A Transition to Advanced Mathematics

Darrin Doud and Pace P. Nielsen

Darrin Doud
Department of Mathematics
Brigham Young University
Provo, UT 84602
doud@math.byu.edu

Pace P. Nielsen
Department of Mathematics
Brigham Young University
Provo, UT 84602
pace@math.byu.edu

Version: 1.03
Date: October 21, 2022

Contents

Preface

This book is intended as the text for the Math 290 (Fundamentals of Mathematics) class at Brigham Young University. It covers several fundamental topics in advanced mathematics, including set theory, logic, proof techniques, number theory, relations, functions, and cardinality. These topics are prerequisites for most advanced mathematics classes, and it seems worthwhile to have a specific course in which they can be learned by students.

The prerequisites for understanding this material are surprisingly light. Typically, only a small amount of college algebra (manipulating simple algebraic expressions) and knowledge of decimal expansions and prime numbers are needed; most other necessary material is covered in the text. The book is designed for a semester-long class; each section contains an appropriate amount of material for an hour long lecture. The exercise sets at the end of each section give problems for the students to use to practice the techniques learned in the section, and to develop their understanding of the material. At BYU there are typically 42 class days in a typical semester; we have included 36 sections in this book. This allows a few days for instructors to review for exams or cover additional topics of their choice.

We are often asked if we will produce a solutions manual for the exercises. For this particular course, a solutions manual is probably not a great benefit to the student. Unlike most mathematics courses that students will have before studying this material, the exercises in this book often do not have a single correct answer. Indeed, as the student progresses further into the book, most of the problems ask for proofs (or disproofs) of statements. Much of the learning in a course such as this comes from the struggle to produce a proof, rather than studying the techniques used by someone else to give a proof. Hence, providing a solutions manual would negate a necessary aspect of the course. In addition, a solutions manual would be of very little help in verifying the correctness of a proof, since there are many different ways to prove almost any given statement, all equally correct.

One aspect of a proof is that it should be a convincing argument that a statement is correct. A student should consider their solution of a proof-type problem to be aimed at an audience of students at their level; if they are unsure if it is a valid proof, then their goal has not been met.

In addition, it is important to note that most of the solutions to exercises in the book will involve much more writing than is usual in previous mathematics classes. In order to adapt to this increase in writing, students may need to change the way they think about problems.

One additional topic that instructors may want to include in a course based on this book is writing mathematics using LaTeX. This is an important skill for mathematicians, engineers, scientists, and mathematics educators. Because technology moves quickly, we have not included instructional material on LaTeX in this text; an internet search can easily find a plethora of such material. In our courses, we typically spend one to two class days instructing students on the use of LaTeX, as well as giving a number of assignments to help students develop their skills in LaTeX.

We thank the many BYU students and instructors who have worked through preliminary versions of this textbook. They have discovered many typographical and other errors, which have been eliminated. Should a reader discover any additional errors in the text, please inform us, so that we can correct them in future printings.

Darrin Doud
Pace P. Nielsen

Chapter I

Set Theory

There is surely a piece of divinity in us, something that was before the elements, and owes no homage unto the sun. Sir Thomas Browne

One of the benefits of mathematics comes from its ability to express a lot of information in very few symbols. Take a moment to consider the expression

$$\frac{d}{d\theta}\sin(\theta).$$

It encapsulates a large amount of information. The notation $\sin(\theta)$ represents, for a right triangle with angle θ, the ratio of the opposite side to the hypotenuse. The differential operator $d/d\theta$ represents a limit, corresponding to a tangent line, and so forth.

Similarly, sets are a convenient way to express a large amount of information. They give us a language we will find convenient in which to do mathematics. This is no accident, as much of modern mathematics can be expressed in terms of sets.

1 Sets, subsets, and set operations

1.A What is a set?

A *set* is simply a collection of objects. The objects in the set are called the *elements*.

We often write down a set by listing its elements. For instance, the set $S = \{1, 2, 3\}$ has three elements. Those elements are 1, 2, and 3. There is a special symbol, $\in$, that we use to express the idea that an element belongs to a set. For instance, we write $1 \in S$ to mean that "1 is an element of S."

For the set $S = \{1, 2, 3\}$, we have $1 \in S$, $2 \in S$, and $3 \in S$. We can write this more quickly as: $1, 2, 3 \in S$. We can express the fact that 4 is not an element of S by writing $4 \notin S$.

Example 1.1. Let P be the set $\{16, -5, 2, 6, 9\}$. Is $6 \in P$? Yes! Is $5 \in P$? No, so we write $5 \notin P$. △

The order of the elements in a set does *not* matter, so we could have written $S = \{1, 2, 3\}$ as $S = \{1, 3, 2\}$, or as $S = \{3, 2, 1\}$. If an element is repeated in a set, we do *not* count the multiplicity. Thus $\{1, 2, 3, 1\}$ is the same set as $S = \{1, 2, 3\}$. We say that two sets are *equal* when they have exactly the same elements.

Not all sets consist of numbers. For instance $T = \{a, b, c, d\}$ is a set whose elements are the letters a, b, c, d. Sets may have words, names, symbols, and even other sets as elements.

Example 1.2. Suppose we want to form the set of Jesus' original twelve apostles. This would be the set

$$\text{Apostles} = \{\text{Peter, James, John the beloved, } \ldots, \text{ Judas Iscariot}\}.$$

We put the 3 dots in the middle to express the fact that there are more elements which we have not listed (perhaps to save time and space). △

The Last Supper, ca. 1520, by Giovanni Pietro Rizzoli.

The next example is a set with another set as an element.

Example 1.3. Let $S = \{1, 5, \{4, 6\}, 3\}$. This set has four elements. We have $1, 5, 3, \{4, 6\} \in S$, but $4 \notin S$. However, $4 \in \{4, 6\}$ and $\{4, 6\} \in S$. △

It can be confusing when sets are elements of other sets. You might ask why mathematicians would allow such confusion! It turns out that this is a very useful thing to allow; just like when moving, the moving truck (a big box) has boxes inside of it, each containing other things.

Advice 1.4. You can think of sets as *boxes* with objects inside. So we could view the set $\{1, 5, \{4, 6\}, 3\}$ from the previous example as the following box, which contains another box:

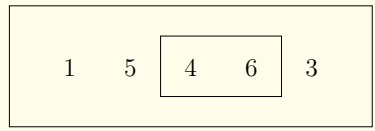


Figure 1.4: A box with a box inside, each containing some numbers.

1.B Naming sets

We've seen that capitalized Roman letters can be used to give names to sets. Some sets are used so often that they are represented by special symbols. Here are a couple of examples.

- The set of *natural numbers* is the set

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

  This is the first example we've given of an *infinite set*, i.e., a set with infinitely many elements.
- The set of *integers* is

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, 4, \ldots\}.$$

  The dots represent the fact that we are leaving elements unwritten in both directions. We use the fancy letter "$\mathbb{Z}$" because the word "integer" in German is "Zahlen."

Some sets are constructed using rules. For example, the set of even integers can be written as

$$\{\ldots, -4, -2, 0, 2, 4, \ldots\}$$

but could also be written in the following ways:

$$\{2x : x \in \mathbb{Z}\} \tag{1.5}$$

$$\{x \in \mathbb{Z} : x \text{ is an even integer}\} \tag{1.6}$$

$$\{x : x = 2y \text{ for some } y \in \mathbb{Z}\}. \tag{1.7}$$

We read the colon as "such that," so (1.5) is read as "the set of elements of the form $2x$ *such that* x is an integer." Writing sets with a colon is called *set-builder notation*. Notice that

$$\{x \in \mathbb{Z} \,:\, 2x + 1\}$$

doesn't make any sense, since "$2x + 1$" is not a condition on x.

Here are a few more examples. The set of prime numbers is

$$\{2, 3, 5, 7, 11, 13, \ldots\} = \{x \in \mathbb{N} \,:\, x \text{ is prime}\}.$$

Similarly, Apostles $= \{x \,:\, x$ was one of the original 12 apostles of Jesus$\}$.

With set-builder notation, we can list a few more very important sets.

- The set of *rational numbers* is

$$\mathbb{Q} = \{a/b \,:\, a, b \in \mathbb{Z}, b \neq 0\}.$$

  Note that there is no problem with the fact that different fractions can represent the same rational number, such as $1/2 = 2/4$. Repetitions do not matter in sets. We will occasionally need the fact that we can always write a rational number as a fraction a/b in *lowest terms*: i.e., so that a and b have no common factor larger than 1. We will prove this in Section 18 (see Exercise 18.6).
- The set of *real numbers* is

$$\mathbb{R} = \{x \,:\, x \text{ has a decimal expansion}\}.$$

  So we have $\pi = 3.14159\ldots \in \mathbb{R}$, $3 = 3.00000\ldots \in \mathbb{R}$, and $\sqrt{2} \in \mathbb{R}$. Later in this book we will prove $\sqrt{2} \notin \mathbb{Q}$.
- The set of *complex numbers* is

$$\mathbb{C} = \{a + bi \,:\, a, b \in \mathbb{R}, i^2 = -1\}.$$

  Is 3 a complex number? Yes, because we can take $a = 3$ and $b = 0$. So we have $3 \in \mathbb{N}$, $3 \in \mathbb{Z}$, $3 \in \mathbb{Q}$, $3 \in \mathbb{R}$, and $3 \in \mathbb{C}$!

Example 1.8. Which of the named sets does $\pi = 3.14159\ldots$ belong to? We have $\pi \in \mathbb{R}$ and $\pi \in \mathbb{C}$. On the other hand, since $3 < \pi < 4$ we have $\pi \notin \mathbb{N}$ and $\pi \notin \mathbb{Z}$. It is true, but much harder to show, that $\pi \notin \mathbb{Q}$. △

There is one more set we will give a special name.

- The *empty set* is the set with no elements. We write it as $\emptyset = \{\,\}$.

