</td>
</tr>
<tr>
<td>( p &lt; 0 ) and ( \frac{4p^3}{27} + q^2 &gt; 0 )</td>
<td>1</td>
</tr>
<tr>
<td>( p &lt; 0 ) and ( \frac{4p^3}{27} + q^2 = 0 )</td>
<td>2</td>
</tr>
<tr>
<td>( p &lt; 0 ) and ( \frac{4p^3}{27} + q^2 &lt; 0 )</td>
<td>3</td>
</tr>
</tbody>
</table>

Your Turn 7. Every depressed cubic \( y = x^3 + px + q \) can be associated to a single point in the \( pq \)-plane, and vice versa.

(a) Using the table above, draw a picture indicating how the \( pq \)-plane may be divided into regions according to the number of \( x \)-intercepts possessed by the corresponding cubics. (Hint: First plot the curve \( \frac{4p^3}{27} + q^2 = 0 \) for \( p < 0 \).)
(b) Suppose we select a depressed cubic at random with the restriction that \( p^2 + q^2 \leq 1 \). What is the most likely number of \( x \)-intercepts for this cubic? (Your solution to part (a) should be helpful.)

1.4. Intercepts, Real Roots, and Multiplicity. The \( x \)-intercepts are arguably the most important features of a polynomial graph since they (partially) enable us to factor the polynomial, as shown in the Fundamental Theorem of Algebra (Theorem 1). In this section we establish a connection among \( x \)-intercepts, roots of polynomials, and the Fundamental Theorem of Algebra.

Consider the polynomial \( f(x) = x^5 + 5x^4 - 4x^3 - 28x^2 - 32x - 192 \). A graph of \( f(x) \) (see Figure 7) indicates two \( x \)-intercepts, including one at \( x = 3 \). (We can verify this by checking that \( f(3) = 0 \).) Meanwhile, the

\[
\text{Fundamental Theorem of Algebra (Theorem 1) tells us that there is a unique set of (not necessarily distinct) complex numbers } \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}, \text{ called the roots of } f(x), \text{ such that}
\]

\[
f(x) = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)(x - \alpha_4)(x - \alpha_5).
\]

Putting Equation (4) together with \( f(3) = 0 \) gives

\[
(3 - \alpha_1)(3 - \alpha_2)(3 - \alpha_3)(3 - \alpha_4)(3 - \alpha_5) = 0,
\]

and so at least one of \( \alpha_1, \ldots, \alpha_5 \) must be 3. However, it is not clear how many times 3 is repeated in the list \( \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\} \) of roots. The number of times a root is repeated in the list of roots is called its multiplicity. We explore this concept below.

**Your Turn 8.** Let \( f(x) \) be a nonconstant polynomial.\(^4\)

\(^4\)This Your Turn anticipates discussion of the Factor Theorem in Section 4.2.
(a) A root of \( f(x) \) is often defined to be a complex number \( \alpha \) satisfying \( f(\alpha) = 0 \). Discuss whether this definition is equivalent to the definition of root given above.

(b) Is every \( x \)-intercept of the graph of \( f(x) \) also a root of \( f(x) \)? Is every root an \( x \)-intercept?

(c) In teaching your precalculus class, you want to show your students an example of a cubic polynomial with roots 2.5, -1, and 7. Find such a cubic, and check your answer with a graphing device.

Your Turn 9. Let \( f(x) = x(x - 1)(x - 2)(x - 3), \ g(x) = x(x - 1)(x - 3)^2, \ h(x) = (x - 3)^3, \) and \( j(x) = (x - 3)^4. \)

(a) Graph the four functions given above with a graphing device.

(b) Conjecture a way to tell from the graph if a real root has multiplicity larger than one.

(c) Conjecture a way to tell when the multiplicity of a root is even, and a way to tell when the multiplicity is odd.

(d) Can the multiplicity of a root exceed the degree of the polynomial? Why?

Referring again to the graph of \( f(x) = x^5 + 5x^4 - 4x^3 - 28x^2 - 32x - 192 \) (Figure 7), we may guess that \( x = 3 \) is a root of multiplicity one while \( x = -4 \) has multiplicity larger than one (see Your Turn 9).

Your Turn 10. By long division, show that 3 is a root of \( x^5 + 5x^4 - 4x^3 - 28x^2 - 32x - 192 \) of multiplicity one, and \(-4\) is a root of multiplicity two. (By long division, you should be able to find the other roots as well.)

Interestingly, we can verify our statements about the multiplicities of these roots without actually factoring \( f(x) \). Let’s first consider \( x = 3 \). Since 3 is a root of \( f(x) \) (say \( \alpha_1 = 3 \) in Equation (4)), by the Fundamental Theorem of Algebra we may write

\[
f(x) = (x - 3)(x - \alpha_2)(x - \alpha_3)(x - \alpha_4)(x - \alpha_5),
\]

which can be shortened to \( f(x) = (x - 3)g(x) \) where

\[
g(x) = (x - \alpha_2)(x - \alpha_3)(x - \alpha_4)(x - \alpha_5).
\]

Taking a derivative gives \( f'(x) = g(x) + (x - 3)g'(x) \), which implies \( f'(3) = g(3) \). Now, either by looking at the graph of \( f(x) \) or by computing \( f'(3) \) directly, we can see that \( f'(3) \neq 0 \). But this implies \( g(3) \neq 0, \)

which, by looking at Equation (5), says that none of \( \alpha_2, \ldots, \alpha_5 \) can be equal to 3. We conclude that \( x = 3 \) is a root for \( f(x) \) of multiplicity one.

We can play the same game to show that \( x = -4 \) is a root for \( f(x) \) of multiplicity two. First write 

\[ f(x) = (x+4)g(x), \]

where \( g(x) = (x-\alpha_2) \cdots (x-\alpha_5) \). Take a derivative to obtain 

\[ f'(x) = (x+4)g'(x) + g(x), \]

which says \( f'(-4) = g(-4) \). Now, unlike the previous case, we observe from the graph of \( f(x) \) or otherwise that \( f'(-4) = 0 \). This says \( g(-4) = 0 \), so that at least one of the \( \alpha_2, \ldots, \alpha_5 \), say \( \alpha_2 \), is equal to \(-4 \). We may then write 

\[ f(x) = (x+4)^2 h(x), \]

where \( h(x) = (x - \alpha_3)(x - \alpha_4)(x - \alpha_5) \). Taking the second derivative gives 

\[ f''(x) = 2h(x) + 4(x+4)h'(x) + (x+4)^2 h(x), \]

so \( f''(-4) = 2h(-4) \). On the other hand, from the rule for \( f(x) \) we may compute \( f''(-4) = -280 \). Therefore \( h(-4) \neq 0 \), and so none of \( \alpha_3, \alpha_4, \alpha_5 \) are equal to \(-4 \). We conclude that \(-4 \) is a root for \( f(x) \) of multiplicity two.

As we began to suspect in Your Turn 9, a root of a polynomial seems to have multiplicity larger than one exactly when the root is also a root of the derivative of the polynomial. This idea is captured in the following theorem:

**Theorem 2.** Let \( f(x) \) be a polynomial of degree \( n \) and \( x_0 \) a root of \( f(x) \). If \( k \leq n \), then \( x_0 \) is a root of multiplicity \( k \) if and only if 

\[ f(x_0) = f'(x_0) = \cdots = f^{(k-1)}(x_0) = 0, \]

but \( f^{(k)}(x_0) \neq 0 \).

And now we return to cubic polynomials:

**Your Turn 11.** In Section 1.3 we counted the number of \( x \)-intercepts for a depressed cubic \( y = x^3 + px + q \); this information is summarized in Figure 8. Re-label Figure 8 so that instead of giving information about the number of \( x \)-intercepts, the figure gives information about the number of real roots of \( y = x^3 + px + q \), being sure to count multiplicity!

**1.5. Exercises.**

1. Transform each cubic polynomial to a depressed cubic polynomial. Be sure to indicate which substitution you are using.
Figure 8. The number of x-intercepts for \( y = x^3 + px + q \), expressed in the \( pq \)-plane.

(a) \( x^3 + 12x^2 - x + 5 \).

(b) \( x^3 - 10x^2 - 3x - 2 \).

(c) \( 2x^3 - 18x^2 + 7x + 1 \).

2. Consider the graph of the cubic polynomial \( f(x) = x^3 + px + q \) given in Figure 9. Using information from the graph, find approximate values for \( p \) and \( q \).

Figure 9. \( f(x) = x^3 + px + q \).

3. Consider the cubic polynomial \( f(x) = (x - 5)(x - 4)(x - 3) \). Find a quick way to depress \( f(x) \) without multiplying out its factors.

4. Why does replacing \( x \) by \( x - \frac{b}{3} \) remove the \( x^2 \)-term from \( x^3 + bx^2 + cx + d \)?

5. In each instance, find a specific example of a cubic polynomial \( p \) satisfying the given criteria:
1. Graphical features of cubic polynomials

(a) Positive $y$-intercept; one $x$-intercept; two local extrema.

(b) Negative $y$-intercept; three $x$-intercepts.

(c) Negative $y$-intercept; two $x$-intercepts; two local extrema.

(d) Positive $y$-intercept; two $x$-intercepts; $\lim_{x \to \infty} p(x) = -\infty$.

6. Give a convincing argument that $\lim_{x \to \infty} x^3 + bx^2 + cx + d = \infty$, and that $\lim_{x \to -\infty} x^3 + bx^2 + cx + d = -\infty$. (Factor out an $x^3$.)

7. Will the shape of a general cubic polynomial polynomial $g(x) = ax^3 + bx^2 + cx + d$ differ significantly from that of a monic cubic polynomial? Explain.

8. Let $f(x) = ax^3 + bx^2 + cx + d$ where $a \neq 0$.

(a) Show that the graph of $f(x)$ has exactly one change in concavity.

(b) Show that in the case that the graph has two local extrema, a change in concavity occurs at a point whose $x$-coordinate is halfway between the $x$-coordinates of the two extrema.

(c) Let $x_0$ be the $x$-coordinate of the inflection point (which you probably found in part (a)). Show that $g(x) = f(x + x_0)$ is a depressed cubic whose point of inflection is on the $y$-axis.

9. Find a specific example of a cubic polynomial with:

(a) a root of multiplicity two at $x = -1$, a root of multiplicity one at $x = 1$, and $y$-intercept 2.

(b) a local minimum at $x = 2$ and a local maximum at $x = 5$.

10. Suppose $f(x)$ is a polynomial of degree $n$. What is the largest number of roots (including multiplicity) that $f(x)$ can possess?

11. Let $f, g : \mathbb{R} \to \mathbb{R}$. The graph of $f(x)$ can be obtained from the graph of $g(x)$ by graphing transformations if and only if $f(x) = cg(ax + b) + d$ for some real constants $a, b, c, d$ (why?). Use this fact to:

(a) Show that $f(x) = x^3 + 2x + 1$ can be obtained from $g(x) = x^3 + x + 1$ by a graphing transformation.

Check your answer using a graphing device.
(b) Consider the graphs of \(x^3\), \(x^3 - x\), and \(x^3 + x\). Show that no two of these graphs can be obtained from one another by graphing transformations.

(c) Show that every cubic polynomial can be obtained from one and only one of \(x^3\), \(x^3 - x\), \(x^3 + x\) via graphing transformations.


13. In the case where \(p < 0\) and \(q > 0\), show that the number of \(x\)-intercepts for \(y = x^3 + px + q\) depends on the sign of \(4p^3 + 27q^2\).

14. Consider the depressed quartic polynomial \(Q(x) = x^4 + bx^2 + cx + d\) where \(b, c, d\) are real numbers. Find conditions on \(b\) and \(c\) indicating that the graph of \(Q(x)\) can take on one of four possible basic shapes. (Begin by looking at \(Q'(x)\) and \(Q''(x)\). Observe that \(Q'(x)\) is a depressed cubic, so your knowledge of roots of depressed cubics will come into play.)

2. Connection: Roots and Coefficients

Our investigation of cubic graphs revealed a connection between the coefficients of a polynomial and its roots. In this section, we continue to explore this important connection, for polynomials of any degree.

Polynomial equations are often hard to solve. For example, if you are handed the equation \(x^3 - x^2 - 2x + 20 = 18\) and told to ‘solve’ it, you might first rewrite this as \(x^3 - x^2 - 2x + 2 = 0\). Then, after scratching your chin, you note that the Fundamental Theorem of Algebra implies

\[
(6) \quad x^3 - x^2 - 2x + 2 = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)
\]

for some complex numbers \(\alpha_1, \alpha_2, \alpha_3\), and that these numbers are precisely the solutions you are looking for.\(^5\)

But what are the values of \(\alpha_1, \alpha_2, \alpha_3\)? Generally, this is a hard question, but we might begin by observing (via multiplication) that the roots \(\alpha_1, \alpha_2, \alpha_3\) depend on the coefficients of the polynomial. In light of this, we set about examining some familiar connections between roots and coefficients.

\(^5\)You might be thinking of the word ‘factoring’. This will be discussed in more detail in Section 4.
2.1. Conjugate Roots. In the case that a polynomial has real coefficients (as has been our default assumption throughout this chapter), you may recall that non-real complex roots must come in conjugate pairs. We investigate this below.

Your Turn 12. Let \( f(x) = (x - (5 + 3i))(x - (a + bi)) \), where \( a, b \in \mathbb{R} \).

(a) If \( f(x) \) is to have real coefficients, what are the possible values of \( a \) and \( b \)? Verify your answer.

(b) Can \( f(x) \) have an alternate factorization \( f(x) = (x - r_1)(x - r_2) \) where \( r_1, r_2 \) are real numbers? Why?

Your Turn 12 indicates that non-real complex roots should appear in conjugate pairs, but the method suggested in part (a) of Your Turn 12 is too unwieldy for polynomials of higher degree. Instead, we rely on basic properties of complex numbers, particularly the fact that taking the complex conjugate of a real number leaves that number unchanged, to prove the following theorem:

**Theorem 3.** If \( f(x) \) is a polynomial (with real coefficients) and \( \alpha \in \mathbb{C} \) is a root of \( f(x) \), then \( \bar{\alpha} \) is also a root of \( f(x) \).

**Proof.** Suppose that \( f(x) = a_0 + a_1x + \cdots + a_nx^n \) where \( a_0, a_1, \ldots, a_n \) are real numbers. Then

\[
f(\bar{\alpha}) = a_0 + a_1\bar{\alpha} + \cdots + a_n\bar{\alpha}^n = \bar{a_0} + \bar{a_1}\alpha + \cdots + \bar{a_n}\alpha^n = \bar{a_0} + \bar{a_1}\alpha + \cdots + \bar{a_n}\alpha^n = \bar{0} = 0.
\]

\[\Box\]

Your Turn 13. Consider the proof of Theorem 3.

(a) What properties of complex conjugates are essential in the proof?

(b) Where is the fact that \( \alpha \) is a root of \( f(x) \) used in the proof?

(c) Is the theorem still true if we do not require \( f(x) \) to have real coefficients? If not, give a counterexample.

Theorem 3, together with the Fundamental Theorem of Algebra (Theorem 1), has an interesting application to writing polynomials with real coefficients as a product of polynomials with real coefficients. Take
for instance the polynomial \( f(x) = x^5 + 5x^4 - 4x^3 - 28x^2 - 32x - 192 \), which we discussed in Section 1.4.

The Fundamental Theorem of Algebra says \( f(x) \) can be ‘completely factored’ and indeed

\[
f(x) = (x - 3)(x + 4)^2(x + 2i)(x - 2i).
\]

Observe that the factorization is consistent with Theorem 3: The non-real roots \( \pm 2i \) occur as a conjugate pair. The corresponding terms \( (x + 2i) \) and \( (x - 2i) \) multiply to give a quadratic \( x^2 + 4 \) with real coefficients, and we may re-write \( f(x) \) as

\[
(7) \quad f(x) = (x - 3)(x + 4)^2(x^2 + 4).
\]

Observe that all of the factors in Equation (7) now have real coefficients, and that the quadratic factor has no real roots (why?). This lends evidence to the following corollary (see Exercise 3), which is popular in high school mathematics curricula.

**Corollary 4.** Any polynomial with real coefficients may be written as a product of linear and quadratic polynomials with real coefficients, where the quadratic polynomials have no real roots.

**2.2. Rational Roots Theorem.** The court magister to the king of Sicily challenged Fibonacci\(^6\) (1180-1240), originator of the famous Fibonacci sequence 1, 1, 2, 3, 5, 8, ..., to solve the cubic equation

\[
(8) \quad x^3 + 2x^2 + 10x - 20 = 0.
\]

While he could not produce an exact solution, Fibonacci still had plenty to say about this equation. Among other things, he showed that Equation (8) has no rational solutions. To do this, he supposes \( p/q \) is a rational solution to Equation (8) with \( \gcd(p,q) = 1 \). Setting \( x = p/q \) in Equation (8) and multiplying by \( q^3 \) to clear the fraction gives \( p^3 + 2p^2q + 10pq^2 - 20q^3 = 0 \), or rather

\[
(9) \quad p^3 = q(20q^2 - 10pq - 2p^2).
\]

This says that \( p \) is a multiple of \( q \). But \( \gcd(p,q) = 1 \), so \( q = \pm 1 \). On the other hand, rewriting Equation (9) as

\[
(10) \quad 20q^3 = p(p^2 + 2pq + 10q^2)
\]

\(^6\)Leonardo Pisano (a.k.a. Fibonacci) was the first prominent Western European mathematician. Born in Pisa, Fibonacci traveled to North Africa and the Middle East where he learned the arithmetic of commerce as well as the Arabic number system and algebra. Fibonacci’s famous sequence is the result of a problem on breeding rabbits.
implies \(20q^3\) is a multiple of \(p\), so \(p|20\) (see Exercise 7). Therefore any rational solution \(p/q\) must be one of \(\{\pm 1, \pm 2, \pm 4, \pm 5, \pm 10, \pm 20\}\). But we can check that none of these values satisfies Equation (8), so there are no rational solutions.\(^7\)

Using Fibonacci’s method described above, we can produce a list of possible rational roots for any polynomial with integer coefficients. This is better known as the Rational Roots Theorem:

**Theorem 5. (Rational Roots Theorem)** Let \(f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0\) be a polynomial with integer coefficients. If \(r,s\) are integers with \(\gcd(r,s) = 1\) and \(r/s\) is a root of \(f(x)\), then \(r|a_0\) and \(s|a_n\).