Warning 1.9. The empty set is not *nothing*. It has no elements, but the empty set is *something*. Namely, it is "the set with nothing in it."

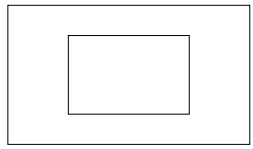
Thinking in terms of boxes, we can think of the empty set as an empty box. The box is *something* even if it has nothing in it.

The symbol $\emptyset$ does not mean nothing. It means $\{\,\}$.

Example 1.10. Sometimes we want the empty set to be an element of a set. For instance, we might take

$$S = \{\emptyset\}.$$

The set S has a single element, namely $\emptyset$. We could also write $S = \{\{\ \}\}$. In terms of boxes, S is the box containing an empty box. Note that *not* all sets have the empty set as an element. △



A box with an empty box inside, representing $\{\emptyset\}$.

1.C Subsets

In many activities in life we don't focus on all the elements of a set, but rather on subcollections. To give just a few examples:

- The set of all phone numbers is too large for most of us to handle. The subcollection of phone numbers of our personal contacts is much more manageable.
- If we formed the set of all books ever published in the world, this set would be very large (but still finite!). However, the subcollection of books we have read is *much* smaller.
- If we want to count how many socks we own, we could use elements of the integers $\mathbb{Z}$, but since we cannot own a negative number of socks, a more natural set to use would be the subcollection of nonnegative integers

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

A subcollection of a set is called a *subset*. When A is a subset of B we write $A \subseteq B$ and if it is not a subset we write $A \nsubseteq B$. There are a couple different ways to think about the concept of A being a subset of B.

Option 1: To check that A is a subset of B, we check that every element of A also belongs to B.

Example 1.11. (1) Let $A = \{1, 5, 6\}$ and $B = \{1, 5, 6, 7, 8\}$. Is A a subset of B? Yes, because we can check that each of A's three elements, 1, 5, and 6, belongs to B.

(2) Let $A = \{6, 7/3, 9, \pi\}$ and $B = \{1, 2, 6, 7/3, \pi, 10\}$. Is $A \subseteq B$? No, because $9 \in A$ but $9 \notin B$. So we write $A \nsubseteq B$.

(3) Let $A = \mathbb{N}$ and $B = \mathbb{Z}$. Is $A \subseteq B$? Yes, every natural number is an integer.

(4) Is $\mathbb{Z}$ a subset of $\mathbb{N}$? No, because $\mathbb{Z}$ has the element -1, which doesn't belong to $\mathbb{N}$. △

Option 2: To check that A is a subset of B, we check that we can form A by throwing out some of the elements of B.

Example 1.12. (1) Let $A = \{1, 5, 6\}$ and $B = \{1, 5, 6, 7, 8\}$. Is A a subset of B? Yes, because we can throw away $7, 8$ from B to get A.
(2) Let $A = \{6, 7/3, 9, \pi\}$ and $B = \{1, 2, 6, 7/3, \pi, 10\}$. Is A a subset of B? No, because as we throw away elements of B, we can never get 9 inside.
(3) Let $A = \mathbb{N}$ and $B = \mathbb{Z}$. Is A a subset of B? Yes, because we can throw away the negative integers and 0 to get the natural numbers.
(4) Is $\mathbb{Z}$ a subset of $\mathbb{N}$? No, because we cannot get -1 by throwing away elements from $\mathbb{N}$. △

When we write $A \subseteq B$, the little line segment at the bottom of "$\subseteq$" means that there is possible equality. (Just like $x \leq y$ means that x is less than *or equal to* y.) Sometimes we do not want to allow equality. We use the following terminology in this case.

Definition 1.13. If $A \subseteq S$ and $A \neq S$, we say that A is a *proper* subset of S, and we write $A \subsetneq S$.

Note that the symbol $\subsetneq$ is different from $\not\subseteq$. If $A \subsetneq B$, then A is a subset of B that is not equal to B, while if $A \not\subseteq B$, then A is not a subset of B.

Example 1.14. We have $\{1, 2\} \subsetneq \{1, 2, 3\}$. Of course $\{1, 2\} \subseteq \{1, 2, 3\}$ is also true. △

Warning 1.15. Some authors use $\subset$ instead of $\subseteq$. Other authors use $\subset$ instead of $\subsetneq$. Thus, there can be a lot of confusion about what $\subset$ means, which is one reason why we will avoid that notation in this book!

Warning 1.16. Many students learning about subsets get confused about the difference between being an element and being a subset. Consider your music library as a set. The elements are the individual songs. Playlists, which are collections of some of the songs, are subsets of your library.

Example 1.17. The elements of $\mathbb{C}$ are complex numbers like $3+6i$ or $-2.7-5.9i$. The subsets of $\mathbb{C}$ are sets of complex numbers like $\{5.4-7.3i, 9+0i, -2.671+9.359i\}$. △

Example 1.18. (1) Let $T = \{1, 2, 3, 4, 5\}$. Is 2 an element or a subset of T? It is an element, since it lives inside T. It is not a subset, since it isn't a set of elements of T.
(2) Let $U = \{-5, 6, 7, 3\}$. Is $\{6\}$ an element or a subset of U? It is not an element of U, since the set $\{6\}$ isn't in its list of elements. It is a subset because it is a box whose elements come from U.

(3) Let $X = \{\{6\}, \{7, 8\}, \{5, 8\}\}$. Is 7 an element or a subset of X? Neither! It is not one of the three elements listed in X, and it is not a box of elements in X either.
Is $\{7, 8\}$ an element or a subset of X? It is an element, since it is one of the three listed elements. It is not a subset, even though it is a box, since it has elements which don't belong to X.
(4) Let $Y = \{5, \{5\}\}$. Is $\{5\}$ an element or a subset of Y? It is both! It is an element, since it is the second element listed inside Y. It is also a subset of Y, since it is a box containing the first element of Y. △

It can be useful to construct sets satisfying certain properties in relation to one another. In the following example we show how this can be done.

Example 1.19. We will find three sets A, B, C satisfying the following conditions:
(1) $A \subseteq B$,
(2) $A \in C$, and
(3) $C \subseteq B$ with $C \neq B$ (i.e., $C \subsetneq B$).

One method to solve this problem is to start with the simplest sets possible and modify them as needed. So let's start with

$$A = \{\ \}, \qquad B = \{\ \}, \qquad C = \{\ \}.$$

We see that condition (1) is fulfilled, but condition (2) is not. To force condition (2) to be true, we must make A an element of C. Thus, our new sets are

$$A = \{\ \}, \qquad B = \{\ \}, \qquad C = \{A\}.$$

Condition (1) still holds, and condition (2) is now true. However, condition (3) doesn't hold. To make (3) true, we need B to have all the elements of C and at least one more. So we take

$$A = \{\ \}, \qquad B = \{A, 1\}, \qquad C = \{A\}.$$

We double-check that all of the conditions hold (which they do), and so we have our final answer. △

1.D Cardinality

The number of elements of a set is called its *cardinality*. For instance, the set $S = \{1, 2, 3\}$ has 3 elements. We write $|S| = 3$ to denote that S has cardinality 3. Note that $|\emptyset| = 0$ but $|\{\emptyset\}| = 1$. A set is *finite* if its cardinality is either 0 or a natural number, and it is *infinite* otherwise. In a later section in the book, we will talk about a better way to define cardinality for infinite sets.

Example 1.20. If $T = \{5, \{6, 7, 8\}, \{3\}, 0, \emptyset\}$, the cardinality is $|T| = 5$. △

In mathematics, we sometimes use the same symbols for two different things. The meaning of the symbols must be deduced from their context. For instance, if we write $|-3.392|$ this is certainly not the cardinality of a set, but instead is probably referring to the absolute value of a number. In the next example, we use $|\cdot|$ in two different ways.

Example 1.21. If $T = \{x \in \mathbb{Z} : |x| < 4\}$, what is $|T|$?
(Hint: It is bigger than 4.) △

1.E Power sets

In this section we define the power set and give some examples.

Definition 1.22. Let S be a set. The *power set* of S is the new set $\mathscr{P}(S)$ whose elements are the subsets of S. In other words, $A \in \mathscr{P}(S)$ exactly when $A \subseteq S$.

The next example determines the power set of a small set S.

Example 1.23. Can we list all of the subsets of $S = \{1, 2, 3\}$? If we think about subsets as "boxes containing only elements of S", we just have to list all possibilities. They are as follows:

$$\mathscr{P}(S) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

Why is the empty set one of the subsets? Is it really a box containing only elements of S? Yes, its elements (there are none!) all belong to S. Thinking about it in terms of "throwing away" elements of S, we threw all of them away.

Why is $S \subseteq S$? Because S is a box containing only elements of S. Thinking in terms of "throwing away" elements, we threw away none of the elements. △

If S is a finite set, we can determine the size of the power set $|\mathscr{P}(S)|$ from $|S|$.

Theorem 1.24. *If $|S| = n$, then $|\mathscr{P}(S)| = 2^n$.*

Here is a sketch of why this is true. To form a subset of S, for each element in S we choose to keep or throw away that element. Thus, there are 2 choices for each element. Since there are n elements, this gives 2^n options. □

Example 1.25. For the set $S = \{1, 2, 3\}$ we have $|S| = 3$. Thus the power set has cardinality $|\mathscr{P}(S)| = 2^3 = 8$. This is exactly the number of elements we listed in Example 1.23. △

Example 1.26. How many elements will the power set of $U = \{1, \emptyset\}$ have? The set U has two elements, so there should be $2^2 = 4$ subsets. They can be listed as:

$$\mathscr{P}(U) = \{\emptyset, \{1\}, \{\emptyset\}, U\}.$$

Which of these are proper subsets of U? (All of them except U itself.) △

Example 1.27. List three elements of $\mathscr{P}(\mathbb{N})$, each having different cardinality, and one being infinite.

Here is one possible answer: $\{1, 7\}$, $\{67, 193, 91948\}$, and $\{2, 4, 6, 8, 10, \ldots\}$. There are many other correct choices. △

1.F Unions and intersections

There are multiple ways to modify sets. When there are two sets S and T, we can put them together to form a new set called the *union*, and we write

$$S \cup T = \{x \,:\, x \in S \;\text{ or }\; x \in T\}.$$

This is the set of elements which belong to S or T or both of them. (When we use the word "or" in this book, we will almost always use the inclusive meaning.) Pictorially, we can view this set using a Venn diagram as follows.

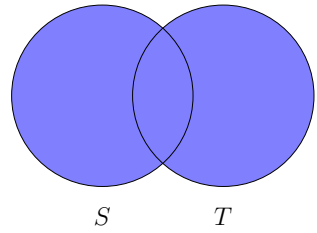


Figure 1.28: The union of S and T.

Similarly, given two sets S and T we can form the set of elements that belong to both of them, called the *intersection*, and we write

$$S \cap T = \{x \,:\, x \in S \;\text{ and }\; x \in T\}.$$

The Venn diagram is the following.

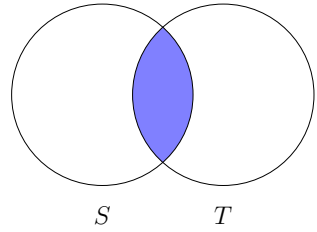


Figure 1.29: The intersection of S and T.

Example 1.30. Let $A = \{1, 6, 17, 35\}$ and $B = \{1, 5, 11, 17\}$. Then

$$A \cup B = \{1, 5, 6, 11, 17, 35\}, \qquad A \cap B = \{1, 17\}. \qquad \triangle$$

Example 1.31. Find sets P, Q with $|P| = 7$, $|Q| = 9$, and $|P \cap Q| = 5$. How big is $|P \cup Q|$?

We start by letting P be the easiest possible set with 7 elements, namely $P = \{1, 2, 3, 4, 5, 6, 7\}$. Since Q must share 5 of these elements, but have 9 elements total, we could write $Q = \{1, 2, 3, 4, 5, 8, 9, 10, 11\}$.

For the example we constructed, we have $P \cup Q = \{x \in \mathbb{N} : x \leq 11\}$, so $|P \cup Q| = 11$. If we chose other sets P and Q, could $|P \cup Q|$ be different? (Answer: No. The given numbers determine the cardinality of each piece in the Venn diagram.) $\triangle$

1.G Complements and differences

Let S and T be sets. The *difference* of T and S is

$$T - S = \{x : x \in T \text{ and } x \notin S\}.$$

The Venn diagram is as follows.

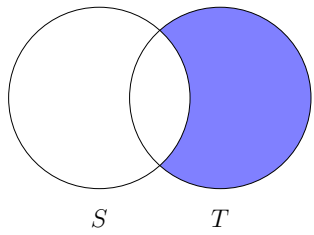


Figure 1.32: The difference of T and S.

Example 1.33. Let $S = \{1, 2, 3, 4, 5, 6, 7\}$ and $T = \{6, 7, 8, 9\}$. We find $T - S$ is the set

$$T - S = \{8, 9\}.$$

Notice that we do not need to worry about those elements of S which do not belong to T. We only have to take away the part the two sets share. So $T - S = T - (S \cap T)$.

Also notice that $S - T = \{1, 2, 3, 4, 5\}$ is different from $T - S$. $\triangle$

Example 1.34. Let A and B be sets. Assume $|A| = 16$ and $|B| = 9$. If $|A \cap B| = 2$, what are $|A - B|$ and $|B - A|$?

There are only two elements that A and B share, thus $|A - B| = 14$ and $|B - A| = 7$.

Can you now figure out $|A \cup B|$? (Hint: Draw the Venn diagram.) $\triangle$

Occasionally we will be working inside some set U, which we think of as the *universal set* for the problem at hand. For instance, when solving quadratic equations, such as $x^2 - x + 2 = 0$, your universal set might be the complex numbers $\mathbb{C}$.

Given a subset S of the universal set U, we write $\overline{S} = U - S$, and call this the *complement* of S (in the universal set U). The Venn diagram follows.

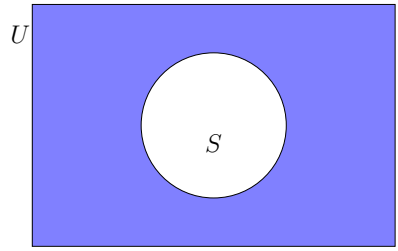


Figure 1.35: The complement of a set S inside a universal set U.

Example 1.36. Let $U = \mathbb{N}$ and let $P = \{2, 3, 5, 7, \ldots\}$ be the set of primes. What is $\overline{P}$? This is the set of composite numbers and 1, or in other words $\overline{P} = \{1, 4, 6, 8, 9, \ldots\}$. △

1.H Exercises

Exercise 1.1. Each of the following sets is written in set-builder notation. Write the set by listing its elements. Also state the cardinality of each set.

(a) $S_1 = \{n \in \mathbb{N} : 5 < |n| < 11\}$.
(b) $S_2 = \{n \in \mathbb{Z} : 5 < |n| < 11\}$.
(c) $S_3 = \{x \in \mathbb{R} : x^2 + 2 = 0\}$.
(d) $S_4 = \{x \in \mathbb{C} : x^2 + 2 = 0\}$.
(e) $S_5 = \{t \in \mathbb{Z} : t^5 < 1000\}$. (This one is slightly tricky.)

Exercise 1.2. Rewrite each of the following sets in the form

$$\{x \in S : \text{some property on } x\},$$

just as we did in (1.6) above, by finding an appropriate property.

(a) $A_1 = \{1, 3, 5, 7, 9, \ldots\}$ where $S = \mathbb{N}$.
(b) $A_2 = \{1, 8, 27, 64, \ldots\}$ where $S = \mathbb{N}$.
(c) $A_3 = \{-1, 0\}$ where $S = \{-1, 0, 1\}$.

Exercise 1.3. Write the following sets in set-builder notation.

(a) $A = \{\ldots, -10, -5, 0, 5, 10, 15, \ldots\}$.
(b) $B = \{\ldots, -7, -2, 3, 8, 13, 18, \ldots\}$.
(c) $C = \{1, 16, 81, 256, \ldots\}$.
(d) $D = \{\ldots, 1/4, 1/2, 1, 2, 4, 8, 16, \ldots\}$.

Exercise 1.4. Give specific examples of sets A, B, and C satisfying the following conditions (in each part, separately):

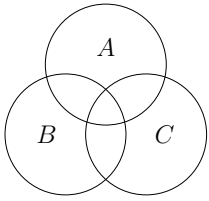
(a) $A \in B$, $B \in C$, and $A \notin C$.
(b) $A \in B$, $B \subseteq C$, and $A \nsubseteq C$.
(c) $A \subsetneq B$, $B \in C$, and $A \in C$.
(d) $A \cap B \subseteq C$, $A \nsubseteq C$, and $B \nsubseteq C$.
(e) $A \cap C = \emptyset$, $A \subseteq B$, and $|B \cap C| = 3$.

Exercise 1.5. Let $A = \{1, 2\}$. Find $\mathscr{P}(A)$, and then find $\mathscr{P}(\mathscr{P}(A))$. What are the cardinalities of these three sets?

Exercise 1.6. Let $a, b \in \mathbb{R}$ with $a < b$. The closed interval $[a, b]$ is the set $\{x \in \mathbb{R} : a \leq x \leq b\}$. Similarly, the open interval (a, b) is the set $\{x \in \mathbb{R} : a < x < b\}$. Let $P = [3, 7]$, $Q = [7, 9]$ and $R = [-3, 8]$. Give simple descriptions of the following sets.

(a) $P \cap Q$.
(b) $P \cup Q$.
(c) $P - Q$.
(d) $Q - P$.
(e) $(R \cap P) - Q$.
(f) $(P \cup Q) \cap R$.
(g) $P \cup (Q \cap R)$.

Exercise 1.7. Consider the following blank Venn diagram for the three sets A, B, C.



For each of the following sets, copy the Venn diagram above, and then shade in the named region:

(a) $A - (B \cap C)$.
(b) $A - (B - C)$.
(c) $B - (A - C)$.
(d) $(B \cap C) \cap (B \cup A)$.
(e) $(A - B) \cup (A - C)$.

Exercise 1.8. Two sets S, T are *disjoint* if they share no elements. In other words $S \cap T = \emptyset$. Which of the following sets are disjoint? Give reasons.

(a) The set of odd integers and the set of even integers.
(b) The natural numbers and the complex numbers.
(c) The prime numbers and the composite numbers.
(d) The rational numbers and the irrational numbers (i.e., real numbers which are not rational).

Exercise 1.9. Find some universal set U and subsets $S, T \subseteq U$, such that $|S - T| = 3$, $|T - S| = 1$, $|S \cup T| = 6$, and $|\overline{S}| = 2$. (Write each of U, S, and T by listing their elements.)

2 Products of sets and indexed sets

2.A Cartesian products

Sets are *unordered* lists of elements. There are situations where order matters. For instance, you probably don't want to put your shoes on before your socks. To give a more mathematical example, if we square a number and then take its cosine, that is not the same as first taking the cosine and then squaring:

$$\cos(x^2) \neq (\cos(x))^2.$$

There are other situations where we want to keep things ordered. We will write (x, y) for the *ordered pair* where x occurs first and y occurs second. Thus $(x, y) \neq (y, x)$ even though $\{x, y\} = \{y, x\}$. Also, an element can be *repeated* in an ordered list, such as $(1, 1)$, while sets do not count repetitions.

There is a very nice notation for sets of ordered pairs.

Definition 2.1. Let S and T be two sets. The *Cartesian product* of these sets is the new set

$$S \times T = \{(s, t) \,:\, s \in S,\ t \in T\}.$$

This is the set of all ordered pairs such that the first entry comes from S and the second entry comes from T. We will often refer to $S \times T$ just as the *product* of S and T.