**Your Turn 14.** Let \(f(x) = 6x^5 - 7x^4 + 2x^3 - 6x^2 + 7x - 2\).

(a) Use the Rational Roots Theorem to find a list of possible rational roots for \(f(x)\). Use this list to determine the actual rational roots of \(f(x)\).

(b) By looking for a pattern in the coefficients, find a factorization of \(f(x)\). Then use this factorization to find the rational roots of \(f(x)\).

### 2.3. Descartes’ Rule of Signs.

While the Rational Roots Theorem gives useful information about the roots of a polynomial in terms of its coefficients, it takes some work to employ. What can we tell just by looking at the coefficients of a polynomial? One thing we can readily determine by looking at the rule of a polynomial is the number of sign changes in its coefficients. For example, the polynomial \(f(x) = 2x^7 - 6x^6 + x^4 + 2x^3 - 5x - 1\) has three sign changes in its coefficients, since the list \(2, -6, 1, 2, -5, -1\) of nonzero coefficients contains three sign changes. Further, if we graph \(f(x)\) (see Figure 10) we find that the number of positive roots for \(f(x)\) is less than the number of sign changes in the coefficients of \(f(x)\). This is not just a strange coincidence, as we shall soon see.

**Your Turn 15.** Argue that the number of positive roots of a quadratic polynomial \(p(x) = x^2 + bx + c\) \((b,c \in \mathbb{R})\) is always less than or equal to the number of sign changes in its coefficients. (Let \(r_1, r_2\) be the roots of \(p(x)\). Then \(b = -(r_1 + r_2)\) and \(c = r_1r_2\) (why?). Is the supposition true when both \(r_1, r_2\) are positive? When one of \(r_1, r_2\) is positive?)

---

\(^7\)This reproduction of Fibonacci’s solution is not entirely faithful. Actually, Fibonacci first shows that any solution must lie strictly between 1 and 2. Then, by showing that \(q = \pm 1\) as above, he concludes that any rational solution \(p/q\) must be an integer. These two facts imply that the equation has no rational solution.
Your Turn 15 is a special case of Descartes’ Rule of Signs:

**Theorem 6.** (Descartes’ Rule of Signs) The number of positive real roots of a nonzero polynomial with real coefficients (counted with multiplicity) is less than or equal to the number of sign changes in its coefficients.

In verifying Descartes’ Rule of Signs, we find many ideas worth exploring. The main thrust is as follows (for details, see Exercise 12): We apply induction on the degree of the polynomial. The result holds for (nonzero) constant polynomials, since there are no roots and no sign changes. Now let \( n > 0 \) be an integer, let \( p(x) = a_n x^n + \cdots + a_0 \) be a polynomial of degree \( n \) with \( a_n > 0 \), and assume the rule of signs holds for polynomials of degree \( n - 1 \). We want to show that the rule of signs works for \( p(x) \). If the constant term \( a_0 > 0 \), then the number of coefficient sign changes in \( p(x) \) must be even, since any sign change to a negative coefficient must be accompanied by a corresponding change back to a positive coefficient. In addition, since the graph of \( p(x) \) crosses the positive \( x \)-axis an even number of times (see Figure 11), the number of positive roots is even. If \( a_0 < 0 \), we can argue similarly that the number of sign changes and the number of positive roots are both odd. Thus the number of coefficient sign changes and the number of positive roots have the same parity. For a contradiction, suppose that \( p(x) \) has more positive roots than sign changes. Then, because of parity considerations, \( p(x) \) must have at least two more roots than sign changes. Since \( p' \) has a root between each pair of roots of \( p(x) \) (Figure 11), this says that \( p' \) has at least one more positive root than \( p(x) \) has sign changes. But \( p'(x) \) has fewer sign changes than \( p(x) \) (why?), and so we conclude that \( p'(x) \)
has more positive roots than coefficient sign changes, contradicting the induction hypothesis. To fill in the
gaps in our justification of Descartes’ Rule of Signs, there is still some work to do:

**Your Turn 16.** Consider the case $a_0 > 0$ as in Figure 11. Is there some way the graph of $p(x)$ could
touch the $x$-axis which was not considered in the figure? In this case, do you think $p(x)$ will still have an
even number of roots if we count repeated roots with multiplicity? Discuss. (Also, see Exercise 14)

**Your Turn 17.** According to Descartes’ Rule of Signs, what is the maximum number of positive real
roots that can be possessed by a depressed cubic $x^3 + px + q$? Does your answer contradict the fact that some
cubics have three positive real roots? Discuss.

### 2.4. Symmetric functions and Viete’s Theorem.

At the beginning of Section 2, we saw that solving $x^3 - x^2 - 2x + 2 = 0$ boils down to finding $\alpha_1, \alpha_2, \alpha_3$ with

$$x^3 - x^2 - 2x + 2 = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3).$$

You might try to solve for $\{\alpha_1, \alpha_2, \alpha_3\}$ by multiplying the factors in the righthand side to obtain

$$x^3 - x^2 - 2x + 2 = x^3 - (\alpha_1 + \alpha_2 + \alpha_3)x^2 + (\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3)x - \alpha_1\alpha_2\alpha_3.$$

Equating coefficients gives a system of equations

$$-1 = -(\alpha_1 + \alpha_2 + \alpha_3) \quad -2 = (\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3) \quad 2 = -\alpha_1\alpha_2\alpha_3$$
that you might try to solve for \( \alpha_1, \alpha_2, \alpha_3 \). While this doesn’t turn out to be very helpful for finding \( \alpha_1, \alpha_2, \alpha_3 \), it does illustrate the fact that, while it is hard to find roots in terms of coefficients, the coefficients of a polynomial can be readily expressed in terms of its roots! This result is known as Viète’s Theorem:

\[ \text{Theorem 7. (Viète’s Theorem)} \]

If \( \alpha_1, \ldots, \alpha_n \) are the roots of \( x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n \), then

\[
\begin{align*}
\alpha_1 + \alpha_2 + \cdots + \alpha_n &= -a_1 \\
\alpha_1\alpha_2 + \alpha_1\alpha_3 + \cdots + \alpha_{n-1}\alpha_n &= a_2 \\
& \vdots \\
\alpha_1\alpha_2 \cdots \alpha_n &= (-1)^n a_n.
\end{align*}
\]

The connection between roots and coefficients given in Viète’s Theorem provides a new perspective on polynomials. For instance, since the coefficient of the \( x^2 \)-term in the depressed cubic \( x^3 + px + q \) is zero, Viète’s Theorem tells us that the complex roots of any depressed cubic must sum to zero. (Compare with part (d) of Your Turn 3 in Section 1.2.)

Viète’s Theorem gives rise to multivariable functions \( s_1, s_2, \ldots, s_n \) defined by

\[
\begin{align*}
s_1(\alpha_1, \alpha_2, \ldots, \alpha_n) &= \alpha_1 + \alpha_2 + \cdots + \alpha_n \\
s_2(\alpha_1, \alpha_2, \ldots, \alpha_n) &= \alpha_1\alpha_2 + \alpha_1\alpha_3 + \cdots + \alpha_{n-1}\alpha_n \\
& \vdots \\
s_n(\alpha_1, \alpha_2, \ldots, \alpha_n) &= \alpha_1\alpha_2 \cdots \alpha_n
\end{align*}
\]

These functions are called the elementary symmetric polynomials in \( n \) variables.

**Your Turn 18.** Let \( s_1, s_2, s_3, s_4 \) denote the elementary symmetric polynomials in four variables.

(a) Find formulas for \( s_1, s_2, s_3, \) and \( s_4 \).

(b) Compute \( s_3(1, 2, 3, 4) \), \( s_3(2, 1, 4, 3) \), and \( s_3(4, 2, 3, 1) \). What do you observe about the outcome of these calculations?

---

Francois Viète (1540–1603) revolutionized symbolic algebra, and was among the first to realize that new mathematics can be created by mechanically manipulating and combining existing formulas. For example, we can apply the known formula \((a + b)^2 = a^2 + 2ab + b^2\) to obtain a formula for \((a + b + c)^2\). This sort of thinking, which seems to have originated with Viète, was a huge leap forward in mathematics.
(c) Discuss the meaning of the word symmetry. Why do you think the functions $s_1, s_2, s_3, s_4$ are called symmetric?

The elementary symmetric functions are ‘elementary’ in that they can be used to express other symmetric functions (i.e., other functions whose output does not depend on the order of the input). For example, the function $\alpha_1^2 + \alpha_2^2 + \alpha_3^2$ is a symmetric function of three variables, and it may be written as

$$\alpha_1^2 + \alpha_2^2 + \alpha_3^2 = (\alpha_1 + \alpha_2 + \alpha_3)^2 - 2(\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3) = s_1^2 - 2s_2,$$

where $s_1, s_2$ are the first two of the three elementary symmetric functions in three variables. We say that $\alpha_1^2 + \alpha_2^2 + \alpha_3^2$ is a “polynomial in the elementary symmetric functions” $s_1$ and $s_2$, since it is obtained by substituting $x_1 = s_1$ and $x_2 = s_2$ in the two-variable polynomial $x_1^2 - 2x_2$. Without proof, we give the famous theorem about symmetric functions:

**Theorem 8.** (Fundamental Theorem on Symmetric Functions) Any symmetric polynomial in variables $\alpha_1, \ldots, \alpha_n$ is a polynomial function of the elementary symmetric polynomials $s_1, \ldots, s_n$ in $n$ variables (that is, a sum of products of the elementary symmetric polynomials). Moreover, any rational (quotient of polynomials) symmetric function in $\alpha_1, \ldots, \alpha_n$ is a rational function of the elementary symmetric polynomials.

2.5. Exercises.

**Conjugate roots**

1. List all the roots of a sixth degree polynomial with real coefficients possessing roots $6 + i$, $2 - 2i$, and $3 + 2i$.

2. List all the monic, fifth degree polynomials with real coefficients whose only roots are $4$, $2 + 3i$, and $2 - 3i$.

3. Prove Corollary 4. (Assume the version of the Fundamental Theorem of Algebra given in Theorem 1.)

**Rational Roots Theorem**
4. Use the Rational Roots Theorem to factor \(2x^3 - 5x^2 - x + 6\) into a product of linear factors.

5. Use the Rational Roots Theorem to show that \(\sqrt{p}\) is irrational whenever \(p\) is a prime number.

6. Let \(p(x) = 2x^{10} + 7x^9 - 4x^8 - 2x^2 - 7x + 4\).
   
   (a) Do you see a quick way to factor \(p(x)\)?
   
   (b) According to the Rational Roots Theorem, what are the possible rational roots of \(p(x)\)?
   
   (c) What are the actual rational roots of \(p(x)\)? Show how you arrived at your answer.
   
   (d) Does \(p(x)\) appear to have any (real) irrational roots? Explain how you arrived at your answer.
   
   (e) Find all the roots of \(p(x)\).
   
   (f) Factor \(p(x)\) into a product of first degree polynomials.
   
   (g) Factor \(p(x)\) into a product of first degree polynomials (with real coefficients) and second degree polynomials (with real coefficients but nonreal roots).

7. If \(\gcd(p, q) = 1\) and \(p | 20q^3\), why must \(p(x)\) be a divisor of 20?

8. Show that Fibonacci’s equation \(x^3 + 2x^2 + 10x - 20 = 0\) has no solutions of form \(\sqrt{p}\), where \(p\) is prime. *(Show that the equation can be rewritten as \(x = 2\left(\frac{10 - x^2}{10 + x^2}\right).\) Then argue by contradiction.)*

9. Prove the Rational Roots Theorem.

Descartes’ Rule of Signs

10. Use Descartes’ Rule of Signs to determine the maximum number of positive real roots for the following polynomials. Check your answers using a graphing device.
   
   (a) \(f(x) = 2.3x^3 - 12.16x^2 + 15x - 2\).
   
   (b) \(g(x) = 4 - 85x + 115x^2 - 56.36x^3 + 11.7x^4 - 0.87x^5\).
   
   (c) \(h(x) = -0.6x^5 + 7.73x^4 - 35.55x^3 + 69.02x^2 - 48.6x + 7\).
   
   (d) \(j(x) = -0.006x^5 - 0.825x^4 - 0.4x^3 + 6.96x^2 - 20.57x + 20\).

11. Using Descartes’ rule of signs, determine whether it is possible that the graph of \(x^8 + 2x^7 - x + 1\) has the shape given in Figure 12.
12. The following questions pertain to the discussion immediately following Theorem 6 (Descartes’ Rule of Signs).

(a) We assumed $a_n > 0$. If $a_n < 0$, what could we do to the polynomial to force the leading coefficient to be positive? Will this change the number of positive roots?

(b) What happens if $a_0 = 0$?

(c) What theorem from calculus ensures that $p'(x)$ has a root between each pair of roots of $p(x)$?

13. How could Descartes’ Rule of Signs be used to find an upper bound on the negative real roots of, say, $f(x) = 2x^7 - 6x^5 + x^3 + 2x^3 - 5x - 1$? Does Figure 10 indicate that your method works?

14. Let $p(x) = a_n x^n + \cdots + a_0$ be a polynomial with real coefficients and $a_n, a_0 > 0$. By the Fundamental Theorem of Algebra, we may write

$$p(x) = (x + n_1) \cdots (x + n_t)(x - q_1) \cdots (x - q_s)f(x)$$

where the $n$’s and $q$’s are positive real numbers and $f(x)$ is a product of quadratic polynomials possessing no real roots and having positive leading coefficients. Use this decomposition to show $p(x)$ has an even number of positive roots. ($a_0 = p(0) = (-1)^s n_1 \cdots n_t q_1 \cdots q_s \cdot C$ where $C$ is the constant term of $f(x)$. Argue that $C > 0$, and so $s$ (the number of positive roots) must be even.)

15. In Exercise 10, Descartes’ Rule of Signs did a better job of approximating the number of positive real roots in parts (a) and (b) than it did in parts (c) and (d). Can you detect any conditions of the coefficients
of a polynomial that seem to affect how well Descartes’ rule of signs predicts the number of positive real roots?

\textit{Viète’s Theorem and Symmetric Polynomials}

16. Use elimination in an attempt to solve the system of equations in (11) for one of $\alpha_1, \alpha_2, \alpha_3$. What disappointing (and perhaps funny, depending on your mood) thing happens?

17. Use Viète’s Theorem to find the unique monic fourth degree polynomial with roots 1, $-1, 2, 0$. (Check your answer by multiplying factors.)

18. Let $g(x)$ be a monic cubic polynomial and suppose $h(x) = g(x - 3)$ is a depressed cubic. What is the sum of the roots of $g(x)$?

19. Use Viète’s Theorem to determine whether the following statements about monic fourth degree polynomials are TRUE or FALSE:
   \begin{itemize}
   \item[(a)] If all the roots are positive, then the polynomial has a negative $x^2$ coefficient.
   \item[(b)] If all the roots are positive, then the polynomial has a negative $x^3$ coefficient.
   \item[(c)] If two roots are positive and two roots are negative, then the polynomial has a negative constant term.
   \end{itemize}

20. Express the following symmetric functions of three variables in terms of the elementary symmetric functions $s_1, s_2, s_3$ of three variables. (Educated trial and error is ok here.)
   \begin{itemize}
   \item[(a)] $x_1^2x_2^2x_3^2 + x_1^2 + x_2^2 + x_3^2 + 2(x_1x_2 + x_1x_3 + x_2x_3)$.
   \item[(b)] $x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2$.
   \end{itemize}

3. Solving Cubic Equations

We have counted and classified roots of cubic polynomials, but up to now we have largely avoided the problem of actually finding these roots. For quadratic polynomials, finding the roots is a routine problem that we have been solving since our secondary school days. However, for cubic (and higher degree) polynomials
this a much harder problem. Historically, the problem of solving polynomial equations (i.e., finding roots of polynomials) has been a driving force in algebra. For example, consider the following quotation:

But of number, cosa, and cubo, however they are compounded..., one cannot give general rules except that, sometimes, by trial,...in some particular cases. And therefore when in your equations you find terms with different intervals without proportion, you shall say that the art...has not given the solutions to this case, even if the case may be possible.

Luca Pacioli, *Summa de arithmetica, geometria, proportioni et proportionalita*, 1494

Pacioli’s quotation may be paraphrased as: *No one (as of 1494) knows a general method for solving cubic equations.* While quadratic equations had been tackled thousands of years beforehand by the Babylonians, the problem of solving general cubic equations (i.e., finding roots for cubic polynomials) remained stubbornly unsolved. It’s not that people weren’t trying: Archimedes, Diophantus, Khayam, and Fibonacci all tried their hand at cubic equations with varying degrees of success, but none arrived at a general method of solution.

All of this was about to change. Within ten years of Pacioli’s statement, the cubic would be solved, and thereby Renaissance mathematicians would significantly exceed the knowledge of the ancients.\(^9\) Perhaps more importantly, the solution of the cubic led to ideas that have shaped algebra as we know it today. In this section we explore the solution of the cubic and its connection with the development of complex numbers.

3.1. Mathematical intrigue: solving the cubic. The first steps in solving cubic equations can be motivated by our experience with quadratic equations:

Let’s Go 2. Consider the quadratic equation \(x^2 + 7x + 5 = 0\).

(a) Substitute \(z = x + 7/2\) in the equation above (that is, replace \(x\) by \(z - 7/2\)) and simplify. Can you easily solve the resulting equation for \(z\)? What familiar technique did you just carry out?

(b) Use your result to part (a) to solve the original equation.

---

\(^9\)Luca Pacioli (1445-1517) was a leading mathematician of his time and close friend of Leonardo da Vinci. Da Vinci provided illustrations for several of Pacioli’s mathematics books.

\(^10\)From the fall of the Roman Empire to the beginning of the renaissance (roughly 450–1450 CE) scholars tended to view themselves as caretakers of ancient knowledge rather than original researchers. The solution of the cubic (and other similar events) helped mark a change in scholarly self-image during the renaissance.
(c) The cubic \( f(x) = x^3 - 14x + 31 \) is the depressed version of \( g(x) = x^3 + 6x^2 - 2x + 5 \). By graphing \( f(x) \) you can see that \( f(x) \) has a root somewhere between \( x = -5 \) and \( x = -4 \). Can you determine this root exactly?