We will now give an example of how to find simple Cartesian products.

Example 2.2. Let $S = \{1, 2, 3\}$ and $T = \{1, 2\}$. What is $S \times T$? It is the set $\{(1, 1), (1, 2), (2, 1), (2, 2), (3, 1), (3, 2)\}$. Notice that 3 can occur as a first coordinate since $3 \in S$, but not as a second coordinate since $3 \notin T$.

While the order matters inside an ordered pair, we could have listed the elements of $S \times T$ in a different order since $S \times T$ is itself just a set (and order is irrelevant in sets). So we could have written

$$S \times T = \{(1, 2), (2, 2), (3, 1), (1, 1), (2, 1), (3, 2)\}.$$

However,

$$S \times T \neq T \times S = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3)\}. \qquad \triangle$$

You might notice that in the previous example we have $|S \times T| = 6 = 3 \cdot 2 = |S| \cdot |T|$. This is not an accident. In fact, the following is true, although we do not as yet have the tools to prove it.

Proposition 2.3. *Let A and B be finite sets, with $|A| = m$ and $|B| = n$. Then $A \times B$ is a finite set, with $|A \times B| = mn$.*

Sets do not need to be finite in order to act as components in products.

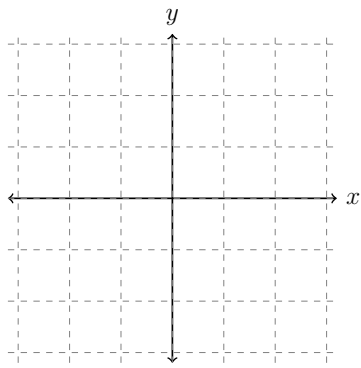
Example 2.4. Let $A = \mathbb{N}$ and $B = \{0, 1\}$. What are the elements of $A \times B$? They are

$$A \times B = \{(1,0), (1,1), (2,0), (2,1), (3,0), (3,1), \ldots\} = \{(n,0), (n,1) \,:\, n \in \mathbb{N}\}.$$

Is $A \times B$ the same set as $B \times A$? No, they have different elements. For instance, $(1,0) \in A \times B$, but $(1,0) \notin B \times A$ since $0 \notin A$. △

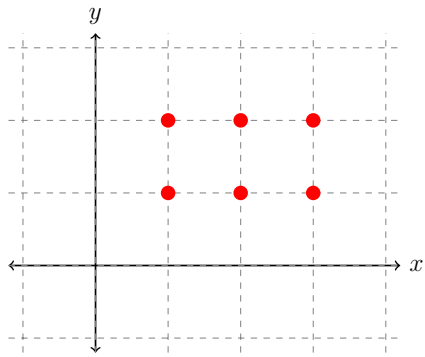
In each of the previous examples, we took the Cartesian product of two *different* sets. If we take the product of a set with itself, we sometimes write $A^2 = A \times A$. The following example is one of the most useful products of a set with itself.

Example 2.5. The set $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is called the *Cartesian plane*. We view elements in this set as points $\{(x, y) \,:\, x, y \in \mathbb{R}\}$.



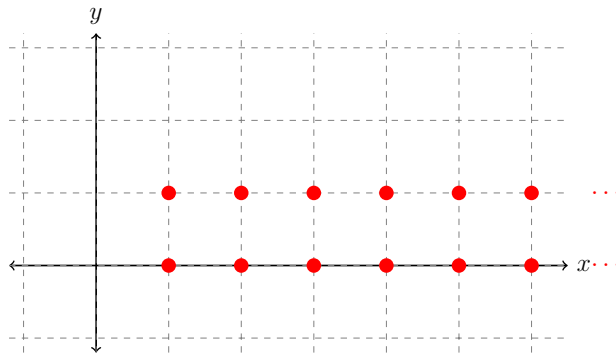
The Cartesian plane, $\mathbb{R} \times \mathbb{R}$

The set $S \times T$ in Example 2.2 is a subset of $\mathbb{R} \times \mathbb{R}$. We can graph it as follows:



$\{1, 2, 3\} \times \{1, 2\}$

Similarly, the set $A \times B$ from Example 2.4 is graphed as:



$\mathbb{N} \times \{0, 1\}$

We can now describe more complicated sets. For instance

$$\{(x, y) \in \mathbb{R} \times \mathbb{R} \,:\, y = 3x + 1\}$$

is a line. The set $\mathbb{R}^2 - \{(0, 0)\}$ is the punctured plane (the plane with the origin removed). Can you describe a simple parabola? △

The following example addresses the question: "What do we do if one of the sets has no elements?"

Example 2.6. We determine $\{1, 2, 3\} \times \emptyset$. Elements of this set are ordered pairs of the form (a, b), with $a \in \{1, 2, 3\}$ and $b \in \emptyset$. Thus, there are no possible choices for b, and so $\{1, 2, 3\} \times \emptyset = \emptyset$. Note that $3 \cdot 0 = 0$, so Proposition 2.3 works in this case too. △

Just as with ordered pairs, we can form the set of ordered triples

$$A \times B \times C = \{(a, b, c) \,:\, a \in A,\ b \in B,\ c \in C\}.$$

We can similarly form ordered quadruples, ordered quintuples, and so forth. The next subsection will give us the tools necessary to talk about even more complicated constructions.

2.B Indices

When we have a large number of sets, rather than writing them using different letters of the alphabet

$$A, B, C, D, \ldots, Z$$

it can be easier to use subscripts

$$A_1, A_2, A_3, A_4, \ldots, A_{26}.$$

This notation is extremely powerful for the following reasons:

- The notation tells us how many sets we are working with, using a small number of symbols. For instance, if we write $A_1, A_2, \ldots, A_{132}$, we know that there are exactly 132 sets. (Try writing them down using different letters of the alphabet!)
- We can even talk about an infinite number of sets $A_1, A_2, A_3, \ldots$. Notice that the subscripts all come from the set $\mathbb{N}$. We refer to $\mathbb{N}$ as the *index set* for this collection.
- Using indices we can form complicated unions, intersections, and Cartesian products.

Example 2.7. Let $A_1 = \{1, 2, 4\}$, $A_2 = \{-3, 1, 5, 9\}$, and $A_3 = \{1, 6, 10\}$. We find

$$\bigcup_{i=1}^{3} A_i = A_1 \cup A_2 \cup A_3 = \{-3, 1, 2, 4, 5, 6, 9, 10\}$$

and

$$\bigcap_{i=1}^{3} A_i = A_1 \cap A_2 \cap A_3 = \{1\}.$$

Those who have seen summation notation

$$\sum_{i=1}^{3} i^2 = 1^2 + 2^2 + 3^2$$

will recognize where this notation comes from. △

Can we form infinite unions and intersections? This is actually a common occurrence.

Example 2.8. Let $B_1 = \{1, -1\}$, $B_2 = \{2, -2\}$, $B_3 = \{3, -3\}$, and so forth. In other words $B_n = \{n, -n\}$ for each $n \in \mathbb{N}$. (Notice that while the subscripts come from $\mathbb{N}$, the elements of the sets B_n come from $\mathbb{Z}$.)

The union is the set of elements which belong to at least one of the sets, thus

$$\bigcup_{n=1}^{\infty} B_n = \{\ldots, -3, -2, -1, 1, 2, 3, \ldots\} = \mathbb{Z} - \{0\}.$$

The intersection is the set of elements which belong to every one of the sets, thus

$$\bigcap_{n=1}^{\infty} B_n = \{\ \} = \emptyset.$$ △

There is an alternative way to write intersections and unions, using index sets. For instance, using the notation in the previous two examples, we could also write

$$\bigcap_{i=1}^{3} A_i = A_1 \cap A_2 \cap A_3 = \bigcap_{i \in \{1,2,3\}} A_i$$

and

$$\bigcup_{n=1}^{\infty} B_n = \bigcup_{n \in \mathbb{N}} B_n.$$

There is nothing to limit our index set, so we can make the following broad definition.

Definition 2.9. Let I be any set, and let S_i be a set for each $i \in I$. We put

$$\bigcup_{i\in I} S_i = \{x : x \text{ belongs to } S_i \text{ for some } i \in I\}$$

and

$$\bigcap_{i\in I} S_i = \{x : x \text{ belongs to } S_i \text{ for each } i \in I\}.$$

The next example shows, once again, how mathematics has the uncanny ability to express information in varied subjects using very simple notation.

Example 2.10. Let $A = \{\text{a,b,c,d,}\ldots,\text{z}\}$ be the "lowercase English alphabet set." This set has twenty-six elements. Let $V = \{\text{a,e,i,o,u}\}$ be the "standard vowel set." Notice that $V \subsetneq A$.

Given $\alpha \in A$, we let W_α be the set of words in the English language containing the letter α. Note that α is a *dummy variable*, standing in for an actual element of A. For instance, if $\alpha = \text{x}$ then we have

$$W_{\text{x}} = \{\text{xylophone, existence, axiom,}\ldots\},$$

while if $\alpha = \text{t}$ then we have

$$W_{\text{t}} = \{\text{terminator, atom, attribute,}\ldots\}.$$

Each set of words W_α is a subset of the universal set of all words in the English language.

Try to answer the following questions:

(1) What is $\bigcap_{\alpha\in V} W_\alpha$?
(2) Is that set empty?
(3) What is $\overline{\bigcup_{\alpha\in V} W_\alpha}$?
(4) Is that set empty?

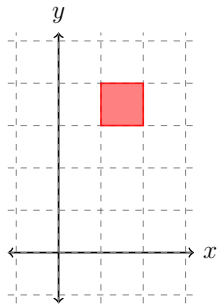
Here are the answers. (Look at them only after you have your own!)

(1) This is the set of words that contain every standard vowel.
(2) It isn't empty, since it contains words like "sequoia," "evacuation," etc.
(3) This is the set of words with no standard vowels. (Don't forget that there is a bar over the union.)
(4) It isn't empty, since it contains words like "why," "tsktsk," etc. △

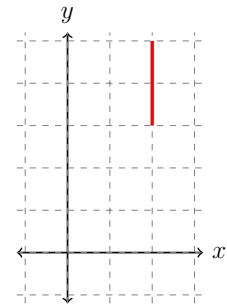
We finish with one more difficult example.

Example 2.11. We determine $\bigcup_{x\in[1,2]} [x, 2] \times [3, x + 3]$.

First, to get a footing on this problem, we try to understand what happens for certain values of x. The smallest possible x value in the union is when $x = 1$. There we get that $[x, 2] \times [3, x + 3] = [1, 2] \times [3, 4]$. This is the set of ordered pairs $\{(x, y) \in \mathbb{R}^2 : 1 \le x \le 2, 3 \le y \le 4\}$. This is just a box in the plane. Its graph is the first graph below.



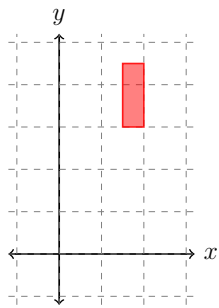
$[1,2] \times [3,4]$



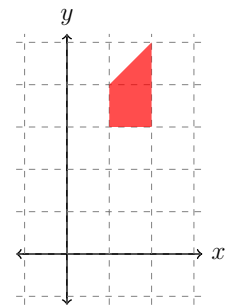
$\{2\} \times [3,5]$

The largest possible x value in the union is when $x = 2$. There we get that $[x, 2] \times [3, x + 3] = [2, 2] \times [3, 5]$. Notice that $[2, 2] = \{2\}$ is just a single point. Now $\{2\} \times [3, 5]$ is a line segment in the plane, where the x-value is 2 and the y-values range from 3 to 5. Its graph is the second graph above.

If we consider the intermediate value $x = 1.5$, we get the box $[1.5, 2] \times [3, 4.5]$, graphed below on the left.



$[1.5, 2] \times [3, 4.5]$



$\bigcup_{x \in [1,2]} ([x, 2] \times [3, x + 3])$

Taking the union over all $x \in [1, 2]$, we get the region graphed above on the right, boxed in by the lines $x = 1$, $y = 3$, $x = 2$, and $y = x + 3$.

△

2.C Exercises

Exercise 2.1. Sketch each of the following sets in the Cartesian plane $\mathbb{R}^2$.

(a) $\{1, 2\} \times \{1, 3\}$.
(b) $[1, 2] \times [1, 3]$.
(c) $(1, 2] \times [1, 3]$. (Hint: If an edge is missing, use a dashed, rather than solid, line for that edge.)
(d) $(1, 2] \times \{1, 3\}$.

Exercise 2.2. Let $A = \{s, t\}$ and $B = \{0, 9, 7\}$. Write the following sets by listing all of their elements.

(a) $A \times B$.
(b) $B \times A$.
(c) A^2.
(d) B^2.
(e) $\emptyset \times A$.

Exercise 2.3. Answer each of the following questions with "True" or "False" and then provide a reason for your answer.

(a) If $|A| = 3$ and $|B| = 4$, then $|A \times B| = 7$.
(b) It is always true that $A \times B = B \times A$ when A and B are sets.
(c) Assume I is an indexing set, and let S_i be a set for each $i \in I$. We always have $\bigcap_{i \in I} S_i \subseteq \bigcup_{i \in I} S_i$.
(d) There exist distinct sets $S_1, S_2, S_3, \ldots$, each of which is infinite, but

$$\bigcap_{i=1}^{\infty} S_i$$

has exactly one element.
(e) The set A^4 consists of ordered triples from A.

Exercise 2.4. Using the notations from Example 2.10, write the following sets (possibly using intersections or unions).

(a) The set of words containing all four of the letters "a,w,x,y."
(b) The set of words not containing any of the letters "s,t,u."
(c) The set of words containing both of the letters "p,r" but not containing any of the standard vowels. (Is this set empty?)

Exercise 2.5. For each number $r \in \mathbb{R}$, consider the "parabola shifted by r" defined as:

$$P_r = \{(x, y) \in \mathbb{R}^2 \, : \, y = x^2 + r\}.$$

Give simple descriptions of the following sets. Your description should not refer to the dummy variable "r" or use any unions or intersections. Also graph the sets in the Cartesian plane.

(a) $\bigcup_{r \in \mathbb{R}} P_r$.
(b) $\bigcup_{r > 0} P_r$.
(c) $\bigcup_{r \neq 0} P_r$.
(d) $\bigcap_{r \in \mathbb{R}} P_r$.
(e) $\bigcap_{r > 0} P_r$.

(When we write $r > 0$ we simply mean that we have restricted the dummy variable r to the subset of positive reals. Similarly, $r \neq 0$ means that we have restricted r to the subset of nonzero real numbers.)

Chapter II

Logic

I am convinced that the act of thinking logically cannot possibly be natural to the human mind. If it were, then mathematics would be everybody's easiest course at school and our species would not have taken several millennia to figure out the scientific method. Neil deGrasse Tyson

At higher levels, mathematics is about proving theorems. A study of logic is central to any endeavor involving proof. In this chapter, we introduce the study of logic, beginning with the basic building blocks (statements) and the different ways of connecting them (logical connectives).

Of particular significance will be the study of implication. Many important theorems in mathematics are stated as implications. For instance, the Mean Value Theorem in calculus is:

> *If a function f is continuous on a closed interval $[a, b]$ and differentiable on the open interval (a, b), then there is some $c \in (a, b)$ such that*
>
> $$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

This is an implication; if a certain condition is true, then some result is implied and can be inferred to be true.

In addition to logical connectives, this chapter will introduce quantifiers, which are also very common in mathematical writing. As an example, the Mean Value Theorem (stated above) includes the quantified phrase "there is some $c \in (a, b)$." Proper understanding and use of quantifiers is essential to understanding mathematical writing.

3 Statements

3.A What is a statement?

Definition 3.1. A *statement* is a declarative sentence that has a truth value.

By a *declarative sentence* we mean a sentence that makes an assertion. This assertion can in fact be false; in that case we say that the statement is false. Examples of statements include the following:

- The number 3 is larger than the number 2.
- The sun is blue.
- There are exactly 100,000 words in this book.
- The number 7 is an odd integer.
- Birds are mammals.

Each of these statements has a truth value. The first and fourth statements are true, the second and fifth are false, and the third can be determined to be true or false by counting the words in this book. Whether we know the truth value of a sentence is irrelevant when deciding whether the sentence is a statement; what is important is that the truth value exists.

Often in mathematics it is desirable to use symbols to denote objects. If we wish to use a symbol to stand for a statement, we will often use the letters P, Q, or R (although any symbol could be used). Since many mathematical statements involve an equals sign, we will not use an equals sign to describe a statement being assigned to a symbol; instead, we will use a colon. Hence, if we write

$$P\text{: 2 is an even number}$$

we are stating that the symbol P will stand for the statement "2 is an even number."

Sentences that are not statements could include questions ("What is the color of the sky?"), commands ("Solve the equation"), or opinions ("Chocolate ice cream is the best"). Moreover, words that don't actually form sensible sentences do not have truth values ("Try apples is"). In addition, sentences involving variables, such as "x is greater than 2," are not statements, unless the value of x is known. (We will talk more about sentences containing variables in the next section.)

3.B Compound statements

Given several statements, we will often want to combine them in ways to create more complex statements. There are a number of different operations that can be used to combine or modify statements.

Definition 3.2. Let P and Q be statements. The *conjunction* of P and Q, written

$$P \wedge Q,$$

is a statement that is true if both P and Q are true, and false otherwise.

When reading the symbols $P \wedge Q$, we will say "P and Q."

Example 3.3. If we have the statements:

$$P\text{: 2 is an even number,}$$

and

$$Q\text{: 5 is an even number,}$$

then the statement $P \wedge Q$ would be

$$\text{2 is an even number and 5 is an even number.}$$

This is a false statement, since Q is false. △

Definition 3.4. Let P and Q be statements. The *disjunction* of P and Q, written

$$P \vee Q,$$

is a statement that is true if P is true, or if Q is true, or if both are true; it is false otherwise.

When reading the symbols $P \vee Q$, we will say "P or Q."

Example 3.5. For the statements P and Q in Example 3.3, $P \vee Q$ would be the statement

$$\text{2 is an even number or 5 is an even number.}$$

In this case $P \vee Q$ is true, since P is true. △

Warning 3.6. If both P and Q are true, then $P \vee Q$ is true. This is not the way the word "or" is usually interpreted in the English language. For instance, if I say to my daughter "You may have an ice cream cone or you may have a candy bar," I typically do not mean that she can have both. Thus, $P \vee Q$ really means "either P or Q or both," but that is too wordy for common use, so we just say "P or Q".

Definition 3.7. Let P be a statement. The *negation* of P, written

$$\neg P,$$

is a statement that is true if P is false, and false if P is true.

The symbols $\neg P$ may be read as "not P." If we wish to express the meaning behind $\neg P$ we might say "it is not the case that P holds," or "P is false."

Example 3.8. If we have the statement

$$P\text{: 2 is an even number}$$

(which is a true statement), then $\neg P$ would be the statement

$$\text{it is not the case that 2 is an even number}$$

or, in other words,

$$\text{2 is not an even number,}$$

which is a false statement. △

Example 3.9. Let's negate the statement

$$P: \text{ It is raining in London.}$$

You probably do not know whether this statement is true or false. Nevertheless we can still negate the sentence. The negation is

$$\neg P: \text{ It is not raining in London.}$$ △

Advice 3.10. To see that the meaning of the negation of P is the same as "P is false" consider two cases.

If P is true, the statement "P is false" is then false. Hence, it has the opposite truth value as P.