You may have recognized the procedure from part (a) of Let’s Go 2 as completing the square, which removes the middle term from a quadratic equation and enables us to solve quadratic equations easily (part (b)). As we saw in Section 2.1, the analogous process for cubic equations is to produce a depressed cubic, but unfortunately this doesn’t immediately lead to solutions of a cubic equation (see part (c) of Let’s Go 2). However, mathematicians in Pacioli’s time wisely suspected that the removal of a middle term would be helpful, and invariably their first step was to trade in an equation \( x^3 + ax^2 + bx + c = 0 \) for the depressed form \( x^3 + px + q = 0 \). This is evident in Pacioli’s quotation above: Number refers to a constant, cosa (literally ‘thing’) means the object we seek, say \( x \), and cubo means the cube of the ‘thing’, \( x^3 \). So, but of number, cosa, and cubo, however they are compounded... translates to \( x^3 + px + q = 0 \).

In the early 1500’s, Scipione del Ferro and Niccolo Tartaglia independently discovered a ‘general’ method for solving depressed cubic equations. In one of the most ignominious events in mathematics history, Tartaglia grudgingly divulged his solution to Girolamo Cardano on the condition that Cardano not reveal the secret, but Cardano eventually published the method in his work *Ars Magna*.

The method for cubic equations given in *Ars Magna* is roughly as follows: Starting with a cube of side length \( u \) which is dissected as in Figure 13, we observe that

\[(u - v)^3 + 3uv(u - v) - (u^3 - v^3) = 0.\]

\(12\)Scipione del Ferro, a mathematician at the university of Bologna, collaborated with Pacioli. Perhaps Pacioli whetted his appetite for cubic polynomials. Tartaglia (1499–1557) was a self-taught man of humble beginnings who made early contributions to mechanics. As a boy of 12 or 13, he was wounded in the face and throat during the French attack of Brescia. This resulted in a speech impediment, hence the name ‘Tartaglia’ (literally ‘stammerer’).

\(13\)Cardano (1501–1576) was a colorful character and a true ‘renaissance man’, having been a doctor, astrologer, and mathematician. For a time Cardano was the court astrologer to the boy king of England, Edward VI. Cardano predicted long life for the king, but unfortunately the boy died at age 15.

\(14\)Ars Magna (literally ‘The Great Art’) was published in 1545.

\(15\)In the fifteenth century, algebraic symbolism was nearly nonexistent. There was no such thing as an ‘operational calculus’ with symbols (that is, people weren’t manipulating formulas to obtain new formulas), and graphing an equation was still 150 years in the future. As a result, people tended to think geometrically. For example, when Khayyam talked about completing the square, he was thinking of an honest-to-goodness square! Likewise, cubic equations were interpreted in terms of cubes.
Equation (12) holds not only when $u > v$ (as suggested by Figure 13), but also when $u \leq v$ (see Exercise 7).

Now, suppose we wish to solve the depressed cubic equation $x^3 + px + q = 0$. If $u, v$ satisfy

$$
\begin{align*}
  u^3 - v^3 &= -q \\
  uv &= p/3,
\end{align*}
$$

then, by substituting Equation (13) into Equation (12), we see that $x = u - v$ will be a solution to $x^3 + px + q = 0$:

$$(u - v)^3 + p(u - v) + q = (u - v)^3 + 3uv(u - v) - (u^3 - v^3) = 0.$$  

To finish, we need to express the solution $x = u - v$ in terms of the known coefficients $p$ and $q$. Substituting $v = \frac{p}{3u}$ into the top equation of (13) and then multiplying by $u^3$ gives the quadratic equation $(u^3)^2 + qu^3 - \frac{q^2}{27} = 0$. Using the quadratic formula we find that

$$
u^3 = \frac{q}{2} + \frac{1}{2} \sqrt{q^2 + \frac{4p^3}{27}}.$$

Assuming that $q^2 + \frac{4p^3}{27} \geq 0$, one of the possible values for $u$ is

$$
u = \sqrt[3]{-\frac{q}{2} + \frac{1}{2} \sqrt{q^2 + \frac{4p^3}{27}}}.$$
Then, using the fact that $v = \frac{p}{3u}$ (see Exercise 1), we find that the corresponding solution $x = u - v$ for $x^3 + px + q = 0$ is

\[
x = \sqrt[3]{-\frac{q}{2} + \frac{1}{2} \sqrt{\frac{q^2}{4} + \frac{4p^3}{27}}} - \sqrt[3]{q^2 + \frac{4p^3}{27}}.
\]

Equation (15) is the culmination of the method, giving a solution to many depressed cubic equations.

**Your Turn** 19. Use Cardano’s method described above to find a solution to $x^3 + 6x = 20$. (This particular equation appears in Cardano’s Ars Magna.) Argue that the solution you obtain is equal to 2. *(Hint: You may wish to look at a graph of $y = x^3 + 6x - 20$).*

**Your Turn** 20. A quadratic equation plays a key role in Cardano’s method. Find this quadratic equation, determine where it comes from, and explain why it is important that the equation is quadratic.

**Your Turn** 21. The method seems to provide only one solution to a cubic equation.

(a) How many solutions should there be? How do you know?

(b) Given one solution to a cubic, how might you find others?

### 3.2. More on the solution of the cubic: complex numbers.

Complex numbers were unknown in the early 1500’s, so there was food for thought whenever $q^2 + 4p^3/27 < 0$ in the formula (15). For example, consider the equation $x^3 = 15x + 4$. It is not difficult to check that $x = 4$ is a solution, and in fact all three solutions to the equation are real (see Your Turn 11). However, when we apply Equation (15), we obtain

\[
x = \sqrt[3]{2 + \frac{1}{2} \sqrt{-484}} - \sqrt[3]{-2 + \frac{1}{2} \sqrt{-484}},
\]

which not only doesn’t look like 4, but also involves $\sqrt{-484}$. Could the expression in Equation (16) be a disguised version of 4 in much the same way that our solution to $x^3 - 6x - 20 = 0$ (see Your Turn 19) turned out to be a disguised version of 2? Raphael Bombelli (1526-1573) thought so, and, taking motivation from Diophantus,\(^{16}\) he was the first to devise an arithmetic on the complex numbers. To handle the cubic numbers.

---

\(^{16}\)The Greek mathematician Diophantus (2nd or 3rd century CE) was the first to perform arithmetic with negative numbers, and devised the multiplicative rule of signs preserving the distributive property: $(+) \cdot (+) = (+), (+) \cdot (-) = (-), (-) \cdot (+) = (-)$, and $(-) \cdot (-) = (+)$. Bombelli, likely the first European mathematician to examine the work of Diophantus closely, produced a similar rule of signs for operating with $\pm \sqrt{-1}$, and complex arithmetic was born.
equation $x^3 = 15x + 4$, Bombelli noticed that

$$(2 + i)^3 = 2 + 11i = 2 + \frac{1}{2}\sqrt{-484},$$

and similarly $(-2 + i)^3 = -2 + \frac{1}{2}\sqrt{-484}$. Therefore, if we regard $2 + i$ and $-2 + i$ as the cube roots in Equation (16), then

$$x = \sqrt[3]{2 + \frac{1}{2}\sqrt{-484}} - \sqrt[3]{-2 + \frac{1}{2}\sqrt{-484}} = (2 + i) - (-2 + i) = 4.$$ 

The story ends happily, provided that $2 + i$ and $-2 + i$ are chosen as the ‘cube roots’ of $2 + 11i$ and $-2 + 11i$, respectively. But each complex number possesses three cube roots, so which one should be chosen? One way to handle this problem is to consider only principal roots of complex numbers (denoted by fractional exponents).

**Your Turn 22.** The expression $q^2/4p^3/27$ should be vaguely familiar. Where else have you encountered it in this chapter? In what context(s)?

**Your Turn 23.** Recall (see Section 2 of Chapter 11) that $z^{1/q}$ denotes the principal $q$-th root of $z$. Is $2 + i = (2 + 11i)^{1/3}$? Is $-2 + i = (-2 + 11i)^{1/3}$?

**Your Turn 24.** Given that square roots of negative numbers sometimes arise when trying to solve quadratic equations, why do you think cubic equations turned out to be the catalyst for developing complex numbers?

### 3.3. Solution of the cubic: finale.

We have already solved the cubic $x^3 + px + q = 0$ when $q^2 + 4p^3/27 \geq 0$, and to finish we now address the case $q^2 + 4p^3/27 < 0$. Just as before, our solution is $x = u - v$, where $u$ and $v$ satisfy the conditions given in Equation (13). Following the practice of only taking principal roots as established in Section 3.2, we wind up with

$$u = \left[ -\frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right]^{1/3}.$$
To finish we must compute \( v \). Letting \( v = \frac{p}{3u} \), we have

\[
(17) \quad v = \frac{p}{3} \left( -\frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right)^{-1/3}
\]

\[
(18) \quad = \frac{p}{3} \left( \frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right)^{-1/3} \left( -\frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right)^{-1/3}
\]

\[
(19) \quad = \frac{p}{3} \left( \frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right)^{1/3} \left( \frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right)^{-1/3}
\]

\[
(20) \quad = \frac{p}{3} \left( \frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right)^{1/3} \left( \frac{p^3}{27} \right)^{-1/3}
\]

\[
(21) \quad = \frac{p}{|p|e^{(\frac{\pi}{3})}} \left( \frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right)^{1/3}
\]

\[
(22) \quad = -e^{(-\pi/3)} \left( \frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right)^{1/3}
\]

Therefore, in case \( q^2 + 4p^3/27 < 0 \), a solution to \( x^3 + px + q = 0 \) is \( x = u - v \), where

\[
(23) \quad u - v = \left( -\frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right)^{1/3} + e^{(-\pi/3)} \left( \frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2} \right)^{1/3}
\]

Note here that \( e^{(-\pi/3)} \) can be rewritten as \( \frac{1 - i\sqrt{3}}{2} \).

You will be asked to justify two of the equalities above in Exercises 9 and 10.

Equations (15) and (23) may be used to produce a solution to any depressed cubic.

**Your Turn 25.** Use the method described above to find a solution to \( x^3 = 15x + 4 \). Then, argue that your solution is equal to 4. (You may use a calculator that computes principal roots if you wish.)

### 3.4. Exercises.

1. Suppose \( u \) is as in Equation (14) with \( q^2 + 4p^3/27 \geq 0 \). Use fact that \( v = \frac{p}{3u} \) to show that

\[
v = \sqrt[3]{\frac{q}{2} + \frac{1}{2} \sqrt{q^2 + \frac{4p^3}{27}}}
\]

(Multiply the numerator and denominator of \( \frac{p}{3u} \) by \( \sqrt[3]{\frac{q}{2} + \frac{1}{2} \sqrt{q^2 + \frac{4p^3}{27}}} \).)

2. Find at least one solution for each depressed cubic equation using the methods given above. Then, check that the solution you obtain is equal to the given value. (The Rational Roots Theorem may be helpful.)
3. SOLVING CUBIC EQUATIONS

(a) $x^3 - 3x - 2 = 0$ (Check: $x = 2$)
(b) $x^3 - 9x - 80 = 0$ (Check: $x = 5$)
(c) $x^3 + 4x = 80$ (Check: $x = 4$)
(d) $x^3 = 27x + 54$ (Check: $x = 6$)

3. Find all three solutions to each equation in Exercise 2.

4. Let $f(x) = x^3 + 3x + 2$.

(a) Explain why $f(x)$ has exactly one real root.
(b) Explain why this root is not rational.
(b) Use Cardano’s method to find the real root of $f(x)$.

5. Find at least one solution to $x^3 + 12x^2 - x + 5 = 0$. (You must first depress the cubic.)

6. Use a graphing calculator to verify that $x^3 = 15x + 4$ has three real solutions. Then, discuss at least two other ways you could show that $x^3 = 15x + 4$ has three real solutions.

7. Show that $(u - v)^3 + 3uv(u - v) - (u^3 - v^3) = 0$ for all values of $u$ and $v$. (This is an important ingredient of Cardano’s method for solving cubic equations.)

8. Given a depressed cubic $f(x) = x^3 + px + q$ with $q^2 + \frac{4p^3}{27} = 0$, use Cardano’s formula to find the roots of $f(x)$.

9. Recall that familiar “laws” of exponents, such as $(zu)^a = z^au^a$, are often false for non-positive bases (see Section 3 in Chapter 11). So, in asserting the equivalence of Equations (18) and (19), we can’t take for granted that

$$\left(\frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2}\right)^{1/3} \cdot \left( -\frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2}\right)^{1/3} = \left( \left( -\frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2}\right)^{1/3} \cdot \left( \frac{q}{2} + \frac{1}{2} \left( q^2 + \frac{4p^3}{27} \right)^{1/2}\right)^{1/3} \right)^{1/3}$$

Assuming $q^2 + \frac{4p^3}{27} < 0$, show that (24) holds.

10. In Equation (22), explain why $(p^3)^{1/3}$ is not equal to $p$. 
4. Factoring and the Fundamental Theorem of Algebra

As its name suggests, the Fundamental Theorem of Algebra\(^{17}\) (Theorem 1) is a central theorem in algebra. In this section we discuss the origin, importance, and proof of this special theorem.

**Let’s Go 3.** Consider the equation

\[
x^6 + x^5 + x^4 - 2x^2 - 2x - 2 = (x^2 - \sqrt{2})(x^2 + \sqrt{2})(x^2 + x + 1).
\]

(a) Verify Equation (25).

(b) Find all solutions in the complex numbers to \(x^6 + x^5 + x^4 - 2x^2 - 2x - 2 = 0\). Which are real numbers?

(c) What does it mean to factor a polynomial? Do you consider the right side of Equation (25) to be a factored version of the left side?

(d) Discuss the role of factoring in your solution to part (b).

**Let’s Go 4.** Let \(f(x) = x^2 - 14x + 43\).

(a) Is it possible to factor \(f(x)\)?

(b) Find the roots of \(f(x)\) using the quadratic formula.

(c) Do you need to revise your answer to part (a) in light of part (b)?

4.1. Fundamental questions lead to a fundamental theorem. Let’s Go 3 may have reminded you that we often first learn how to solve polynomial equations by factoring. For example, \(x^2 - 3x + 2 = 0\) if and only if \((x - 2)(x - 1) = 0\), so the solutions are \(x = 2\) and \(x = 1\). We can attempt to solve any polynomial equation by factoring, but this is not always easy, and it is far from clear whether it is always possible! Case in point: if you were not given the factorization in Equation (25), could you have found it on your own?

Let’s Go 4 indicates that we don’t need to rely on factoring to solve quadratic equations. Instead, the quadratic formula provides what is known as a solution by radicals. Essentially, a solution by radicals is

\(^{17}\)The Fundamental Theorem of Algebra was first articulated by Albert Girard in *L’invention en algebre* (1629), the first serious attempted proof was given by D’Alembert (1649), and the first correct proof was given by Gauss (1816). The name Fundamental Theorem of Algebra is something of a misnomer—any proof of the theorem must rely on some analytic conditions (e.g., existence of a real root for any odd degree polynomial with real coefficients). A proof of the Fundamental Theorem of Algebra is usually first seen in a complex analysis course.
a specific formula that employs the elementary operations $+, -, \cdot, \div$ together with the extraction of roots (square roots, cube roots, etc.) to express solutions to polynomial equations in terms of the coefficients.

One might wonder whether solutions by radicals exist for higher degree polynomial equations, thus providing a sure-fire alternative to factoring. Earlier in this chapter we saw that a solution by radicals exists for cubic equations (see Equation (15)). Further, in the 1540’s, Ludovico Ferrari used Cardano’s strategy for solving cubic equations to find a solution by radicals for quartic (degree 4) polynomial equations.\footnote{Ferrari was Cardano’s student. Ferrari’s method for quartic was essentially identical to that for the cubic, the three main steps being: depression of the quartic, followed by a system of equations in auxiliary variables, and finally reduction to a cubic equation which could be solved by Cardano’s method.} Unfortunately the same strategy did not seem to work for quintic (5th degree) polynomial equations. The lack of progress in solving the quintic led to two overarching questions that played a key role in the development of modern algebra:

1. Does every polynomial equation with real coefficients have a solution by radicals?

2. To what extent can a given polynomial be factored?

Unfortunately, after nearly three hundred years of work, it turned out that the answer to the first question is NO. In an 1826 paper appearing in the first volume of Crelle’s journal, Niels Abel showed that generally speaking, polynomial equations of degree five and higher cannot be solved by radicals (that is, in a manner similar to that for polynomials of degree four and lower). However, the work on question 1 led to the Fundamental Theorem of Algebra (Theorem 1), thus providing an encouraging answer to question 2. So in situations where we must resort to factoring to investigate the solutions of a polynomial equation, at least we know the factoring is theoretically possible.

There are many different (equivalent) ways to express the Fundamental Theorem of Algebra, including:

**Theorem 9.** (Fundamental Theorem of Algebra: Version 1) If $n$ is a positive integer and $f(x)$ is a polynomial of degree $n$ with complex coefficients, then there is a unique complex number $c$ and a unique collection $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ of complex numbers (not necessarily distinct) satisfying

$$f(x) = c(x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_n).$$
**Theorem 10.** (Fundamental Theorem of Algebra: Version 2) Any nonconstant polynomial with complex coefficients has a complex root.

**Theorem 11.** (Fundamental Theorem of Algebra: Version 3) If \( n \) is a positive integer and \( f(x) \) is a polynomial of degree \( n \) with real coefficients, then there is a unique real number \( c \) and a unique collection \( \{\alpha_1, \alpha_2, \ldots, \alpha_n\} \) of complex numbers (not necessarily distinct) satisfying

\[
f(x) = c(x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_n).
\]

**Theorem 12.** (Fundamental Theorem of Algebra: Version 4) Any polynomial with real coefficients may be written as a product of linear and quadratic polynomials with real coefficients, where the quadratic polynomials have no real roots.

**Theorem 13.** (Fundamental Theorem of Algebra: Version 5) Any nonconstant polynomial with real coefficients and even degree may be written as a product of quadratic polynomials with real coefficients.