If, on the other hand, P is false, then the statement "P is false" is a true statement, so again, it has the opposite truth value as P.

Often, it is easier to understand (or maybe just shorter to say) "P is false" than to say "it is not the case that P." However, inserting the phrase "it is not the case that" at the beginning of a sentence P has the advantage that it will generally form a grammatical sentence when combined with the words comprising P.

Warning 3.11. Typically, if we are asked to negate a statement, the intent is to write a statement that will have the opposite truth value to the original statement, regardless of whether the original statement is known to be true or false. For instance, if asked to negate the statement

$$P\text{: 2 is an even number}$$

(which is true), it would not be correct to write

$$\neg P\text{: The sun is blue}$$

(which is false), because the fact that these two statements have opposite truth values depends on knowing the truth values of each statement. A correct negation of the statement P would be the statement

$$\neg P\text{: 2 is not an even number.}$$

The next concept that we define is one of the most important in mathematics.

Definition 3.12. Let P and Q be statements. Then we construct a new statement, written

$$P \Rightarrow Q,$$

that is true unless P is true and Q is false.

A statement of the form $P \Rightarrow Q$ is called an *implication* or a *conditional statement*. In the implication $P \Rightarrow Q$, the statement P is called the *premise* and Q is called the *conclusion*.

We read $P \Rightarrow Q$ as either "P implies Q," or as "If P, then Q."

Example 3.13. Let P be the statement "2 is an even number" and Q be the statement "5 is an even number," as in Example 3.3. Then the statement $P \Rightarrow Q$ would be the statement

$$\text{If 2 is an even number, then 5 is an even number,}$$

which is false, since P is true and Q is false. On the other hand, the statement $Q \Rightarrow P$ would be the statement

$$\text{If 5 is an even number, then 2 is an even number,}$$

which would be true, since the premise is false. △

Remark 3.14. To better understand why implication is defined as it is, we can think of an implication as a promise, or a contract. Suppose I tell my daughter

If you clean your room, then you will get ice cream.

This is an implication, with the premise being the statement

P: You clean your room

and the conclusion being the statement

Q: You will get ice cream.

We now examine this statement to find out under what conditions I am telling the truth.

First, if my daughter cleans her room and I let her have ice cream, then I have told the truth and kept my promise. In other words, if P is true and Q is true, then $P \Rightarrow Q$ is true.

Second, if my daughter cleans her room and I do not let her have ice cream, then I have lied. In other words, if P is true and Q is false, then $P \Rightarrow Q$ is false (a lie).

Third, if my daughter does not clean her room and I let her have ice cream anyway (perhaps because she did some other duty to deserve the ice cream), then I have not lied, so my statement is true. I am under no obligation to give her ice cream, but I do so anyway. Hence, if P is false and Q is true, then $P \Rightarrow Q$ is true.

Finally, if my daughter does not clean her room and I do not let her have ice cream, then I have not lied; she did not fulfill the condition of the implication, so I did not fulfill the conclusion. If P is false and Q is false, the implication $P \Rightarrow Q$ is true. ▲

Warning 3.15. Many students have difficulty with the idea that the statement $P \Rightarrow Q$ should be true when P is false and Q is true. It can help to think of $P \Rightarrow Q$ as meaning "Whenever P is true, Q must also be true, but if P is false, anything can happen."

We define one final way to combine statements, called the biconditional, below.

Definition 3.16. Let P and Q be statements. Then the *biconditional*

$$P \Leftrightarrow Q$$

is a statement that is true if P and Q have the same truth value, and false otherwise.

We read $P \Leftrightarrow Q$ as "P if and only if Q." We can also think of it as meaning that P is true exactly when Q is true.

Example 3.17. Suppose P is the statement "2 is an even integer" and Q is the statement "5 is an even integer." Then $P \Leftrightarrow Q$ is the statement

2 is an even integer if and only if 5 is an even integer.

This is a false statement, since P is true and Q is false. △

In a sense that we will see later, $P \Leftrightarrow Q$ is the same as saying that $P \Rightarrow Q$ and $Q \Rightarrow P$.

We have discussed a number of ways to combine and modify statements to form more complex statements. We define a term to encompass all of these.

Definition 3.18. A *logical connective* is an operation that modifies or combines statements into more complex statements.

Examples of logical connectives are $\wedge$, $\vee$, $\neg$, $\Rightarrow$, and $\Leftrightarrow$. (Other connectives exist, such as "exclusive or.") We summarize the definitions of these symbols in Table 3.19. In this table, we have a column for each of P, Q, $\neg P$, $P \wedge Q$, $P \vee Q$, $P \Rightarrow Q$, and $P \Leftrightarrow Q$. We have a row for each possible combination of truth values of P and Q, and the entries in each row indicate the truth value of the statement at the top of the column.

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \Rightarrow Q$	$P \Leftrightarrow Q$
T	T	F	T	T	T	T
T	F	F	F	T	F	F
F	T	T	F	T	T	F
F	F	T	F	F	T	T

Table 3.19: Truth values of common compound statements

Extremely complicated statements can be built by using multiple logical connectives. For instance, consider

$$((P \vee Q) \wedge (R \vee \neg(S))) \Rightarrow (P \wedge R).$$

The following definition gives a name to all such statements.

Definition 3.20. A *compound statement* is a statement that has been built by applying at least one logical connective to one or more statements.

Example 3.21. Let P, Q, R, and S be statements. Then the statement

$$(P \wedge (Q \Rightarrow R)) \vee (\neg(P \Rightarrow (R \wedge S)))$$

is a compound statement. We call P, Q, R, and S the *components* of the compound statement. △

Example 3.22. The statement "2 is even and 5 is even" is a compound statement, $P \wedge Q$, with components

$$P\text{: 2 is even}$$

and

$$Q\text{: 5 is even.}$$ △

3.C Logical equivalences

In some cases, we will wish to show that two compound statements are essentially the same. We make this notion precise with the following definition.

Definition 3.23. Two compound statements are *logically equivalent* if they have the same truth value regardless of the truth values of the components. If R and S are compound statements that are logically equivalent, we write $R \equiv S$.

The following example will give two compound statements that are logically equivalent, and we show how to prove this equivalence.

Example 3.24. Let P and Q be statements. Then the statements

$$R:\ \neg(P \wedge Q)$$

and

$$S:\ (\neg P) \vee (\neg Q)$$

are compound statements with components P and Q. In Table 3.25 below, we make a truth table that indicates the truth values of these two compound statements for every possible combination of truth values of P and Q.

P	Q	$P \wedge Q$	$\neg(P \wedge Q)$	$\neg P$	$\neg Q$	$(\neg P) \vee (\neg Q)$
T	T	T	**F**	F	F	**F**
T	F	F	**T**	F	T	**T**
F	T	F	**T**	T	F	**T**
F	F	F	**T**	T	T	**T**

Table 3.25: Truth table showing equivalence of $\neg(P \wedge Q)$ and $(\neg P) \vee (\neg Q)$

We notice immediately that the two boldface columns have identical truth values. This shows that the statements R and S are logically equivalent, or, in symbols, that $\neg(P \wedge Q) \equiv (\neg P) \vee (\neg Q)$. △

When constructing a truth table, it is useful to organize the rows in such a way that you can be sure that all possible combinations of the components are in fact represented. If there are n components there will be 2^n rows. One convenient way to organize them is to write each component as a label for a column of the truth table

(just as we did in Tables 3.19 and 3.25). For the last component (the rightmost one), put alternating truth values of T and F in the column, until there are 2^n of them.

For the second to the last component, fill the corresponding column with alternating blocks of two Ts and Fs, until all the entries have been filled. Continue from right to left, doubling the sizes of the blocks, until for the leftmost component, you write just one block of Ts and one block of Fs.

In this way, you can (in a systematic way) be certain that all possible combinations of truth values are written. An example of how this is to be done with three components is given below in Example 3.27.

Note that a truth table can have as many columns as is convenient to work out the truth values of the statements in which we are interested; typically, we will include a column for each intermediate step.

A number of logical equivalences are important enough to be given standard names.

Theorem 3.26. *Let P, Q, and R be statements.*
(1) $\neg(\neg P) \equiv P$ (*Double negation*),
(2) $\neg(P \wedge Q) \equiv (\neg P) \vee (\neg Q)$ (*De Morgan's law*),
(3) $\neg(P \vee Q) \equiv (\neg P) \wedge (\neg Q)$ (*De Morgan's law*),
(4) $P \Rightarrow Q \equiv (\neg Q) \Rightarrow (\neg P)$ (*Contrapositive*),
(5) $P \wedge (Q \vee R) \equiv (P \wedge Q) \vee (P \wedge R)$ (*Distributivity*),
(6) $P \vee (Q \wedge R) \equiv (P \vee Q) \wedge (P \vee R)$ (*Distributivity*),
(7) $P \wedge (Q \wedge R) \equiv (P \wedge Q) \wedge R$ (*Associativity*),
(8) $P \vee (Q \vee R) \equiv (P \vee Q) \vee R$ (*Associativity*),
(9) $P \wedge Q \equiv Q \wedge P$ (*Commutativity*),
(10) $P \vee Q \equiv Q \vee P$ (*Commutativity*).

Each of the logical equivalences in Theorem 3.26 can be demonstrated by constructing a truth table. For example, (2) is proved in Table 3.25.

Example 3.27. We prove 3.26(7) by forming the appropriate truth table.

P	Q	R	$P \wedge Q$	$(P \wedge Q) \wedge R$	$Q \wedge R$	$P \wedge (Q \wedge R)$
T	T	T	T	**T**	T	**T**
T	T	F	T	**F**	F	**F**
T	F	T	F	**F**	F	**F**
T	F	F	F	**F**	F	**F**
F	T	T	F	**F**	T	**F**
F	T	F	F	**F**	F	**F**
F	F	T	F	**F**	F	**F**
F	F	F	F	**F**	F	**F**

Because the two boldface columns match, the corresponding statements are logically equivalent. △

3.D Tautologies and contradictions

Sometimes the truth value of a compound statement does not depend on the truth value of the components from which it is composed. We give such compound statements special names.