**Your Turn 26.** Suppose \( f(x) = -126 + 27x + 76x^2 - 51x^3 + 19x^4 - 6x^5 + x^6 \). Given that

\[
f(x) = (x - 3i)(x + 3i)(x^2 - 5x + 7)(x - 2)(x + 1),
\]

show that \( f(x) \) satisfies each version of the Fundamental Theorem of Algebra given above.

**Your Turn 27.** Theorems 9 and 11 look remarkably similar.

(a) How do the statements of the two theorems differ?

(b) On the surface, which of the theorems appears to be stronger?

### 4.2. Division algorithm and factor theorem.

In preparation for further discussion of the Fundamental Theorem of Algebra, in this section we precisely define the word factor and establish a bedrock result about factoring known as the Factor Theorem.

Factoring presents us with a problem that is the opposite of, and is much harder than, polynomial multiplication.\(^{19}\) To understand factoring we must first understand division. The division of polynomials is governed by the Division Algorithm:

\(^{19}\)An entire branch of mathematics, known as inverse problems, is devoted to understanding ‘opposite’ problems such as this one.
Figure 14. Long division of polynomials

**Theorem 14. (Division Algorithm)** Given polynomials \( f(x) \) and \( g(x) \) with complex coefficients and \( g(x) \neq 0 \), there exist unique polynomials \( q(x) \) and \( r(x) \) with complex coefficients, called the quotient and remainder, respectively, satisfying

(i) \( f(x) = g(x)q(x) + r(x) \), and

(ii) Either \( \deg r(x) < \deg g(x) \) or \( r(x) = 0 \).

For example, when dividing \( f(x) = 3x^5 + 2x^4 + 2x^3 + 4x^2 + x - 2 \) (the dividend) by \( g(x) = 2x^3 + 1 \) (the divisor), the quotient is \( q(x) = \frac{3}{2}x^2 + x + 1 \) and the remainder is \( r(x) = \frac{5}{2}x^2 - 3 \). Practically speaking, the quotient and remainder can be found by long division (see Figure 14).

In the special case that the Division Algorithm yields \( f = gq \) with \( r = 0 \), we say that \( g(x) \) is a factor\(^{20}\) of \( f(x) \). For example, in Equation (25) we see that \( x^2 + x + 1 \) is a factor of \( x^6 + x^5 + x^4 - 2x^2 - 2x - 2 \). Thus, factoring usually means the process of expressing a polynomial as a product of two or more (lower degree) factors.

**Your Turn 28.** Consider the Division Algorithm for polynomials.

(i) Compare the Division Algorithm for polynomials with the Division Algorithm for integers.

(ii) Do you think the Division Algorithm will be true if we replace ‘complex coefficients’ by ‘rational coefficients’? By ‘integer coefficients’?  

\(^{20}\)There are many equivalent ways to say this, such as: \( g(x) \) is a divisor of \( f(x) \); \( f(x) \) is a multiple of \( g(x) \); \( g(x) \) divides \( f(x) \).
Your Turn 29. Use long division to show that \( x - 4 \) is a factor of \( 2x^6 - 9x^5 + 4x^4 + 3x^3 - 11x^2 - 12x + 32 \). Then, find a solution to \( 2x^6 - 9x^5 + 4x^4 + 3x^3 - 11x^2 - 12x + 2 = 0 \).

Your Turn 29 indicates a relationship between linear factors (specifically, those of the form \( x - \alpha \)) of a polynomial \( f(x) \) and its roots. This connection, first articulated by Rene Descartes in 1637, is known as the Factor Theorem\(^{21}\):

**Theorem 15.** (Factor Theorem) Let \( f(x) \) be a polynomial with complex coefficients and \( \alpha \in \mathbb{C} \). Then \( f(\alpha) = 0 \) if and only if \( x - \alpha \) is a factor of \( f(x) \).

**Proof.** By the Division Algorithm, there exist unique polynomials \( q(x) \) and \( r(x) \) with

\[
f(x) = (x - \alpha)q(x) + r(x),
\]

where either \( \deg r(x) < \deg(x - \alpha) \) or \( r(x) = 0 \). Observe that these conditions force \( r(x) \) to be a constant polynomial. Therefore

\[
f(\alpha) = 0 \iff (\alpha - \alpha)q(\alpha) + r(\alpha) = 0 \iff r(\alpha) = 0 \iff r = 0 \iff (x - \alpha) \text{ is a factor of } f.
\]

Your Turn 30. Construct a polynomial \( f(x) \) with roots \( 2, -1, -1 + i, -1 - i, 5 \) such that \( f(1) = 10 \).

4.3. Euler’s ‘proof’ of the Fundamental Theorem of Algebra. You may be surprised to learn that the Fundamental Theorem of Algebra (Theorem 1) has even more proofs than it has statements. Many famous mathematicians\(^{22}\) have tried their hand at proving the theorem, and successful proofs involve various (superficially unlikely) branches of mathematics, such as complex analysis and topology.\(^{23}\) In this section we sketch a largely algebraic argument due to Leonhard Euler (1707-1783).

\(^{21}\)In *La Geometrie*, Descartes states: *It is evident that the sum of an equation containing several roots is always divisible by a binomial consisting of the unknown quantity diminished by the value of one of the true roots...*. In addition to the Factor Theorem, *La Geometrie* contains some of the first thoroughly modern algebraic notation, including using the letters \( x \) and \( y \) for unknowns.

\(^{22}\)This list includes D’Alembert, Euler, Gauss, Laplace, and Cauchy. The first ‘rigorously correct’ proof is due to Gauss (1799).

\(^{23}\)As mentioned earlier, the most ‘common’ proof of the Fundamental Theorem of Algebra uses complex analysis, following immediately from Liouville’s Theorem: *If a function of one complex variable is complex-differentiable at each point of the complex plane and is also bounded, then the function must be constant.*
Euler began his attempt at the Fundamental Theorem of Algebra by arguing that any fourth degree polynomial with real coefficients could be factored as a product of quadratic polynomials with real coefficients.24 (This is a special case of Theorem 13.) The gist of the argument is as follows: Start with a depressed quartic \( f(x) = x^4 + Bx^2 + Cx + D \) with \( B, C, D \) being real numbers, and (to ensure the annihilation of the cubic term) set
\[
x^4 + Bx^2 + Cx + D = (x^2 - ux + \alpha)(x^2 + ux + \beta).
\]
By multiplying the righthand side of Equation (26), we see that the result will have a nonzero \( x^3 \) term unless \( u = v \). Setting \( u = v \) gives
\[
x^4 + Bx^2 + Cx + D = (x^2 - ux + \alpha)(x^2 + ux + \beta).
\]
The goal now is to show that \( u, \alpha, \) and \( \beta \) can be chosen to be real numbers (but we focus only on \( u \)). Toward this end, Euler assumes that there are four roots \( p, q, r, s \) for \( f(x) \). (In the absence of the Fundamental Theorem of Algebra, we cannot be certain yet that \( p, q, r, s \) are complex numbers!) Then, by the Factor Theorem (Theorem 15) we have
\[
f(x) = (x^2 - ux + \alpha)(x^2 + ux + \beta) = (x - p)(x - q)(x - r)(x - s).
\]
In Equation (28), each of the quadratic terms is necessarily a product of two of the linear terms, so for example we could have
\[
x^2 - ux + \alpha = (x - p)(x - q) = x^2 - (p + q)x + pq,
\]
in which case \( u = p + q \). In all, there 6 possible choices for \( u \):
\[
\begin{align*}
c_1 &= p + q & c_4 &= r + s \\
c_2 &= p + r & c_5 &= q + s \\
c_3 &= p + s & c_6 &= q + r
\end{align*}
\]
Therefore \( u \) must satisfy the equation
\[
(x - c_1)(x - c_4)(x - c_2)(x - c_5)(x - c_3)(x - c_6) = 0.
\]
Further, since \( f(x) = (x - p)(x - q)(x - r)(x - s) \) has no \( x^3 \) term, by Viète’s Theorem (Theorem 7) we find that \( p + q + r + s = 0 \), which implies \( c_4 = -c_1, c_5 = -c_2, \) and \( c_6 = -c_3 \). Therefore, upon multiplying pairs

24 Before Euler, it was not at all clear that fourth degree polynomials admitted such factorizations. For example, Nicolaus Bernoulli (1687–1759) conjectured that \( x^4 - 4x^3 + 2x^2 + 4x + 4 \) could not be factored into a product of quadratic polynomials with real coefficients. Gottfried Leibniz (1646–1716), famous as a founding developer of calculus, believed \( x^4 + 1 \) could not be factored into a product of quadratics with real coefficients.
of terms in Equation (29), we see that $u$ is necessarily a root of

$$F(x) = (x^2 - c_1^2)(x^2 - c_2^2)(x^2 - c_3^2).$$

Euler then appeals to the Fundamental Theorem of Symmetric Functions (Theorem 8) to conclude that

- $c_1 c_2 c_3$ is a real number (see Exercise 9), and
- $F(x)$ has real coefficients.

These items imply that the constant term for $F(x)$ (namely $-(c_1 c_2 c_3)^2$) will be negative, and, since $F(x)$ has real coefficients with a positive leading coefficient, $F(x)$ will be positive for large values of $x$ (see Figure 15). Therefore, by the Intermediate Value Theorem, $F(x)$ has a real root $u_0$. By setting $u = u_0$, quadratic factors in Equation (26) can be chosen to have real coefficients. (It turns out that if $u_0$ is real, then $\alpha$ and $\beta$ may also be chosen to be real.) We conclude that

$$f(x) = x^4 + Bx^2 + Cx + D = (x^2 - u_0 x + \alpha)(x^2 - u_0 x + \beta)$$

is a factorization of $f(x)$ into a product of two quadratic polynomials with real coefficients.

**Your Turn 31.** Euler’s argument is supposed to convince us that any quartic polynomial with real coefficients can be factored into a product of quadratics with real coefficients.

(a) Euler assumes the existence of roots $p, q, r, s$ for $f(x)$. Should we assume these roots are real numbers? Should we assume they are complex numbers?

(b) How does the argument use ideas from calculus?
The same ideas used in Euler’s quartic factorization argument lead to the following more general fact:

Any polynomial of degree $2^n$ with real coefficients can be expressed as a product of two polynomials of degree $2^{n-1}$ with real coefficients. This is of paramount importance because it says that most nonconstant polynomials with real coefficients can be factored considerably. For example, suppose we want to know whether a degree 24 polynomial $p(x)$ with real coefficients can be factored. We first multiply $p(x)$ by $x^8$ to obtain a polynomial of degree 32. Since 32 is a power of two, repeated application of Euler’s general fact above indicates that $x^8p(x)$ can be written as a product of quadratic polynomials. Now, some of these quadratics (perhaps all) will factor into a product of linear terms with real coefficients. At this stage, we cancel the eight linear factors $x^8 = x \cdot x \cdot x \cdot x \cdot x \cdot x \cdot x \cdot x$, and what remains is a factorization of $p(x)$ into quadratic and linear polynomials with real coefficients. Thus we have arrived at the Fundamental Theorem of Algebra.\(^{25}\)

4.4. Exercises.

1. Here we address the Leibniz conjecture that $x^4 + 1$ cannot be factored into a product of two quadratic polynomials with real coefficients.

   (a) Find all the fourth complex roots of $-1$. Use your answer together with the Fundamental Theorem of Algebra (Theorem 1) to express $x^4 + 1$ as a product of four linear factors.\(^{26}\)

   (b) Use part (a) to factor $x^4 + 1$ as a product of two quadratic factors with real coefficients. (Try multiplying linear factors corresponding to conjugate roots.)

2. Here we examine local extrema for the graph of a polynomial.

   (a) Is it possible for a polynomial of degree $n$ with complex coefficients to have more than $n$ roots?

   (b) Why must the graph of an $n$-th degree polynomial $f(x)$ (with real coefficients) have no more than $n - 1$ local extrema?

---

\(^{25}\) Gauss had the same objections to Euler’s proof that you discovered in Your Turn 31! In particular, he felt that the existence of the roots $p, q, r, s$ could not be safely assumed. Later, Leopold Kronecker showed that any polynomial $p(x)$ with coefficients in a field of numbers must have a root in some larger field.

\(^{26}\) Since polar representation of complex numbers had not yet been discovered, Leibniz (and others of his time) could not express $i^{1/2}$ as a complex number. This kept Leibniz from factoring $x^4 + 1$. 
(c) Graph the Bernoulli polynomial \( f(x) = x^4 - 4x^3 + 2x^2 + 4x + 4 \) on the interval \([-10, 10]\). Explain why the graph of \( f(x) \) can not turn around and cross the \( x \)-axis somewhere outside the viewing window.

3. Here we look at the factorization of cubic polynomials.

(a) Using ideas from calculus, explain why the graph of an odd-degree polynomial with real coefficients must cross the horizontal axis.

(b) Without appealing to the Fundamental Theorem of Algebra, explain why a cubic polynomial with real coefficients must factor into a product of a linear polynomial and a quadratic polynomial (both with real coefficients).

4. In this exercise, we see that sometimes two numbers written using radicals may be equal to each other, without this equality being obvious. Let \( p(x) = x^4 - 10x^2 + 1 \).

(a) Let \( r_1 = \sqrt{2} + \sqrt{3} \), \( r_2 = \sqrt{2} - \sqrt{3} \), \( r_3 = -\sqrt{2} + \sqrt{3} \), and \( r_4 = -\sqrt{2} - \sqrt{3} \). Show that each of these numbers is a root of \( p(x) \).

(b) Using the quadratic formula, find expressions for the four roots of \( p(x) \).

(c) Identify each of the roots in (b) as \( r_1 \), \( r_2 \), \( r_3 \), or \( r_4 \). How many strategies can you find?

(d) If you had been given \( p(x) \) and asked to find all the roots, would you have been more likely to find the list in (a) or in (b)?

(e) Using the roots of \( p(x) \), find three different ways of factoring \( p(x) \) into a product of two quadratic polynomials. Do you prefer using the list of roots in (a) or the list in (b)?

5. Let \( q(x) = x^2 - 14\sqrt{2}x + 87 \). Find a fourth degree polynomial \( p(x) \) with integer coefficients whose roots include the roots of \( q(x) \).

6. The following questions pertain to Euler’s factorization of a real quartic polynomial into a product of real quadratics:

(a) What substitution must be made to transform a quartic polynomial \( g(x) = x^4 + ax^3 + bx^2 + cx + d \) into a depressed quartic polynomial?
(b) Suppose we have
\[ x^4 + Ax^2 + Bx + C = (x^2 + bx + c)(x^2 + ax + d). \]
Show that \( b = -a. \)

(c) Let \( p, q, r, s \) be as in Euler’s factorization. Why must \( p + q + r + s = 0? \)

7. Many of us are familiar with the procedure of ‘rationalizing the denominator.’ As a typical example,
\[ \frac{1 + 7\sqrt{2}}{5 + 3\sqrt{2}} = \left( \frac{1 + 7\sqrt{2}}{5 + 3\sqrt{2}} \right) \left( \frac{5 - 3\sqrt{2}}{5 - 3\sqrt{2}} \right) = \frac{-37 + 32\sqrt{2}}{7} = \frac{-37}{7} + \frac{32}{7}\sqrt{2}. \]

(a) What is the purpose of rationalizing the denominator?

(b) Perform a similar computation to simplify \( \frac{1 + 7i}{5 + 3i} \). What is the purpose of this computation?

(c) More generally, suppose \( q \) is a rational number that is not the square of another rational number.

Prove that for every choice of rational numbers \( \{a, b, c, d\} \) with \( cd \neq 0 \), the real number \( \frac{a + b\sqrt{q}}{c + d\sqrt{q}} \) can be written as \( A + B\sqrt{q} \) for some rational numbers \( A \) and \( B \).

(d) How might you ‘rationalize the denominator’ for the expression \( \frac{1}{2 + 7\sqrt{2} + 5\sqrt{3}} \)?

(e) How might you ‘rationalize the denominator’ for the expression \( \frac{1}{1 + \sqrt{2}} \)?

8. Assuming the Fundamental Theorem of Algebra, show that every odd-degree polynomial with real coefficients has at least one real root.

9. Let \( f(x) = (x - p)(x - q)(x - r)(x - s) \) and put \( c_1 = p + q, \ c_2 = p + r, \ \text{and} \ c_3 = p + s. \) Further, assume \( f(x) \) has real coefficients and \( p + q + r + s = 0. \)

(a) Show that \( c_1c_2c_3 = (p + q)(p + r)(p + s) \) is a symmetric function of the roots \( p, q, r, s. \) \( \text{(Note that } p = -(q + r + s).) \)

(b) Euler argued that part (a) together with the Fundamental Theorem of Symmetric Functions (Theorem 8) implies that \( c_1c_2c_3 \) must be a real number. In fact the Fundamental Theorem of Symmetric Functions is not quite enough to guarantee this. Why?

10. Let \( p(x) = a_nx^n + \cdots + a_1x + a_0 \) be a polynomial with complex coefficients, and let \( \bar{p}(x) = \bar{a}_nx^n + \cdots + \bar{a}_1x + \bar{a}_0 \) be the polynomial obtained by conjugating the coefficients of \( p(x). \)
(a) Show that if \( z \) is a complex number, then \( z + \bar{z} \) is a real number.

(b) Show that \( p(x)\overline{p}(x) \) is a polynomial with real coefficients.

(c) By applying Theorem 11 to \( p(x)\overline{p}(x) \), show that Theorem 9 holds for \( p(x) \). (By Theorem 11, \( p(x)\overline{p}(x) \) can be expressed as a product of linear factors with complex coefficients. Each linear factor is either a factor of \( p(x) \) or of \( \overline{p}(x) \).)

11. Here we investigate another method for factoring a depressed quartic with real coefficients into a product of two real quadratics. (This is Euler’s method, and motivated his more general argument given in the text.)

(a) Suppose we set

\[
x^4 + Bx^2 + Cx + D = (x^2 - ux + \alpha)(x^2 + ux + \beta).
\]

Show that \( B = \alpha + \beta - u^2, C = (\beta - \alpha)u, \) and \( D = \alpha\beta \).

(b) Show that \( 2\beta = B + u^2 + C/u \) and \( 2\alpha = B + u^2 - C/u \). (The first two equations in part (a) indicate \( \beta + \alpha = B + u^2 \) and \( \beta - \alpha = C/u \).)

(c) Using the third equation from part (a) together with part (b), show that \( 4D = u^4 + 2Bu^2 + B^2 - C^2/u^2 \).

(d) Multiplying by \( u^2 \) in part (c), we obtain

\[
(31) \quad u^6 + 2Bu^4 + (B^2 - 4D)u^2 - C^2 = 0.
\]

Recalling that \( B, C, D \) are given real numbers, use calculus to explain why the equation above has a real solution \( u = u_0 \). (What happens to the left side when \( u = 0 \)? What happens as \( u \) tends to (positive) infinity?)

(e) Use the existence of a real solution \( u = u_0 \) together with part (b) to show that \( \alpha \) and \( \beta \) can be chosen to be real. Explain how this shows that \( x^4 + Bx^2 + Cx + D \) can be factored into two real quadratic polynomials.
12. Use Exercise 11 to show that the Bernoulli polynomial \( x^4 - 4x^3 + 2x^2 + 4x + 4 \) can be factored into the product
\[
\left[ x^2 - \left( 2 + \sqrt{4 + 2\sqrt{7}} \right) x + \left( 1 + \sqrt{4 + 2\sqrt{7}} + \sqrt{7} \right) \right] \left[ x^2 - \left( 2 - \sqrt{4 + 2\sqrt{7}} \right) x + \left( 1 - \sqrt{4 + 2\sqrt{7}} + \sqrt{7} \right) \right].
\]
(First make a substitution to depress the quartic, then apply Exercise 11. Equation (31) is a disguised cubic equation which can be solved by Cardano’s method.)

5. Application: Newton’s Method and Polynomials

In previous sections we’ve seen that it can be difficult to find the roots of a polynomial exactly. However, there are various numerical techniques that can provide close approximations of zeros for many kinds of functions, including polynomials. One such technique is Newton’s method. In this section we review Newton’s method and see that our knowledge of polynomials can be used to improve its efficiency.

Let’s Go 5. Your graphing calculator can be used to estimate zeros of functions in a number of ways:

(a) Your graphing calculator likely has a built-in program that estimates zeros of functions. Use this program to estimate a positive root for \( f(x) = 3x^4 + x^2 + 3x - 11 \). What information does your calculator require to perform this task? Why do you think your calculator requires this information?

(b) Graph \( f(x) \) on your graphing calculator and use a table or ‘trace’ feature to estimate a positive root for \( f(x) \). How does your estimate compare with that of part (a)?

5.1. Newton’s method: a functional iteration scheme. Newton’s method is a numerical procedure for estimating a zero of a function. The method relies on the idea that the tangent line to a curve at a given point provides the best linear approximation to the curve near that point. Specifically, suppose \( p_0 \) is an initial estimate of a zero for a differentiable function \( f \). Using the tangent line approximation to \( f \) at \( p_0 \), we have
\[
(32) \quad f(x) \approx f'(p_0)(x - p_0) + f(p_0)
\]
for values of \( x \) ‘close’ to \( p_0 \) (see Figure 16). Using this approximation, \( f(x) = 0 \) becomes

\(^{27}\)We say that \( a \in \mathbb{R} \) is a zero of a function \( f(x) \) if \( f(a) = 0 \).

\(^{28}\)The method is sometimes called the Newton-Raphson method after Joseph Raphson (1648–1715). While the method first appears in Raphson’s Universal Analysis of Equations (1691), Newton wrote about the method twenty years earlier in his Methods of Fluxions, which was not published until 1736.
Figure 16. The tangent to \( y = f(x) \) at \( x = p_0 \).

\[
(33) \quad f'(p_0)(x - p_0) + f(p_0) = 0
\]

(the equation determining the \( x \)-intercept of the tangent line), and solving for \( x \) in Equation 33 gives

\[
(34) \quad x = p_0 - \frac{f(p_0)}{f'(p_0)}
\]

Let \( p_1 \) denote the value of \( x \) satisfying Equation (34). We may then repeat the whole process described above with \( p_1 \) in place of \( p_0 \) to obtain a third estimate \( p_2 \) for a zero of \( f \), and so on (see Figure 17).\footnote{Graphing calculators use a variant of Newton’s method, known as the secant method, that uses a difference quotient in place of the derivative (the reason being that calculating a derivative can be hard.)}

The end product is a sequence \( \{p_0, p_1, p_2, \ldots \} \), which in favorable cases converges to a zero of \( f \).

In addition to the successive tangent viewpoint described above and shown in Figure 17, Newton’s method may also be viewed in terms of function iteration (that is, successive composition of a function with itself). To see this, we use the righthand side of Equation (34) to define a function \( G(x) = x - \frac{f(x)}{f'(x)} \). If \( p_0 \) is
our initial estimate of a zero for \( f \), then \( p_1 = G(p_0) \), \( p_2 = G(p_1) = G(G(p_0)) \), and in general \( p_k = G^k(p_0) \) where \( G^k \) denotes the \( k \)-fold composition of \( G \) with itself. Observe that a number \( p \) (with \( f'(p) \neq 0 \)) is a zero for \( f \) if and only if \( G(p) = p \):

\[
G(p) = p \iff p = p - \frac{f(p)}{f'(p)} \iff \frac{f(p)}{f'(p)} = 0 \iff f(p) = 0.
\]

Such a point \( p \) is called a fixed point for \( G \).

As an illustration, we approximate a solution to the equation \( \sin x = 1 - x^2 \) (see lefthand panel of Figure 18). This is the same as approximating a zero for \( f(x) = \sin x + x^2 - 1 \) (righthand panel of Figure 18). Newton’s method is invoked by computing

\[
G(x) = x - \frac{f(x)}{f'(x)} = x - \frac{\sin x + x^2 - 1}{\cos x + 2x},
\]

and then selecting an initial estimate \( p_0 \) (say \( p_0 = 1 \)) ‘nearby’ one of the zeros. Repeated application of \( G \) gives

\[
p_0 = 1 \quad p_1 = G(p_0) = 0.6687516352 \quad p_2 = G(p_1) = 0.6370680331 \quad p_3 = G(p_2) = 0.6367326888,
\]

and \( p_3 \) is already accurate to seven decimal places (see Exercise 3).

Newton’s method begs a number of significant questions, including:

- For a given differentiable function \( f \) and initial estimate \( p_0 \), does \( \{p_0, p_1, \ldots\} \) always converge to a zero of \( f \)? (Unfortunately the answer is ‘no.’)
- Provided that \( \{p_0, p_1, \ldots\} \) converges, how ‘fast’ does it converge?
• When do we stop? (Specifically, which \( p_k \) do we consider to be our final approximation?)

• How do we accurately and efficiently employ Newton’s method?

While the first three questions are interesting and important, we choose to focus on the last question. Specifically, in the next subsection we see that our knowledge of polynomials can be used to improve the efficiency of Newton’s method.

Your Turn 32. Recall the graph of \( f(x) = \sin x + x^2 - 1 \) from above (Figure 18).

(a) Graph \( f' \) using a graphing device.

(b) There are two zeros of \( \sin x + x^2 - 1 \), one positive and one negative. By looking at the graphs of \( f \) and \( f' \), determine which values of \( p_0 \) will force Newton’s method to converge to the positive zero, and which \( p_0 \) force convergence to the negative zero.

(c) How is the behavior of \( f' \) related to your answer?

5.2. Newton’s method, polynomials, and the Remainder Theorem. In order to use Newton’s method, a function \( f \) and its derivative \( f' \) must be evaluated repeatedly. In the case that \( f \) is a polynomial function, we see that this can be done efficiently by using synthetic division.

Let’s Go 6. Let \( f(x) = 3x^4 + x^2 + 3x - 11 \).

(a) Without using a calculator, evaluate \( f(2) \) by first substituting \( x = 2 \) into the rule for \( f \), and then by synthetic division. (If you don’t recall synthetic division, consult a classmate.)

(b) How much ‘work’ (in terms of the number of multiplications and additions) is required to evaluate \( f(2) \) using the rule for \( f \)? How much work is required to evaluate \( f(2) \) using synthetic division? (Do not count exponentiation as a single operation.)

Let \( f(x) \) be a polynomial function and \( x_0 \in \mathbb{R} \). By the Division Algorithm (Theorem 14) there are unique polynomials \( q(x) \) and \( r(x) \) satisfying

\[
(35) \quad f(x) = (x - x_0)q(x) + r(x)
\]
where the \( r(x) \) is a \textit{constant} polynomial (why?). Evaluating both sides of Equation (35) at \( x = x_0 \) yields \( f(x_0) = r(x_0) \). So, to evaluate \( f \) at \( x = x_0 \) it suffices to find the remainder\(^{30} \) upon dividing \( f(x) \) by \( x - x_0 \).

With visions of lengthy long-division calculations dancing in our heads, it might seem grossly inefficient to evaluate \( f(x_0) \) by computing a remainder. However, when the divisor is linear (like \( x - x_0 \)), the process of synthetic division shortens the computation dramatically:

\textbf{Theorem 16. (Synthetic Division)} Suppose \( a_0, a_1, \ldots, a_n \in \mathbb{R} \) and

\[ f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0. \]

Given \( x_0 \in \mathbb{R} \), put

- \( b_n = a_n \),
- \( b_k = a_k + b_{k+1} x_0 \) for \( 0 \leq k \leq n - 1 \), and
- \( q(x) = b_n x^{n-1} + b_{n-1} x^{n-1} + \cdots + b_2 x + b_1 \).

Then \( f(x) = (x - x_0) q(x) + b_0 \) and \( f(x_0) = b_0 \).

Verification of Theorem 16 is requested in Exercise 4. As an illustration the theorem, we evaluate

\[ f(x) = 3x^5 + 2x^4 - 4x^2 + x + 9 \text{ at } x = -2 \text{ (see Figure 19).} \]

\[ 3x^5 + 2x^4 - 4x^2 + x + 9 = (x + 2)(3x^4 - 4x^3 + 8x^2 - 20x + 41) - 73, \quad f(-2) = -73, \]

\(^{30}\)This is sometimes called the \textit{Remainder Theorem}. 

\textbf{Figure 19.} Theorem 16 and the process of synthetic division

The upshot of the computation is that
and that all of this information only cost us five multiplications and five additions.

This leads to the moral of Theorem 16 and the subsequent computation: For a polynomial \( f \) of degree \( n \) with all nonzero coefficients, **evaluation of \( f \) via synthetic division requires only \( n \) additions and \( n \) multiplications**, as opposed to the \( \frac{n(n+1)}{2} \) multiplications and \( n \) additions required by using the rule for \( f \) (see Exercise 8). This makes synthetic division a useful tool in numerical procedures (such as Newton’s method) that call for function evaluation.

Newton’s method also demands that we evaluate \( f' \), and again synthetic division comes to the rescue. To see why, first observe that the synthetic division process used to evaluate \( f(x_0) \) also determines a polynomial \( q(x) \) with

\[
\tag{36}
f(x) = q(x)(x - x_0) + f(x_0). 
\]

(In the case \( f(x) = 3x^5 + 2x^4 - 4x^2 + x + 9 \) and \( x_0 = -2 \), we saw in the example above that \( q(x) = 3x^4 - 4x^3 + 8x^2 - 20x + 41 \), where the coefficients for \( q(x) \) are found on the bottom row of the table given in Figure 19.) By differentiating both sides of Equation (36) we obtain \( f'(x) = q(x) + q'(x)(x - x_0) \), and evaluation at \( x = x_0 \) gives

\[
f'(x_0) = q(x_0). 
\]

Therefore, we may evaluate \( f'(x_0) \) by evaluating \( q(x_0) \) (again by synthetic division), where \( q(x) \) is the quotient upon dividing \( f(x_0) \) by \( (x - x_0) \).

All of this suggests a ‘double’ synthetic division process when evaluating both \( f \) and \( f' \). For example, suppose we want to apply Newton’s method to \( f(x) = 3x^5 + 2x^4 - 4x^2 + x + 9 \) with \( p_0 = -2 \). The ‘double’ synthetic division in Figure 20 indicates that \( f(-2) = 73 \) and \( f'(-2) = q(-2) = 193 \), so that

\[
\tag{37}
p_1 = G(p_0) = G(-2) = -2 - \frac{f(-2)}{f'(-2)} = -2 - \frac{-73}{193} \approx -1.62176
\]

To obtain the next iterate \( p_2 \) for Newton’s method, we simply repeat the double synthetic division process above with \(-1.62176\) in place of \(-2\).

**Your Turn 33.** Let \( f(x) = 3x^5 + 2x^4 - 4x^2 + x + 9 \).

(a) Complete two more iterations (i.e., find \( p_2 \) and \( p_3 \)) of Newton’s method with \( p_0 = -2 \) to approximate a zero for \( f \), being sure to use synthetic division to evaluate \( f \) and \( f' \).
Figure 20. Evaluating both \( f \) and \( f' \) using synthetic division

(b) Graph \( f \) and use a trace feature to approximate the zero for \( f \) lying between \(-2\) and \(-1\). How does the result compare with your answer to part (a)?

5.3. Exercises.

1. Using the initial estimate \( p_0 = 0 \), employ three iterations of Newton’s method to approximate the (unique) zero of \( g(x) = \cos x - 2x \). (The zero of \( g \) features prominently in Your Turn 32.)

2. Using the initial estimate \( p_0 = 2 \), employ three iterations of Newton’s method to approximate a zero of \( f(x) = 3x^4 + x^2 + 3x - 11 \). In your calculation, use synthetic division to evaluate \( f \) and \( f' \).

3. Earlier in this section we claimed that 0.6367326888 agrees with a true zero for \( f(x) = \sin x + x^2 - 1 \) up to seven decimal places. Argue that this is the case by computing both \( f(0.63673268) \) and \( f(0.63673260) \) on a calculator. (What does the Intermediate Value Theorem tell us?)

4. With notation as in Theorem 16, verify that \( f(x) = (x - x_0)g(x) + b_0 \).

5. Suppose \( f : \mathbb{R} \to \mathbb{R} \) is an odd differentiable function. (Recall that \( f \) is odd when \( f(-x) = -f(x) \).)
   
   (a) Show that \( f(0) = 0 \).
   
   (b) Show that \( f' \) is an even function (i.e., that \( f(-x) = f(x) \)).

6. (Here we show that Newton’s method need not converge for every \( f \) and every choice of \( p_0 \).) Suppose \( f \) is an odd differentiable function and that \( p_0 \in \mathbb{R} \) satisfies \( \frac{f(p_0)}{f'(p_0)} = 2p_0 \).
(a) Show that Newton’s method applied to \( f \) with initial approximation \( p_0 \) will not converge. (What is \( G(p_0) \)? \( G(G(p_0)) \)? Exercise 5 may be helpful.)

(b) Find an example of a function \( f \) satisfying the conditions above.

7. Use the fact that the third degree Maclaurin approximation of \( \sin x \) is \( x - \frac{x^3}{6} \) to estimate the solution of \( \sin x = 1 - x^2 \) by

(a) Cardano’s method.

(b) Newton’s method.

8. Let \( f \) be a polynomial of degree \( n \) with all coefficients being nonzero, and let \( x_0 \in \mathbb{R} \). Show that evaluating \( f(x_0) \) by using the formula for \( f \) requires \( \frac{n(n+1)}{2} \) multiplications and \( n \) additions.

9. Consider the following fixed-point theorem: Suppose that \( G : \mathbb{R} \to \mathbb{R} \) is continuous on \( [a, b] \), differentiable on \((a, b)\), and there exists a positive constant \( k < 1 \) with \( |G'(x)| \leq k \) for all \( x \in (a, b) \). Then for any number \( p_0 \) in \([a, b]\) the sequence \( \{p_n\} = \{G^n(p_0)\} \) converges to the (unique) fixed point of \( G \) in \([a, b]\).

(a) Prove the fixed-point theorem.

(b) Suppose \( f \) is continuous on \([a, b]\), differentiable on \((a, b)\), and \( G(x) = x - \frac{f(x)}{f'(x)} \) satisfies the hypotheses of the fixed-point theorem on \([a, b]\). Deduce conditions on \( f \) and \( f' \) that will ensure the success of Newton’s method on \([a, b]\).
CHAPTER 13

Measurement

Let’s Go 1. Put the objects in Figure 1 in order from largest to smallest. Justify your conclusion.

1. What Is Measurement?

1.1. Constructing a definition. Many of our early experiences in school and in life involve measuring, and the concept of measurement lies behind Let’s Go 1, so let’s begin by considering what it means to measure.
Your Turn 1.

(a) Construct a definition of what it means “to measure.”

(b) List at least ten things that one learns to measure in school. (You might consider things that arise in mathematics, but also things that are studied in other subjects, such as science.)

(c) Check to see if the definition you constructed in part (a) works for all items on the list you generated in part (b). If not, revise your definition accordingly.

Any serious attempt to answer Let’s Go 1 must confront the ambiguity in the notion of “size” of the objects sketched there. What exactly should we measure to gauge their “size”? We realize that two different entities are involved in every measurement activity: an object and an attribute.

Often a child’s early experiences involve measurements of themselves: age, height, weight, or body temperature. In this case, the object that is being measured is the child; age, height, weight, and temperature are measurable attributes. A rectangle is a geometric object whose size can be measured; its measurable attributes include lengths of the sides, perimeter, and area.

Your Turn 2. Identify each item on the list you created in Your Turn 1(b) as an object or an attribute.

One way to measure is to use direct comparison. For instance, two children can quickly determine which child is taller by standing next to each other and directly comparing their heights. Two apples can be placed on a balance scale, and the heavier apple can be identified visually as it falls while the lighter apple rises. Several children can run a race and the fastest child easily identified by observing which child first reaches the goal or crosses the finish line. Measuring by direct comparison is useful and helps establish order, but doesn’t quantify the attribute or produce a measure. In order to quantify the measurement, we need to identify appropriate units of measure.

Your Turn 3.

(a) For each attribute you listed in Your Turn 1(b), identify several units that can be used to measure that attribute.

(b) Discuss what makes a unit “appropriate.”
(c) Suppose you are measuring the length of your bicycle. Identify three units that would be appropriate to use. Then identify three units that would be inappropriate and explain why.

Typically, measurement is described as a three-step process:

(1) Identify the attribute to be measured.
(2) Choose an appropriate unit of measure for the attribute.
(3) Compare the attribute with the unit by counting the number of units needed to represent the attribute.