Definition 3.29. A compound statement that is false for every possible combination of truth values of its components is called a *contradiction*.

Example 3.30. Let P be a statement. The compound statement $P \wedge (\neg P)$ is then false, regardless of whether P is true or false. (Try it! Form the truth table to see that this statement is always false. The truth table should have two rows.) △

Definition 3.31. A compound statement that is true for every possible combination of truth values of its components is called a *tautology*.

Example 3.32. Let P be a statement. Then the compound statement $P \Rightarrow P$ is a tautology. (Try it!) △

Example 3.33. Let P and Q be statements. The compound statement

$$(P \wedge (P \Rightarrow Q)) \Rightarrow Q$$

is a tautology. This can be seen from the following truth table.

P	Q	$P \Rightarrow Q$	$P \wedge (P \Rightarrow Q)$	$(P \wedge (P \Rightarrow Q)) \Rightarrow Q$
T	T	T	T	T
T	F	F	F	T
F	T	T	F	T
F	F	T	F	T

This particular tautology is so useful it has a name: *modus ponens*. △

Remark 3.34. All contradictions are logically equivalent to each other. Similarly, all tautologies are logically equivalent to each other. (Can you see why?) ▲

3.E Exercises

Exercise 3.1. Determine whether the given sentence is a statement. If it is, indicate its truth value (if you can).

(a) The number 0 is an even integer.
(b) Let $x = 2$.
(c) If 2 is an even integer.
(d) Either 2 is even or 4 is odd.
(e) George Washington had seven children.
(f) There are 7254 different species of ants in the United States.

Exercise 3.2. Let P, Q, and R be statements. Construct a truth table showing the possible truth values for each of the following compound statements.

(a) $(P \wedge Q) \Rightarrow P$.
(b) $P \Rightarrow (P \vee Q)$.
(c) $\neg(P \Rightarrow Q) \wedge (\neg P)$.
(d) $(P \vee Q) \wedge R$.
(e) $P \vee (Q \wedge R)$.

Exercise 3.3. Use truth tables to prove the given logical equivalences.

(a) $\neg(P \wedge Q) \equiv (\neg P) \vee (\neg Q)$.
(b) $P \Rightarrow Q \equiv (\neg P) \vee Q$.
(c) $(P \vee Q) \Rightarrow R \equiv (P \Rightarrow R) \wedge (Q \Rightarrow R)$

Exercise 3.4. Determine whether the two given compound statements are logically equivalent.

(a) $\neg(P \Rightarrow Q)$ and $P \wedge (\neg Q)$.
(b) $(P \wedge Q) \Rightarrow R$ and $P \Rightarrow (\neg Q \vee R)$.
(c) $P \Rightarrow (Q \vee R)$ and $(P \wedge \neg Q) \Rightarrow R$.
(d) $(P \vee Q) \Rightarrow R$ and $(\neg R \wedge P) \Rightarrow (\neg Q)$.
(e) $P \Leftrightarrow Q$ and $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$.

Exercise 3.5. Let P, Q, and R be statements. Identify each of the following statements as a tautology, a contradiction, or neither.

(a) $((P \Rightarrow Q) \wedge (\neg Q)) \Rightarrow (\neg P)$.
(b) $((P \vee Q) \wedge (\neg P)) \Rightarrow Q$.
(c) $(P \Rightarrow Q) \Rightarrow (P \Rightarrow R)$.
(d) $((\neg Q) \Rightarrow (\neg P)) \wedge P \wedge (\neg Q)$.

Exercise 3.6. Let P and Q be statements.

(a) Prove that the compound statement $P \Rightarrow Q$ is *not* logically equivalent to $Q \Rightarrow P$.
(b) The statement $Q \Rightarrow P$ is called the *converse* of $P \Rightarrow Q$. Give an example of statements P and Q for which $P \Rightarrow Q$ is true, but $Q \Rightarrow P$ is false. For the specific statements P and Q in your example, state their truth values.

Exercise 3.7. The following exercise shows that the biconditional can be a useful logical connective (if you are ever lost in grue infested woods):

After getting lost in the woods, you stumble upon a path. As you follow the path it comes to a fork, and a grue blocks your way. There is a sign which reads:

> This grue either always tells the truth or always lies. You may ask the grue a single question. Any more than that, and it will eat you.

You are positive that one of the paths leads home and the other leads to certain death. Explain why asking the grue the question

> Is the statement "the left path is the way home if and only if you are a truth teller" true?

can help you decide which path to take. (Note that the question is **not** "Is the left path the way home?" Also note that the Grue's answer might be a lie.)

4 Open sentences

4.A Open sentences

In many cases, we wish to write a sentence whose truth value depends on *variables*, where the variables can take on many different values. Typically, we want each variable to take on values in some specific set. We call each such set of values the *domain* of the corresponding variable.

Definition 4.1. An *open sentence* is a declarative sentence containing one or more variables such that when all of the variables are assigned values in their respective domains, the resulting sentence has a truth value.

Some authors do not explicitly write all open variables that occur in an open sentence. We will always specify all open variables.

Example 4.2. Let x and y be variables, both with the same domain $\mathbb{R}$. We may define the open sentence

$$P(x, y)\colon\ x > y.$$

With this definition we see that $P(2, 3)$ is false (since the statement $2 > 3$ is false), $P(5, 2)$ is true, and so on. Note that if we try to plug in a value of x or y outside the domain, the resulting sentence may be meaningless. For instance $P(\text{apples}, \text{oranges})$ is meaningless since the sentence "apples $>$ oranges" makes no sense (you can't compare apples and oranges). △

Open sentences are very common throughout mathematics. They can express quite important and general principles. Here are some examples.

Example 4.3. Let S be the set of triangles and let x be a variable with domain S. Then we have an open sentence

$$P(x)\colon \text{ The sum of the (interior) angles of } x \text{ is } 180^\circ.$$

The open sentence $P(x)$ happens to be true for all choices of $x \in S$. This fact is a very famous (and very old) theorem.

Again taking x to be a variable with domain S, we may write the open sentence

$$Q(x)\colon \text{ One of the angles of } x \text{ is a right angle.}$$

In this case, $Q(x)$ is not true for all $x \in S$; instead it is true exactly when x is a right triangle. △

Warning 4.4. An open sentence is much like a function from calculus. It can be considered to be a rule in which, when we plug in a value for the variable (or variables), we obtain a truth value. Even if an open sentence is true for each value of the variable, it does not become a statement. For instance, if we let x be a variable with domain equal to the real numbers, the open sentence

$$P(x)\text{: } x^2 > -1$$

is true for all possible values of x. Nevertheless, it is not a statement. It is still a rule that assigns a truth value to every real number. Compare this to the constant function defined by the rule

$$f(x) = 2$$

from algebra. The function f is not equal to the number 2; rather it is a rule that associates to each value of x the number 2.

We can use logical connectives to combine open sentences much as we combine statements. The result of combining open sentences with any of the logical connectives $\wedge$, $\vee$, $\neg$, $\Rightarrow$, or $\Leftrightarrow$ is again an open sentence. This open sentence obtains a truth value when all the variables in the component open sentences are given values. Once these values are fixed, each of the component open sentences becomes a statement, and the truth value of the compound sentence can be evaluated from the truth values of the components.

Example 4.5. Let x be a variable with domain the set of all triangles, and define the following open sentences:

$$P(x)\text{: The sum of the angles of } x \text{ is } 180^\circ$$

and

$$Q(x)\text{: One of the angles of } x \text{ is a right angle.}$$

Then, if we plug in a specific right triangle t in place of x, the statement $P(t) \wedge Q(t)$ is true; however, if we take x to be an equilateral triangle e (so that it cannot be a right triangle), the statement $P(e) \wedge Q(e)$ is false (since $Q(e)$ is false). $\triangle$

Example 4.6. Let x be a variable with domain the set of real numbers. We define two open sentences:

$$P(x)\text{: } x > 3, \quad Q(x)\text{: } x < 5.$$

Consider the following compound open sentences:

(1) $P(x) \wedge Q(x)$,
(2) $P(x) \vee Q(x)$,
(3) $P(x) \Rightarrow Q(x)$,
(4) $P(x) \Leftrightarrow Q(x)$,
(5) $\neg P(x) \wedge \neg Q(x)$.

Below, we write the set of places these statements are true. Try it for yourself before looking at the answers.

We will work out (3) in detail. Remember that for a given value of x, we have

$$P(x) \Rightarrow Q(x) \equiv \neg P(x) \vee Q(x).$$

We know that $\neg P(x)$ is true for $x \in (-\infty, 3]$. Also $Q(x)$ is true for $x \in (-\infty, 5)$. Thus, $\neg P(x) \vee Q(x)$ is true for $x \in (-\infty, 3] \cup (-\infty, 5) = (-\infty, 5)$.

All Answers:
(1) $(3, 5)$,
(2) $\mathbb{R}$,
(3) $(-\infty, 5)$,
(4) $(3, 5)$,
(5) $\emptyset$. △

Remark 4.7. The previous example demonstrates why logical equivalence is so useful. We know that

$$P(x) \Rightarrow Q(x) \equiv \neg P(x) \vee Q(x)$$

even before we know the value of x, hence before we know the truth value of $P(x)$ or $Q(x)$. Thus, we can simplify some compound open sentences, before we even know what the components say. ▲

4.B Quantifiers

We have seen the expression "for all $x \in S$" several times. This expression is common enough that we create a symbol for it.

Definition 4.8. Let x be a variable with domain S and let $P(x)$ be an open sentence. The expression

$$\forall x \in S,\ P(x)$$

is then a statement. It is true if $P(x)$ is true for each choice of $x \in S$, and it is false if $P(x)$ is false for one (or more) choice of $x \in S$.