Thus, a measurement consists of two key pieces of information: a number and a unit.

Measurement is understood quite differently in abstract mathematics (on the one hand) and in scientific fields and everyday life (on the other hand). In mathematics, we can consider a right triangle with side lengths 3, 4, and 5 (in whatever units of length we choose); we know that such a triangle exists, since $3^2 + 4^2 = 5^2$, satisfying the Pythagorean Theorem. We happily discuss this triangle, sometimes with the aid of a rough sketch on paper or blackboard, untroubled whether the side lengths in our sketch are exactly 3, 4, and 5 units long, or whether the angle we’ve drawn is exactly a right angle, or even whether the sides are exactly straight. In the real world, the situation is very different. Instead of dealing abstractly with a segment that is 3 units long, in the real world, we have to deal with physical objects, and the physical process of measuring them. When we measure segments, it is rare to find a segment whose length seems to be exactly 3 cm. Indeed, scientists don’t even pretend to know the exact measurements of objects; to a scientist measuring a line segment, the length might be described as $3.13 \pm 0.05$ cm. This “measurement” shows confidence that the length is between 3.08 and 3.18 cm, but uncertainly about the location of the segment’s endpoints (fuzzy objects, if drawn with a pencil, and undoubtedly falling between the tick marks on a ruler), along with an acknowledgement of the fallibility of the scientist’s eyes and fingers. As mathematicians, we consider it incorrect to replace a length of $\sqrt{2}$ with 1.41 since these numbers are not equal; in science or engineering, it may be highly appropriate to replace one number with another very close number, if the difference between them is dwarfed by the uncertainties involved in making the physical measurements. In the activities that
arise in this chapter, you should stay alert as to whether an exact answer or a reasonable approximation is appropriate to the setting.

1.2. Units.

Let’s Go 2. Consider the tasks given in Figures 2 and 3.

(a) In what ways are these two tasks similar?

(b) In what ways do these two tasks differ?
Figure 2. Figure for Let’s Go 2, Task 1
Once an attribute that is to be measured is identified, an appropriate choice of a unit of measure must be made. Often there are many units from which to choose, so one must determine the best unit of measure from several possible options. Let’s consider what makes a unit of measure appropriate, and examine the criteria used to decide which is the best unit of measure in a given situation.

The most basic requirement for an appropriate unit of measure is that the unit must possess the same attribute as what is being measured. If one is measuring volume, the unit of measure must itself possess the attribute of volume. If one is measuring time, the unit of measure must possess the attribute of time. Thus, the volume of a box cannot be measured in inches or in square meters or in seconds, as inches, square meters,
and seconds don’t possess the attribute of volume. Since volume measures space in three dimensions, the unit used to measure volume must also represent space in three-dimensions.\(^1\)

So in measuring the distance from New York to San Diego, a unit must be chosen that itself possesses the attribute of distance, or length. But there are many units of length from which to choose—inches, feet, miles, centimeters, and kilometers, to name a few. What determines the choice of unit? We know that the distance between cities typically is represented in miles—why are miles the usual unit choice? Though this distance could be given in inches or feet, the number of inches or feet would be large and perhaps hard to envision or comprehend. Use of a smaller unit results in a larger number of units, while use of a larger unit requires a smaller number of units. Choosing a unit that produces a measure where the number of units is neither too big nor too small is of practical use and contributes to our definition of what makes a unit of measure appropriate.

Suppose that you consult a map of Europe to identify the distance between Paris and Rome. Knowing that the metric system of measurement is used in Europe, you should not be surprised to find that this distance is not given in miles, but in kilometers. Thus, another aspect of what makes a unit of measure appropriate may depend on history, experience, or custom. Depending on where he lives, what his customs are, and how much land he owns, one farmer may describe the area of his field in square feet, another in square miles, another in square kilometers, and yet another in acres.

The central concept underlying the tasks presented in Let’s Go 2 is the relationship between the size of the unit and the number of units. In both cases, a given quantity is to be measured. There are several choices available for the size of the unit—in the first problem the units are paces of different lengths and in the second problem the units are balls of different sizes. If one chooses a large unit, a small number of units is needed to represent the quantity. If one decides to use a smaller unit to represent the same quantity, more of these smaller units will be needed. Recognizing this inverse relationship between the size of the unit and the number of units is critical to understanding the concept of measurement and the procedures commonly used to perform unit conversions.

\(^{1}\)A local pizza company plays off this critical component of measurement when it advertises that a deep-dish pizza is “over three pounds deep!” On a more serious note: if you have studied relativity in a physics class, you know know that what makes the subject so mind-bending is that combined spacial-and-time separation can be regarded as a single attribute.
Your Turn 4. Suppose that you are a teacher and that you overhear two students in your class discussing how to represent 1200 minutes in seconds. One student quickly says, “Well, there are 60 seconds in a minute.” The other student says, “Yeah, so we either multiply or divide by 60. I wonder which one works? I’ll multiply by 60 and you divide.” The students successfully perform the computations, but are still uncertain about which answer was “right.”

(a) Which computation correctly represents 1200 minutes in seconds: 1200 × 60 or 1200 ÷ 60?

(b) Discuss how understanding the inverse relationship between the number of units and the size of the unit helps one determine whether to multiply or divide by 60 when converting 1200 minutes to seconds.

(c) Use this method to convert 30 cm to inches, and 16 in to centimeters. (One inch equals 2.54 centimeters.)

Your Turn 5. Another method used to perform unit conversions is commonly referred to as the factor/label method. The following example illustrates how to use the factor/label method to represent 1200 minutes in seconds:

\[
1200 \text{ minutes} \times \frac{60 \text{ seconds}}{1 \text{ minute}} = 72000 \text{ seconds}
\]

(a) Explain why the factor/label method works.

(b) Use the factor/label method to convert 1200 minutes to hours.

(c) Use this method to convert 30 cm to inches and 16 in to centimeters.

Your Turn 6. Yet another strategy for unit conversion involves proportional reasoning. For example, one can argue that since one hour equals 60 minutes, it follows that five hours equals 5 × (60 minutes) = 300 minutes.

(a) Use proportional reasoning to convert 1200 minutes to hours. (Is this as easy as converting hours to minutes?)

(b) Can you use this method to convert 30 cm to inches and 16 inches to cm?
Your Turn 7. Compare and contrast the mathematical prerequisites of these methods. Which method would you recommend? Does your answer depend on the age of the student?

1.3. Exercises.

Measurable attributes

1. List some measurable attributes of a circle. How do the measurable attributes of a circle compare to the measurable attributes of an ellipse? Of a square? Of a parallelogram?

2. What are some measurable attributes of a cube? Of a cylinder? Of a cone? Of a sphere?

3. For regions in the plane, give as many different attributes as you can that might reasonably be called the “size” of the region. Discuss why each of these notions of “size” is important in real-world settings.

4. We compare competing notions of the “size” of a triangle.

(a) Draw two triangles, such that the area of the second is larger than the area of the first, but the perimeter of the second is smaller than the perimeter of the first. How is this exercise related to Let’s Go 1? (In this problem, you are asked physically to draw triangles on paper, then accurately measure perimeter and area with aid of a ruler. Be certain to include the units of measurement. As you formulate your strategy, consider qualitatively what sort of triangle has large (or small) area relative to its perimeter.)

(b) Define two triangles, such that the area of the second is larger than the area of the first, but the perimeter of the second is smaller than the perimeter of the first. (In this problem, you are asked to define the two problems abstractly, by choosing the side lengths of the two triangles. For what sorts of triangles is it easy to calculate the area from the side lengths?)

5. Suppose that at the end of the season a basketball coach is asked to identify the player who has shown the most improvement during the season.

(a) In what ways could this be considered a measurement task? (Your answer should make use of the definition of measurement.)
(b) List some attributes that the coach might measure in making his/her decision, along with units appropriate to the attribute.

(c) How would the coach determine improvement, as opposed to performance?

6. Is it more challenging to measure a pyramid than a prism? Discuss, making clear what attributes you are measuring, and whether you are determining measurements physically or through calculation.

7. The most natural way to measure money is by its usual “monetary value,” but there are other reasonable measures. For example, the measure of a collection of coins could be simply the number of coins in the collection (ignoring the monetary value).

In this spirit, we define a function $f : \mathbb{N} \to \mathbb{N}$, by the rule $f(n) =$ the smallest number of coins that can be collected together to make a total of $n$ cents, using pennies, nickels, dimes, and/or quarters. For example, $f(9) = 5$, since one can make nine cents most efficiently with a nickel and four pennies (a total of five coins).

(a) Compute $f(10)$, $f(30)$, $f(39)$, and $f(100)$.

(b) Is $f$ injective? Explain.

(c) Is $f$ surjective? Explain.

Units and unit conversion

8. Consider the line segment in Figure 4.

(a) Give reasonably accurate measurements of the length of the segment in centimeters, inches, feet, miles, nanometers, and light-years.

b) In completing (a),

(i) How many times did you physically measure the segment?

(ii) How did you obtain the “measurements” in the units for which you did not perform a physical measurement?

(iii) Which of the units is reasonable for expressing the length of the segment?

(iv) Did you use a decimal for each answer, or did you use fractions? Did you ever choose to use scientific notation (an expression like $7.8306 \times 10^{-17}$ or $3.52 \times 10^{40}$)?
(v) In the answers you obtained through calculation, how many significant figures did you choose to include?

**Figure 4. Figure for Exercise 8**

9. Suppose that you are building a new garden with dimensions 20 feet by 25 feet and are trying to decide how much topsoil to buy. A truckload of dirt holds approximately 10 cubic yards of topsoil. How deep will the topsoil from one truckload be after you have distributed it evenly over your new garden? What choice of unit is reasonable for expressing your answer?

10. Use two different methods to convert 35 km/sec to miles per hour.

11. Temperature conversions.

   (a) Construct the relationship between the Fahrenheit and Celsius temperature scales, using that water boils at 212 degrees Fahrenheit and at 100 degrees Celsius, and water freezes at 32 degrees Fahrenheit and freezes at 0 degrees Celsius.

   (b) The high temperature for the day is forecast to be 35° Celsius in Rome. Using your answer to (a), convert this temperature to a temperature in the Fahrenheit scale.

   (c) Compare the method one uses to convert temperature from Celsius to Fahrenheit with the methods used to convert inches to feet. How does the method for performing temperature conversion differ from the methods used to convert inches to feet?

12. Suppose a rectangle has side lengths of 8 cm and 10 cm.

   (a) What is the area of the rectangle when measured in 1-cm square units?

   (b) What is the area of the rectangle when measured in 1-mm square units?
(c) Beginning with a sketch of an 8 cm × 10 cm rectangle, how can you modify the sketch to illustrate your calculations in (a) and (b)? (How do “square centimeters” and “square millimeters” appear naturally in this problem?)

13. Mrs. D. draws a large right triangle for her students, and shows that the two legs are 3 and 4 feet long, and the hypotenuse is 5 feet long. She computes that the area is 6 square feet. As an aside, she remarks that using inches, the side lengths would be 36, 48, and 60 inches.

Jimmie wonders whether the area of the triangle would be 72 square inches since $12 \times 6 = 72$ (but he’s afraid to ask). Is he correct? Explain. (Does your explanation feature a calculation or a drawing?)

14. The next day, Mrs. D. discusses square inches, so Jimmie gets his question answered. But he starts to wonder...If we use square inches for the area of squares, why don’t we use circular inches for the areas of circles?

(a) What would be a reasonable definition of “circular inches”?

(b) What makes rectangles more useful than circles in defining and computing areas?

2. Measuring Geometric Objects

Let’s Go 3. Consider the following classroom episode: A student is given a ruler and asked to determine the area of the right triangle in Figure 5. She recalls learning that the area of a triangle is half the area of a rectangle. She is confident that the area of a rectangle can be found by multiplying the rectangle’s length times the rectangle’s width, so she concludes that the area of the triangle must be half the length times the width. But then she looks at the triangle and says to herself, “But what’s the length of the triangle? And what’s its width?” She then concludes that she must not have remembered the formulas correctly and gives up.

(a) In what ways is the student’s thinking correct?

(b) In what ways is the student’s thinking confused or incorrect? What appears to be the source of the confusion?

(c) Use a ruler to determine the area of the given triangle.
2. MEASURING GEOMETRIC OBJECTS

2.1. Measuring One-dimensional and Two-dimensional Geometric Objects. A line, a ray, and a line segment are geometric objects with “one dimension.” The measurable attribute of a one-dimensional geometric object is typically called length. The length of a line or a ray is infinite, but the length of a line segment is finite and can be measured using units that also possess the attribute of length. Standard units used to measure length include inches, feet, miles, centimeters, and kilometers.

Your Turn 8. Mathematicians also consider a curve, such as the graph of \( y = x^2 \) for \( 0 \leq x \leq 3 \), to be a one-dimensional object. What is a reasonable way to define the length of a curve? (You might have an intuitive picture in mind, or perhaps a rigorous definition from a math course.)

Several terms are used when referring to the length of a line segment, particularly when the line segment is a component of a two or three-dimensional geometric object.

Your Turn 9. How would you personally use the terms “length,” “width,” “base,” or “height” in describing the rectangles in Figure 6? What principles seem to affect your choice? Are there any other words you would choose to use?
Likewise, consider the rectangular prism shown in Figure 7. Many would choose to call the lengths of the two perpendicular sides of the rectangle that comprise the base (sides $a$ and $b$) the “length” and the “width” (in no particular order), and would use the term “height” for the length of the third (vertical) side. Another way to describe the rectangular prism is to refer to the measures of side $b$ as the “length,” side $c$ as the “height”, and side $a$ as the “depth.”
Your Turn 10. On what principles do these choices of words rely? How do these notions change as the figure is rotated?

The use of these various terms to describe linear measure may have confused the student whose thinking was presented in Let’s Go 3. She readily used the terms “length” and “width” when recalling a familiar formula for the area of a rectangle, but couldn’t reconcile how these terms related to the linear measures of a triangle or the formula for the area of a triangle, in which linear measure was referred to as the “base”\(^2\) and the “height.” The fact that triangles and rectangles are so often drawn with one side horizontal only adds to the confusion.

Measuring “two-dimensional” objects is a major component of the school mathematics curriculum from elementary grades through calculus. There are several attributes of two-dimensional geometric objects whose measures are important—both in school mathematics and in everyday life. One measure is the length of the boundary of a two-dimensional object; this boundary actually is a one-dimensional object. For most two-dimensional objects the measure of the boundary is called the perimeter; for a circle, this boundary measure is called the circumference. Since the boundary of two-dimensional objects is comprised of lines or curves, the appropriate units used to measure the boundary are linear units.

The measure of the space inside the boundary of a two-dimensional figure, the area, is another important measure. To find the area of a two-dimensional figure, we think of filling or covering the space inside the figure with units that also contain the attribute of area. The goal is to use exactly enough appropriate units to cover the space with no overlap of units and no space left uncovered. Thus, the area of a floor, in square feet, should be imagined as the number of 1 foot × 1 foot tiles that would be needed to cover the floor.

Your Turn 11.

(a) Suppose \(a\) and \(b\) are positive integers. Explain why the area of an \(a \times b\) rectangle is \(ab\).

(b) Jimmie is trying to figure out the area of a rectangle that is 3.5 feet by 2 feet. Mrs. D. says the area is seven square feet, but it seems to him that you would need eight tiles to cover the region.

What is his misunderstanding?

---

\(^2\)Regrettably, “base” commonly is used to describe both a side of a triangle (a segment) and the length of the side (a number).
(c) In general, the area of an $a \times b$ rectangle is $ab$, where $a$ and $b$ are any positive real numbers. Show that to cover a $\frac{4}{3} \times \frac{4}{5}$ rectangle with tiles requires exactly $\frac{32}{15}$ tiles.

**Your Turn 12.** We return to the topic of Let’s Go 3, considering triangles that don’t contain a right angle.

(a) Draw a scalene acute triangle, and identify one of its sides as the “base.” Then draw a rectangle, having the same “base” and the same “height” as the triangle. Explain why the triangle has half the area of the rectangle.

(b) Repeat (a) for a scalene obtuse triangle. This time, draw three sketches, with each side of the triangle taking its turn as the “base.” (What is qualitatively different about the case of an obtuse triangle?)

**2.2. Measuring Three-dimensional Geometric Objects.**

**Let’s Go 4.** First, visualize some familiar three-dimensional geometric figures. Turn the figures around in your mind so that you can see them from several different perspectives.

(a) Choose two three-dimensional figures by naming them and/or sketching them.

(b) Using only words, describe each of these figures to another person who cannot see them. Do not use the figure’s name in your description.

(c) Have the other person name or describe the three-dimensional figure that you had described to them. Compare.

One way to describe a three-dimensional object is to focus on the surfaces that comprise its boundary; in essence this is what one needs to do in order to draw the three-dimensional object on a two-dimensional piece of paper. For instance, a right-circular cylinder has two bases that are disks; the body of the cylinder is a “rolled” rectangle. The boundary of a pyramid contains several triangles and a base that can be a triangle, a square, or essentially any polygon. A rectangular prism (box) has six sides - three pairs of congruent rectangles with the two members of a pair located opposite one another. When drawing the three-dimensional figure, the two-dimensional shapes that make up the boundary must be distorted a bit in
order to add perspective. And when viewing a three-dimensional figure drawn on a two-dimensional surface, one must interpret these distortions - for example, all six sides of the rectangular prism shown in Figure 7 are rectangles, though some of these rectangles literally look like parallelograms.

It is no surprise, then, that the measure of the surface of a three-dimensional figure is related to the measures of the two-dimensional figures that form its boundary. The \textit{surface area} of a three-dimensional figure is simply the sum of the areas of the two-dimensional figures that form its boundary, and is measured in square units. The measure of the space inside the boundary of a three-dimensional figure, the \textit{volume}, is another important measure. To find the volume of a three-dimensional figure, we can envision filling the space inside the figure with units that also contain the attribute of volume. The goal is to use exactly enough units to fill the space with no overlap of units and no space left uncovered. Often it is useful to think of filling the space “from the bottom up.” For instance, in the rectangular prism \((3 \times 4 \times 5)\) shown in Figure 7, one can think of filling the bottom layer with 12 cubes, then continue filling the space with layers until there are 5 layers of cubes. Thus, the formula for the volume of a rectangular prism can be constructed as

\[
V = \text{length} \times \text{width} \times \text{height},
\]

or

\[
V = \text{area of base} \times \text{height}.
\]

Similar reasoning can be used in developing the formula for the volume of a right-circular cylinder (see Exercises).

\section*{2.3. Exercises.}

1. \textit{Distance between two points}.

   \begin{itemize}
   \item[(a)] Determine the length of the line segment with endpoints \((-2, 1)\) and \((7, 4)\) by plotting the points, constructing a right triangle, and using the Pythagorean theorem.
   \item[(b)] Determine the length of the line segment with endpoints \((x_1, y_1)\) and \((x_2, y_2)\) by plotting the points, constructing a right triangle, and using the Pythagorean theorem.
   \end{itemize}

2. \textit{Altitudes of a triangle}.
(a) Draw an acute scalene triangle (all angle measures less than 90°, and no sides of equal length).

Carefully draw the three altitudes of the triangle. (If you draw carefully, you will illustrate an important theorem from geometry.)

(b) Repeat with a scalene right triangle.

(c) Repeat with a scalene obtuse triangle.

3. Categories of triangles. Using what you know about the measures of sides and angles of triangles, construct an example for each of the following type of triangle. Identify the length of the sides of each triangle in your example. If it is not possible to construct an example, explain why.

   (a) a right, isosceles triangle
   (b) a right, equilateral triangle
   (c) an obtuse, isosceles triangle

(Did you draw triangles and measure with a ruler, or create the examples abstractly?)

4. Geometric definitions via measurement concepts. Give definitions of a circle, rectangle, a square, a rhombus, a parallelogram, and a trapezoid. For which of these objects can you give a definition that relies on a measurement concept (e.g. side length or angle measure)?

5. Investigating Scale Changes in Two-dimensions.

   (a) Construct two rectangles that are similar.
   (b) Determine the ratio of the perimeter of the larger rectangle to the perimeter of the smaller rectangle.
   (c) Determine the ratio of the area of the larger rectangle to the area of the smaller rectangle.
   (d) Compare the ratios of the linear factors (e.g., the sides lengths) of the two rectangles, the perimeter of the two rectangles, and the area of the two rectangles.
   (e) Repeat parts (a)–(d) for a pair of similar triangles and a pair of circles.
   (f) Based on your work above, form a conjecture about the ratio of the perimeters and the ratio of the areas of two-dimensional similar figures.
   (g) Prove your conjecture in the case of two similar rectangles.
6. **Investigating Scale Changes in Three-dimensions.**

(a) Construct two rectangular prisms that are similar.

(b) Based on your findings in Exercise 5, predict the ratio of the surface area of the larger prism to the surface area of the smaller prism.

(c) Determine the ratio of the surface area of the larger prism to the surface area of the smaller prism.

(d) Predict the ratio of the volume of the larger prism to the volume of the smaller prism.

(e) Determine the ratio of the volume of the larger prism to the volume of the smaller prism.

(f) Compare the ratios of the linear factors (e.g., the sides lengths) of the two rectangular prisms, the surface area of the two prisms, and the volume of the two prisms.

(g) Prove that the relationships identified in part (f) between the ratios of the linear factors, the surface area, and the volume exist for any two rectangular prisms.

7. **Lengths of smooth curves.** Consider the curve \( y = x^2 \) from the points \((0, 0)\) to \((3, 9)\).

(a) Approximate the length of this curve using a single line segment with endpoints \((0, 0)\) and \((3, 9)\). Is this approximation greater than or less than the actual length of the curve?

(b) Improve your initial approximation by using three line segments using the points where \(x = 0, x = 1, x = 2,\) and \(x = 3\).

(c) Find the length of the curve using the arc length formula from calculus.

(d) How does the formula for arc length relate to the approximations you found in parts (a) and (b)?

8. **Circle measurement.** An advertisement offers “Family Size Pizza for the price of a Medium Pizza” and claims that this deal gives you “75% more pizza FREE!” If a family size pizza is 16 inches in diameter and a medium pizza is 12 inches in diameter, verify that this claim is true.

9. **Subdividing a rectangle.** Suppose that when teaching about fractions, a teacher asks students to divide a rectangle into fourths. His students construct the figures show here, each of which contains four congruent rectangles. (See Figure 8.) One student draws the diagonals of the rectangle and claims that she has shown “fourths.” (See Figure 9.) Other students disagree that this illustrates “fourths,” as the four triangles formed are not congruent.
Explain why Figure 9 indeed illustrates fourths, even though the four triangles are not congruent.

10. *Interior angle measure of regular polygons.*

(a) Find the measure of one of the angles in a regular hexagon.

(b) Find the measure of one of the angles in a regular octagon.
(c) What is the limit of the measure of one of the angles in a regular polygon as the number of sides approaches infinity?

11. Ambiguous terminology. Consider how the term base is used in the following two formulas.

\[
\text{Area of a parallelogram} = \text{base} \times \text{height} \\
\text{Volume of a pyramid} = \frac{1}{3} \times \text{base} \times \text{height}
\]

(a) Describe how the use of the term base is similar in the two formulas.

(b) Describe how the use of the term base is different in the two formulas.

(c) Explain the relationship between these two forms of the volume formula for a right circular cylinder:

\[ V = Bh \text{ and } V = \pi r^2 h. \]

12. Optimization. Suppose that you work for a packaging company and your job is to create the container with the greatest volume from a piece of cardboard that measures 8 feet by 10 feet.

(a) Create four different three-dimensional containers and compare their volumes.

(b) Determine which of your containers has the maximum volume.

(c) Determine how much cardboard was wasted/unused when constructing each container.

(d) Discuss the relationship between the volumes of the containers and the waste generated from making the containers.

13. Optimization. Using a spreadsheet,

(a) List all rectangles with perimeter 72 units having integer dimensions.

(b) Compute the area of each rectangle.

(c) Identify the rectangle with maximum area from the spreadsheet.

(d) If possible, find a rectangle with non-integer dimensions whose area is greater than the rectangle identified in (c).

(e) Construct a real-world situation in which finding the maximum area for a given perimeter is necessary and useful.
14. Optimization. Repeat parts (a)–(d) of Exercise 13 using rectangles with perimeter of 75 units instead of 72 units. Discuss how your response to part (d) in Exercise 14 compares to your response in Exercise 13.

15. Optimization. Using a spreadsheet,

(a) List all rectangles with area 72 square units having integer dimensions.
(b) Compute the perimeter of each rectangle.
(c) Identify the rectangle with minimum perimeter from the spreadsheet.
(d) If possible, find a rectangle with non-integer dimensions whose perimeter is less than the rectangle identified in (c).
(e) Construct a real-world situation in which finding the minimum perimeter for a given area is necessary and useful.

16. Optimization. Construct all possible rectangular prisms with volume of 60 cubic units whose side lengths are integers. From your list, identify the prism with maximum surface area and the prism with minimum surface area.

17. Optimization. Construct several regions in the plane with perimeter 60 inches and compare their areas. Continue constructing figures until you believe that you have found the figure with maximum area. Explain why you believe this figure has maximum area for the given perimeter of 60 inches.

18. Optimization. The U.S. Post Office uses a rate system based on linear measure in determining the mailing cost for packages. Consider the following statement: Parcel Post pieces measuring over 84 inches in combined length and girth, but not more than 108 inches in combined length and girth, and weighing less than 15 pounds are mailable at the rate equal to that of a 15-pound parcel for the zone to which the parcel is addressed.

(a) Explain the meaning of the statement “combined length and girth.” It may be helpful to construct a diagram as part of your explanation.
(b) Represent the statement “combined length and girth” algebraically.
(c) Construct four different packages (i.e., rectangular prisms) with “combined length and girth” of 84 inches.

(d) Compute and compare the volumes of the four packages you constructed above.

(e) Construct another package with combined length and girth of 84 inches whose volume is greater than the four packages you constructed above.

(f) Describe how you could find the dimensions of the package with combined length and girth of 84 inches with maximum volume.

19. *Diagnosing student misconceptions.* Suppose that you are teaching a middle school class. Your students are using protractors to measure angles. Suddenly one of your students asks, “How do I find the area of an angle?” The question suggests that the student has misconceptions related to understanding area or understanding angle.

(a) Based on the question, identify three misconceptions about area and/or angles that this student might have.

(b) Describe how you would respond to the student’s question.

(c) What do you hope that this student would understand about area and about angles as a result of this classroom interaction?

20. *Volume of a cone.* Explain why the volume of a cone equals a third of the product of the height and the area of the base. (Your calculus book may be helpful.)

21. *Lateral surface area of a cone.* A cone is illustrated in Figure 10.

(a) Find, and justify, a formula for the lateral surface area of the cone. You may use any of the variables \( R, S, H, \) and \( \theta. \)

(b) Suppose a physical cone is measured and it determined that \( R = 2.37 \) ft and \( S = 4.15 \) ft. What is the lateral surface area? *(You will need to exploit relationships among the variables.)*
22. *On the definition of π.* Derive the formula area = $\pi R^2$ for the area of a circular disk of radius $R$. (Approximate the circle by an inscribed regular polygon. You are free to use the fact that the circumference of the circle is $2\pi R$, which is tantamount to the definition of $\pi$.)

23. *A survey of measurement techniques from calculus.* Show that the volume of a ball of radius $R$ is $\frac{4}{3}\pi R^3$. Use each of the following techniques:

   (a) By using the *disk method* from calculus.

   (b) By using the *shell method* from calculus.

   (c) By assuming that the volume of a spherical shell of radius $r$ is approximately $4\pi r^2$ times the thickness of the shell. (Why is this a reasonable assumption?)

   (d) By a double integral over a disk.

   (e) By a triple integral in spherical coordinates.

24. *How independent are area and perimeter?* For certain geometric figures (e.g., circles, squares, and equilateral triangles), area and perimeter convey the same information. This happens because a single number (radius of a circle, side length of a square or equilateral triangle) determines the figure up to congruence.
(a) Find an equation that relates the area and the perimeter (circumference) of a circle. (Express each in terms of the radius $r$, then use algebra to eliminate $r$.)

(b) Find an equation that relates the area and perimeter of a square.

(c) Find an equation that relates the area and perimeter of an equilateral triangle.

25. Area, perimeter, and congruence of rectangles. Let $P$ and $A$ be positive numbers.

(a) Suppose a rectangle has side lengths $x$ and $y$ (with $x \geq y$), and perimeter $P$ and $A$. Using algebra, solve for $x$ and $y$ in terms of $P$ and $A$. This shows that perimeter and area together determine at most one rectangle, up to congruence.

(b) Examine your solution to (a). How are $P$ and $A$ related if $x = y$?

(c) Examine your solution to (a). Can $x$ and $y$ fail to be real numbers? Conclude that for any rectangle, $P$ and $A$ satisfy a certain inequality (what is it?).

26. Area, perimeter, and congruence of right triangles. Let $P$ and $A$ be positive numbers.

(a) Suppose a right triangle has legs $x$ and $y$ (with $x \geq y$), and perimeter $P$ and $A$. Using algebra, solve for $x$ and $y$ in terms of $P$ and $A$. This shows that perimeter and area together determine at most one right triangle, up to congruence. The solution is similar to Exercise 25 but the algebra is harder.

(b) Examine your solution to (a). How are $P$ and $A$ related if $x = y$?

(c) Examine your solution to (a). Can $x$ and $y$ fail to be real numbers? Conclude that for any right triangle, $P$ and $A$ satisfy a certain inequality (what is it?).

27. Continuation of Exercise 26. If two triangles (not necessarily right triangles) have the same area and the same perimeter, must they be congruent? Give a proof or a counterexample.

28. An area-perimeter inequality for triangles. It turns out that given any triangle, if the area is $A$ and the perimeter is $P$, then $A$ and $P$ satisfy a certain inequality. Make a conjecture about this inequality. If you can, prove your conjecture.
3. Exploring Areas of Triangles

Let’s Go 5. Perhaps the most familiar method for computing the area of a triangle is by the formula “area is half the product of the base and the height.” Using this formula, compute the area of the triangle in the coordinate plane with vertices (2, 1), (6, 5), and (4, 9).

Do you see any easier method to compute the area of this triangle?

In upper elementary grades or early middle school, students learn to determine the area of a triangle. Ideally, students commit the formula for the area of a triangle to memory (“area of a triangle is half of the base times the height”), and also understand how this formula is related to the area formulas of rectangles and parallelograms. In mathematics textbooks, finding the area of a triangle is frequently a fairly routine task (once the formula is known), since typically the base and the height of a triangle are given, or can be easily determined. In the real world, the base and the height can usually be measured. Thus, at first glance, we may dismiss the problem of “finding the area of a triangle” as a straightforward task easily handled by pre-algebra students.

However, solving the problem posed in Let’s Go 5 might have caused you to reconsider the alleged simplicity of finding the area of a triangle. What makes finding the area of this triangle so challenging? The triangle wasn’t defined in terms of the lengths of its base and height. The three vertices of the triangle were stated as ordered pairs in the first quadrant of the coordinate plane; from this information, the lengths of a base and a height must be found in order to use the familiar area formula. So first we must identify one side as the base and determine its length. Then the height of the triangle can be found, and the familiar formula for determining the area of a triangle can be applied. This solution process depends on successful execution of several algebraic procedures that typically are learned only in high school mathematics (e.g. applying the distance formula, finding the equation of a line, finding the slope and equation of a perpendicular line to a given line, finding the intersection point of two lines). Were these the strategies you employed?

The problem teaches us an important lesson: a formula is only useful if one has the “ingredients” of the formula. It also raises an interesting question: are there other formulas for the area of a triangle that are
better suited to some problems (meaning, they define the area in terms of some data other than the base and the height)? We now see that this is the case.

3.1. Side Lengths and Angles. One method for finding the area of a triangle depends on knowing the lengths of two sides of a triangle and the measure of the angle between them.

Proposition 1. If a triangle has side lengths of $A$, $B$, and $C$, and opposite angles with measures $\alpha$, $\beta$, and $\gamma$ as in Figure 11, then the area of the triangle is given by

\[
\text{area} = \frac{1}{2} BC \sin \alpha = \frac{1}{2} AC \sin \beta = \frac{1}{2} AB \sin \gamma.
\]

Figure 11. Angles and sides

Proof. From Figure 11 we see that if we take $B$ to be the base of the triangle, then the altitude is $A \sin \gamma$. Since the area is half of the base times the height, we have that the area is $\frac{1}{2}AB \sin \gamma$. The other equations are provided similarly.

Your Turn 13. Use the formula above to determine the area of the triangle given in the Let’s Go 5. Compare this answer with your previous answer. (Hint: Use the Law of Cosines to find the measure of an angle.)
3.2. Heron’s Formula. *Heron’s Formula*, due to Heron of Alexandria, gives the area of an arbitrary triangle in terms of the lengths of the three sides\(^3\). The proof is elementary, requiring only the Law of Cosines and repeated use of “difference-of-two-squares” factoring.\(^4\)

**Theorem 2.** Heron’s Formula. Given a triangle with side lengths \(A\), \(B\), and \(C\), let \(S\) be half of the perimeter \((S = \frac{A+B+C}{2})\). The area of the triangle is given by

\[
\text{area} = \sqrt{S(S-A)(S-B)(S-C)}.
\]

**Proof.** First, by the Law of Cosines, we have \(C^2 = A^2 + B^2 - 2AB \cos \gamma\), so

\[
\cos \gamma = \frac{A^2 + B^2 - C^2}{2AB}.
\]

\(^3\)Heron of Alexandria lived sometime around A.D. 75. He was interested largely in applied mathematics problems, and in his *Dioptra* describes how one may use elementary geometry to build tunnels. However, he is best known for his work *Metria*, which contains the formula bearing his name.

\(^4\)This is not the method of Heron’s very clever geometric proof, which did not employ the Law of Cosines. The argument given above does have the virtue of being shorter.
Next, let’s rewrite the square of the right side of Heron’s Formula, without the use of the semiperimeter:

\[ S(S - A)(S - B)(S - C) = \left( \frac{A + B + C}{2} \right) \left( \frac{A + B + C}{2} - A \right) \left( \frac{A + B + C}{2} - B \right) \left( \frac{A + B + C}{2} - C \right) \]
\[ = \frac{(A + B + C)(A + B + C - 2A)(A + B + C - 2B)(A + B + C - 2C)}{16} \]
\[ = \frac{(A + B + C)(-A + B + C))((A - B + C)(A + B - C))}{16} \]

(now interchange the second and fourth factors)
\[ = \frac{(A + B + C)(A + B - C))((A - B + C)(-A + B + C))}{16} \]
\[ = \frac{(A + B)^2 - (C^2)}{16} \]
\[ = \frac{(A^2 + B^2 - C^2) + 2AB}{16} \]
\[ = \frac{(A^2 + B^2 - C^2)}{16} + \frac{2AB}{16} \]
\[ = \frac{(A^2 + B^2 - C^2)}{16} - (2AB)^2 \]
\[ = \left( \frac{(2AB)^2}{16} \right) \left( (A^2 + B^2 - C^2)^2 - (2AB)^2 \right) \]
\[ = \left( \frac{(2AB)^2}{16} \right) \left( \cos^2 \gamma - 1 \right) \]
\[ = \frac{(AB)^2}{4} \left( 1 - \cos^2 \gamma \right) \]
\[ = \left( \frac{AB \sin \gamma}{2} \right)^2 \]

Hence, the right side of Heron’s Formula is equal to \( \frac{AB \sin \gamma}{2} \), which by Proposition 1 is equal to the area of the triangle. \( \square \)

**Your Turn 14.** Use Heron’s Formula to determine the area of the triangle given in Let’s Go 5. Compare your answer to your previous answer.

### 3.3. Areas from coordinates.
We still haven’t found a solution to Let’s Go 5 that makes direct use of the data in the problem, namely, the coordinates of the vertices. In this section, we’ll show how to do exactly this. (Our discussion will focus on parallelograms, but will apply easily to triangles too.)

Let’s begin with an example: we take the parallelogram associated to the points \((3, 1)\) and \((2, 5)\). By this, we mean that these two points are opposite (not adjacent) vertices, and that the origin \((0, 0)\) is also a vertex. This forces \((3, 1) + (2, 5) = (5, 6)\) to be the fourth vertex. (See Figure 12.)
We will find the area of the parallelogram by transforming it to a rectangle of the same area, in two steps. First, we shear off the triangular region inside the parallelogram to the right of $x = 3$, and attach it to the left of the parallelogram. We obtain a new parallelogram (see Figure 13).

Here we need to find the $y$-intercept of the line that contains the top edge. Using the points $(2, 5)$ and $(5, 6)$ and point-slope form, we find that the $y$-intercept is $13/3$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure12.png}
\caption{Vertices of a parallelogram}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure13.png}
\caption{Shearing a parallelogram}
\end{figure}
Second, we shear off the top of the parallelogram, and attach it to the bottom (see Figure 14). This rectangle has base 3 and height 13/3, so its area, and the area of the original parallelogram, are both 13.

We now turn to the general case of a parallelogram determined by points \((a, b)\) and \((c, d)\) (see Figure 15).
Figure 16. Area of a parallelogram

If our figure is accurate, we can compute easily that the $y$-intercept of the line containing the top edge is $d - \frac{bc}{a}$, and then it is clear that the area is $ad - bc$. However, this approach is misleading, since the two points might lie in any quadrant of the coordinate plane. We take a somewhat different approach instead. The area of the parallelogram is equal to the length of the base (the segment joining $(0,0)$ and $(a,b)$) times the altitude (the segment that starts at $(c,d)$ and meets the base at right angles). (See Figure 16.)

The length of the base is $\sqrt{a^2 + b^2}$. For the altitude, we need to find the distance between the point $(c,d)$ and the line $bx - ay = 0$. By a well-known formula (see Chapter 2), that distance is $\frac{|0 - (bc - ad)|}{\sqrt{b^2 + (-a)^2}} = \frac{|ad - bc|}{\sqrt{a^2 + b^2}}$. We have proven:

**Theorem 3.** The area of the parallelogram determined by the points $(a,b)$ and $(c,d)$ is $|ad - bc|$. 

Note that $ad - bc$ is the determinant of the matrices \( \begin{pmatrix} a & b \\ c & d \end{pmatrix} \) and \( \begin{pmatrix} a & c \\ b & d \end{pmatrix} \). The fact that “determinants give change in area” is commonly taught in linear algebra, and is crucial for the change-of-variable formulas for two-dimensional integrals in multivariate calculus.

**Your Turn 15.** Use Theorem 3 to solve Let’s Go 5. (Hints: Recall that you must find the area of a triangle, not a parallelogram. Also, you will need to compensate for the fact that the triangle does not have $(0,0)$ as one of its vertices.)
3.4. Exercises.

1. This exercise uses Figure 17. Length is measured in centimeters.

   (a) Estimate the coordinates of the three points, up to tenths of centimeters. Then use this information to estimate the area of the triangle.
   (b) Use a ruler to estimate the lengths of the three sides, up to tenths of a centimeter. Use these measurements and Heron’s Formula to estimate the area of the triangle.
   (c) Use a ruler to estimate the lengths of the three altitudes of the triangles. Use the formula area = \( \frac{1}{2} \times \text{base} \times \text{height} \), compute three more estimates for the area of the triangle. (Note: Using a compass, you can accurately construct the altitudes.)

2. Choose a right triangle with integer side lengths.\(^5\) Find the area of the triangle in the following ways:
   (a) Using the formula area = \( \frac{1}{2} \times \text{base} \times \text{height} \).
   (b) Using Heron’s formula.

3. Find the area of a triangle with side lengths 3, 5, and 6.

\(^5\)Take your favorite Pythagorean triple.
4. Find the area of an equilateral triangle with side length 1 using Heron’s formula. Check your answer by finding the area by another method.

5. Find a formula for the area of a right triangle in terms of its side lengths. (Do you need Heron’s formula?)

6. Perhaps the most elementary method for solving Let’s Go 5 is by embedding the triangle in a rectangle whose sides are parallel to the coordinate axes.

(a) Plot the points (4, 9), (6, 5), and (2, 1). Draw the minimal rectangle with sides parallel to the coordinate axes, that includes the three points. Identify the coordinates of the vertices of the rectangle. By computing the area of the rectangle and several right triangles, solve Let’s Go 5.

(b) More generally, use this method to find the areas of the triangles in Figure 18. (Note the reappearance of determinants.)

7. Consider the triangle with vertices (1, 3), (7, 24), and (15, −1). Does this triangle have a right angle? Use a strategy that would be convincing and understandable to high school students.
8. Use the method of “shearing” to find the area of the parallelogram determined by \((-2, 1)\) and \((1, 3)\). Check your answer using the determinant formula.

9. There are several parallelograms that have \((0, 0), (4, 1),\) and \((2, 3)\) as vertices. Find all of them (by finding the fourth vertex). Sketch them. Are the parallelograms all congruent? Do they all have the same area? Discuss.

10. Find all parallelograms that have \((4, 7), (2, 8),\) and \((-1, 5)\) as vertices (along with a fourth vertex, not \((0, 0))\). Find the area of each.

11. Find the area of the triangle with vertices \((3, -2), (4, 5),\) and \((0, 0)\).

12. Find the area of the triangle with vertices \((2, 3), (4, 10),\) and \((5, 7)\).

13. Heron’s Formula give a formula for the area of a triangle in terms of its side lengths. Would you expect there to be a formula for the area of a parallelogram in terms of its side lengths? Give clear reasoning to support your answer.

14. If \(C = A + B\), what does Heron’s formula give for the area of the triangle? Is this the correct answer?

15. Suppose someone hands you a pentagon (five-sided polygon) made from a sheet of plexiglass. Assume that it is not a regular pentagon, meaning, it does not have equal side lengths. Give the easiest strategy you can to determine its area with good accuracy. Explain clearly how to implement your strategy. Be certain to identify any tool(s) you need, and specify which measurements you will make with them.

16. Construct a right triangle on the coordinate plane where none of the sides is parallel to a coordinate axis. Identify the three vertices of the triangle. Determine the area of the triangle in the following ways:

   (a) using the formula area = \(\frac{1}{2} \times \text{base} \times \text{height}\).

   (b) using Heron’s formula.

   (c) using a determinant.
17. Consider the lines \( y = 0, y = x, y = 1, \text{ and } y = x + b \). What must \( b \) be in order for the four lines to bound a parallelogram of area 1?

18. In this exercise, we take the first steps toward a geometric proof of Heron’s formula without using the Law of Cosines. Consider a triangle \( ABC \) with side lengths \( a, b, c \), and an inscribed circle of radius \( r \) centered at \( O \) (called the incenter of \( ABC \)). Let \( s \) be the semiperimeter of the triangle. Show that the area of \( ABC \) is \( rs \). (Hint: Draw segments from \( O \) to each point where the circle is tangent to the triangle. These segments, each of length \( r \), form the heights of three triangles which together make up all of \( ABC \).)

19. Given three lines with equations \( A_1x + B_1y = C_1, A_2x + B_2y = C_2, \text{ and } A_3x + B_3y = C_3 \), find the area of the triangle whose vertices are the (pairwise) intersection points of the three lines. (A computer algebra system will be helpful.)

4. Measuring Figures with Non-integer Dimension

Let’s Go 6. In Calculus, you learned a variety of approaches for measuring curved objects: formulas for the length of curves, areas of surfaces, and volumes of solids. Recall as many of these as you can. What information did you need to have about the curve/surface/solid to use the formula?

Let’s Go 6 reminds us of the power of calculus to measure many geometric objects. Note, though, that before we can use any of these formulas, we must clearly understand whether the object is a curve, a surface, or a solid. Essentially, we measure the dimension of the object before we perform the additional measurement of length, area, or volume. In this section, we consider some geometric objects whose “dimension” requires careful consideration.

4.1. The notion of dimension. Throughout most of the history of mathematics, dimension was measured in non-negative integer values, and defined in relation to “direction,” or more precisely, the number of independent directions in which one can move along the object.\(^6\)

If an object consists of a single point, then one can’t move at all within the object, so we declare the dimension of a single point to be zero.

\(^6\)A physicist might speak of the “degrees of freedom” in moving along the object. Mathematically, this notion of dimension finds its rigorous treatment in the notion of a smooth manifold.
A line (whether it is the number line, a line in the plane, or a line in space) is considered to be one-dimensional, since one can move forward or backward along it, but not in any perpendicular direction. The same principle applies to, say, a circle in the plane: one can move clockwise or counterclockwise along it, but not in any perpendicular direction. Thus a circle—or any curve—is deemed to be one-dimensional.

In contrast, a plane should be considered two-dimensional, since at any point on it, one can choose a direction of travel, and then a second direction of travel that is perpendicular to the first. Likewise, familiar surfaces like spheres are two-dimensional because we again can find two independent directions of travel from any point. (See Figure 19.)

**Figure 19. Directions in two-dimensional geometric figures**

Finally, solid regions of space are three-dimensional, since one can move within the solid in three mutually perpendicular directions (for example, parallel to the $x$, $y$, and $z$ axes).

**Your Turn** 16. *The concept of boundary is important in geometry.*

(a) The boundary of an object generally has dimension one less than the dimension of an original object. Check this principle, by describing the boundary of each of the following objects:

(i) *a solid rectangular prism.*

(ii) *a solid cone.*

(iii) *a solid sphere.*

(iv) *a solid circle.*

---

Thus, at each point, there are infinitely many directions of travel. We identify the dimension as two, because we can find two directions that are perpendicular to each other.
(v) a solid circle.
(vi) a line segment.

(b) What notion of boundary did you use in answering these questions? Try to formulate a formal definition of boundary.

(c) What is the boundary of a hollow sphere? A circle?


If we looked at a map in an highway Atlas, we would see that while the coast has a few sharp turns, it certainly looks like a curve (and thus, should have dimension one). If we wanted a closer look, we might look at a compilation of local maps. These would reveal many more twists and turns that were smoothed over when the highway map was created. If we had the pleasure of actually visiting the coast, we would see even more roughness: peninsulas that were ignored by the mapmakers, house-sized outcroppings jutting into the water, surrounded by boulders, pebbles, grains of sand...and then we reach a microscopic stage that might be even harder to understand.

This calls into question whether the coastline is truly a curve, and whether it is correct to deem it one-dimensional. Certainly the coastline can be approximated by a curve, as done coarsely by the mapmaker, more finely by a hiker trekking the coast, or even more accurately by a crawling insect. But it seems odd to discuss a “direction of travel” when one is constantly altering course around boulders, pebbles, sand grains, atoms...

Even were we to accept the coastline as a curve, we would have difficulty assigning a meaningful length to it. We could make an estimate from the highway atlas, but this estimate would go up considerably with the added detail from local maps, and it would go up enormously with each added detour around boulders or pebbles or sand grains. Perhaps it is not meaningful to assign a value to the length of the coastline.

Mathematicians have developed more subtle definitions\(^8\) of dimension than the one presented to Section 4.1, to more accurately capture the complexity of some geometric objects. We will restrict our discussion to a few mathematical objects that are self-similar, meaning, the object resembles itself when magnified. We will give a non-integer-valued definition of dimension that reflects the complexity of the objects.

\(^8\)These include Lebesgue Covering Dimension, Hausdorff Dimension, Inductive Dimension, and Minkowski Dimension.
Example 1. The Koch curve is constructed using an iterative process:

Start with a line segment (say, of length one), divide it into three congruent pieces, and remove the middle piece. Then add two pieces, congruent to the piece that was removed, to the remaining line segments as shown below. Repeat this process of removing the middle third from each line segment and replacing it with two congruent pieces.

Below, we show several stages of the Koch curve. In a way that can be made precise, this process approaches a limit, which we define to be the Koch curve.

Stage 0

Stage 1

Stage 2

Stage 3

Example 2. The Cantor set is constructed by the following iterative process:
Divide a line segment (say, of length one) into three congruent pieces and remove the middle piece. Repeat this process for each of the remaining line segments.

Several stages in the construction of the Cantor set are shown below. The limit\(^9\) of this repetitive process is the Cantor set.

Stage 0

Stage 1

Stage 2

Stage 3

Example 3. The Sierpinski triangle is constructed by the following iterative process:

Divide an equilateral triangle (say, with side length one) into four congruent pieces and remove the middle piece. Repeat this process for each of the remaining triangles. The limit (intersection) of the stages is the Sierpinski triangle.

Stage 0

Stage 1

Stage 2

\(^9\)The “limiting process” for the Cantor set is easier than for the Koch curve; here the limit is simply the intersection of the stages.
5. The dimension of a self-similar set

Suppose that a set can be split into pieces, each of which is a rescaled copy of the original set, with scaling factor \( m \). We define the fractal dimension of the set to be the number \( d \) such that \( n = m^d \).

As a first example, consider a line segment. It is (highly!) self-similar, in that it can be divided into two subintervals (hence \( n = 2 \)), each scaled down by a factor of two (hence \( m = 2 \)). From the equation \( n = m^d \), we conclude that \( d = 1 \). This should reassure us that our definition is reasonable, since we expected a line segment to have dimension one.

Next, consider a solid rectangle. It can be divided into four smaller rectangles (by bisecting the sides); each is similar to the original rectangle, but scaled down by a factor of two. From \( n = 4 \) and \( m = 2 \), we conclude \( d = 2 \), as we would hope for a region in the plane, which has two dimension, in the sense of Section 4.1.

**Your Turn 17.** Suppose we considered a finer subdivision of the rectangle, say by dividing each side into five. Show that the formula \( n = m^d \) again gives \( d = 2 \).

**Your Turn 18.** Prove that the fractal dimension of a solid box in space is 3.

Let’s use the relationship \( n = m^d \) to describe the dimension of the Cantor set. We think of stage zero as the interval \([0, 1]\), which contains the Cantor set. The Cantor set can be divided into two pieces, namely the portion between 0 and 1/3 and the portion between 2/3 and 1 (there is no portion between 1/3 and 2/3!). Each of these is a copy of the Cantor set, scaled down by a factor of three. Thus we have \( 2 = 3^d \), so \( d = \log_3 2 = \ln 2 / \ln 3 \), or approximately 0.63. As we might expect, the dimension of the Cantor set is greater than 0 but less than 1, reflecting that is is too small to be a true curve.

A similar analysis gives the fractal dimension of the Koch curve. Perhaps looking at the Stage 3 version for inspiration, it is easy to see how to break the Koch curve into four pieces, each a third the size of the full

---

10Here, a subset of a line, plane, or three-dimensional space.
Koch curve. From \( n = 4, m = 3 \), we conclude that \( d = \ln 4 / \ln 3 \). This value is between 1 and 2, reflecting that the Koch curve is too thick to be merely a curve, but thinner than a solid region in the plane.

**Your Turn 19.** Compute the dimension of the Sierpinski triangle. (Between what two whole numbers do you expect the dimension to fall?)

5.1. Measuring the length of fractal curves. Is it possible to assign a length to the Koch curve? Recall that the length of the line segment used to create the Koch curve was 1. After removing the middle third and adding two congruent pieces, the length of the curve increases to \( 4/3 \). During the second iteration, the middle third of each of the four \( 1/3 \) pieces is removed, and the sum of the remaining segments is \( 8/9 \). Adding eight new pieces, each of length \( 1/9 \), gives a perimeter of \( 8/9 + 8/9 \) or \( 16/9 \). Thus the length of the Koch curve increases from \( 4/3 \) to \( 16/9 \) during this second iteration. In general, the length of a stage is four thirds times the length of the previous stage, hence the length of the \( n \)th stage is \( (4/3)^n \). Hence the limit of the lengths of the stages is infinity, and it is reasonable to regard the Koch curve as having infinite length.

**Your Turn 20.** Give a similar analysis to assign values to the length of the Cantor set, and the area of the Sierpinski triangle.

5.2. Exercises.

1. **Measuring the length of the Koch curve.** Verify that the length of the stage ten Koch curve is nearly 18, and over 300 at stage 20.

2. At what stage of the construction of the Cantor set does the length drop below 0.001? At what stage of the construction of the Sierpinski triangle does the area drop below 0.001?

3. A student believes that since the area of the Sierpinski triangle is zero, it must contain only finitely many points. Is this true? Explain.

4. Suppose that the interval used at the 0th stage of the Cantor set is the closed interval \([0, 1]\).
   
   (a) Prove that every rational number between 0 and 1, whose denominator is a power of three, lies in the Cantor set.
5. The Dimension of a Self-Similar Set

(b) Is every point in the Cantor set of the type described in (a)?

A closed figure called a Koch Snowflake is created by the following iterative process: Begin with an equilateral triangle. Construct a Koch curve on each side of the triangle, as shown in Figure 20.

Figure 20. Koch Snowflake

5. Measuring the dimension of the Koch Snowflake.

(a) Draw the next two iterations of the Koch Snowflake.

(b) Compute the fractal dimension of the Koch Snowflake.

6. Measuring the perimeter of the Koch Snowflake. Define the measure of a side of the initial triangle used to construct the Koch Snowflake to be 1 linear unit.

(a) Determine the perimeter of the Koch snowflake at each of the first four iterations.

(b) Construct a general equation to describe the perimeter of the Koch Snowflake at the nth iteration.

(c) Find the limit of the perimeter of the Koch Snowflake as \( n \to \infty \).

7. Measuring the area of the Koch Snowflake. Define the area of the initial triangle used to construct the Koch Snowflake to be 1 triangular unit.

(a) Determine the area of the Koch snowflake at each of the first four iterations.
(b) Construct a general equation to describe the area of the Koch Snowflake at the $n$th iteration.

(c) Find the limit of the area of the Koch Snowflake as $n \to \infty$.

8. Research the fractal known as the Menger Sponge. Give its definition, and calculate its fractal dimension and its volume.

6. Exercises Involving Student Work: Measuring Two-dimensional Figures

For Problems 1. and 2. below,

• Complete the given task yourself. Compare your responses with those of a partner or small group.
• Read through the student responses provided.
• Comment on the quality of each student response. Identify ways in which the students’ thinking is correct and ways in which the student’s thinking is incorrect or incomplete.

1. Is it possible for a rectangle to have the same perimeter and area? If so, construct such a rectangle. If not, explain why this is not possible.

2. Is it possible for a rectangle and a triangle to have the same area? If so, construct a rectangle and a triangle that have the same area. If not, explain why this is not possible.