We read the statement $\forall x \in S,\ P(x)$ as "for all x in S, $P(x)$."

Example 4.9. Let S be the set of all triangles and let x be a variable with domain S. Let $P(x)$ be the open sentence

$$P(x)\text{: The sum of the angles of } x \text{ is } 180^\circ.$$

Then the statement "$\forall x \in S,\ P(x)$" means "For all x in the set of triangles, the sum of the angles of x is 180°." We may simplify this a bit by reducing it to the sentence "The sum of the angles of any triangle is 180°." without changing the meaning. We note that, in this case, the statement $\forall x \in S,\ P(x)$ is a true statement. △

Example 4.10. Again letting x be a variable with domain S equal to the set of all triangles, we will let $Q(x)$ be the open sentence

$$Q(x)\text{: One of the angles of } x \text{ is a right angle.}$$

If we examine the statement $\forall x \in S,\ Q(x)$, we find that it is false, since it is not the case that all triangles contain a right angle; for example, none of the angles in an equilateral triangle are right angles. However, if we let T be the subset of S consisting of right triangles, then the statement $\forall x \in T,\ Q(x)$ is a true statement. △

We call the symbol $\forall$ the *universal quantifier*; the expression $\forall x \in S,\ P(x)$ makes the assertion that the open sentence $P(x)$ is universally true for any x in the domain S. There is another quantifier, $\exists$, called the *existential quantifier*.

Definition 4.11. Let x be a variable with domain S and let $P(x)$ be an open sentence. The expression

$$\exists x \in S,\ P(x)$$

is then a statement. It is true if $P(x)$ is true for at least one choice of $x \in S$, and it is false if $P(x)$ is false for each choice of $x \in S$.

The statement $\exists x \in S,\ P(x)$ reads as "There exists some x in S such that $P(x)$."

Example 4.12. Let x be a variable with domain $\mathbb{R}$ and let $P(x)$ be the open sentence

$$P(x)\text{: } x^2 = 2.$$

Then the statement $\exists x \in \mathbb{R},\ P(x)$ is a true statement, because there is at least one choice of $x \in \mathbb{R}$, namely $x = \sqrt{2}$, for which $P(x)$ is a true statement (in fact there are two choices, since we could take $x = -\sqrt{2}$ as well). However, the statement $\exists x \in \mathbb{Z},\ P(x)$ is false, since there are no integers x such that $P(x)$ is a true statement. △

Example 4.13. Let A and B be sets and let x be a variable with domain A. We will use quantifiers to describe some relationships between the sets A and B.

Suppose that the statement

$$\forall x \in A,\ x \in B$$

is true. This means that every element of A is an element of B. This is what it means to say that $A \subseteq B$.

Suppose that the statement

$$\forall x \in A,\ x \notin B$$

is true. In other words, every element of A is *not* an element of B. Then we know that A and B have no elements in common; they are disjoint. In other words, $A \cap B = \emptyset$.

Suppose that the statement

$$\exists x \in A,\ x \in B$$

is true. In other words, there is at least one element in A that is also in B. This tells us that $A \cap B \neq \emptyset$; so A and B are not disjoint. △

Note that, in the case of open sentences with *multiple* variables, we can define quantification in a way similar to Definitions 4.8 and 4.11. In this case, we may need multiple quantifiers in order to turn the open sentence into a statement. In general, one quantifier will be needed for each distinct variable in the open sentence. As an example, if we let x and y be variables with the domain $\mathbb{R}$ and let $P(x, y)$ be the open sentence

$$P(x, y) : \ x > y,$$

then the sentence $\forall x \in \mathbb{R},\ P(x, y)$ is not a statement, because it still has a variable that has not been specified or quantified (namely y). In order to make it a statement we need to evaluate y; we could say $\forall y \in \mathbb{R}, \forall x \in \mathbb{R},\ P(x, y)$, which means that

For all real numbers y and for all real numbers x, it holds that $x > y$.

Note that this is a false statement, since taking $y = 2$ and $x = 1$ gives $1 > 2$.

Example 4.14. Suppose we wish to express the statement "Every even integer is a sum of two odd integers" in symbolic logic. We can do this using multiple quantifiers. Let Even denote the set of even integers and let Odd denote the set of odd integers. The statement

$$\forall x \in \text{Even}, \exists y \in \text{Odd}, \exists z \in \text{Odd},\ x = y + z$$

then means that for each even integer x, there is an odd integer y and there is an odd integer z such that $x = y + z$. In other words, any even integer x is the sum of two odd integers, y and z. (This is true.) △

Example 4.15. Suppose we wish to express the statement "Every positive real number has a positive square root" in symbolic logic. Let $\mathbb{R}_{>0}$ be the set of positive real numbers. We might write

$$\forall x \in \mathbb{R}_{>0}, \exists y \in \mathbb{R}_{>0},\ x = y^2.$$

Interpreted, this means that for each x in the positive real numbers, there exists a positive real number y, such that $x = y^2$. In other words, y is a positive real square root of x. △

We will return to a much more detailed discussion of multiple quantifiers in the next section. They are quite important and appear in many parts of mathematics.

4.C Implication and open sentences

If we have two open sentences $P(x)$ and $Q(x)$, and we join them with $\Rightarrow$, we have seen that the resulting sentence $P(x) \Rightarrow Q(x)$ is again an open sentence. In order to make a statement of it, we may either plug in a value of x from the domain or use one of the two quantifiers $\forall$ or $\exists$.

Statements of the form

$$\forall x \in S,\ P(x) \Rightarrow Q(x)$$

are so commonplace that there are multiple ways to express them. Here are some common ways this is done.

- Given $x \in S$, if $P(x)$, then $Q(x)$.
- Let $x \in S$. If $P(x)$, then $Q(x)$.
- If x is an arbitrary element of S satisfying $P(x)$, then $Q(x)$.
- If x is an element of S satisfying $P(x)$, then $Q(x)$.
- If $P(x)$, then $Q(x)$.

The last bulleted option can only be used when the domain S of x is understood. However, it could be confused with the (unquantified) open sentence

$$P(x) \Rightarrow Q(x),$$

and so in general it is best to quantify statements in some way. (Similar considerations hold for the biconditional.) It is best practice to always use some word signifying the universal scope of the quantifier $\forall$, such as: "arbitrary", "any", "every", "all", "given", and "let" (as in the above bulleted list). (The words we associate with $\exists$ are: "for some", "exists", "fix", etc.)

If we talk about the *open sentence* $P(x) \Rightarrow Q(x)$, it is clear that no quantifier is intended (since the quantified version of the sentence would be a *statement* not an open sentence).

Example 4.16. Let $P(x)$ be the open sentence

$$P(x)\text{: } x \text{ is even}$$

and let $Q(x)$ be the open sentence

$$Q(x)\text{: } x - 2 \text{ is even}$$

(with the domain of x being the integers in both cases). If we wish to write the quantified statement

$$\forall x \in \mathbb{Z},\ P(x) \Rightarrow Q(x),$$

we may write the following:

$$\text{If an (arbitrary) integer } x \text{ is even, then } x - 2 \text{ is even.}$$

Notice that no explicit quantifier is stated, but the quantifier is nevertheless understood.

Similarly the biconditional

$$\text{An integer } x \text{ is even if and only if } x - 2 \text{ is even}$$

means

$$\forall x \in \mathbb{Z},\ P(x) \Leftrightarrow Q(x),$$

even though we didn't use the words "for all $x \in \mathbb{Z}$." △

Example 4.17. Inclusion of sets can be understood through implication.

Let A and B be subsets of a universal set U. Let x be a variable with domain U, and assume that

$$\forall x \in U,\ x \in A \Rightarrow x \in B$$

is true. Since the implication must be true for all elements of U, we see that whenever $x \in A$, it must be the case that $x \in B$. (When $x \notin A$, we know nothing about whether it is in B.) This is exactly what it means to say that $A \subseteq B$. Hence, the statement $A \subseteq B$ can be interpreted (in terms of symbolic logic) as the statement above.

Occasionally the universal set U is understood from context, and we will write $\forall x,\ x \in A \Rightarrow x \in B$. More simply, $A \subseteq B$ could be written as $\forall x \in A,\ x \in B$, without the use of implication or a universal set U. △

4.D The meaning of implication

We now discuss what it means for a statement of the form

$$\forall x \in S,\ P(x) \Rightarrow Q(x)$$

to be true.

We first examine how it could be false. We see that the only way for it to be false is for there to be one or more values of x for which $P(x)$ is true and $Q(x)$ is false. Therefore, in order for the statement $\forall x \in S,\ P(x) \Rightarrow Q(x)$ to be true, it must be the case that for each $x \in S$, either $P(x)$ is false or $Q(x)$ is true. Another way to put this is that whenever $P(x)$ is true, it must happen that $Q(x)$ is also true. In other words, if we know that $P(x)$ is true, then $Q(x)$ must be true. Hence, for any given $x \in S$ we have "If $P(x)$, then $Q(x)$."

There are a number of different ways of writing a sentence with the meaning "If P, then Q." Some of them are:

If P, then Q.
P implies Q.
P only if Q.
Q if P.
P is sufficient for Q.
Q is necessary for P.
Whenever P is true, then Q is true.

For example, if we say "P is sufficient for Q," this has the meaning that P being true is sufficient for us to conclude that Q is true. In other words, "If P is true, then Q is true."

You should think about the meaning of the other phrases above, and satisfy yourself that they all have the same meaning as $P \Rightarrow Q$.

Another sentence that can be expressed in several different ways in English is $P \Leftrightarrow Q$. Some of the ways that it can be expressed are:

P if and only if Q.
P is equivalent to Q.
P is necessary and sufficient for Q.
P holds exactly when Q holds.

4.E Translating between English and symbolic logic

We now give some examples of translating statements from English to symbolic logic.

Example 4.18. The statement "If a real number is not zero, then its square is not zero" can be written as

$$\forall x \in \mathbb{R},\ x \neq 0 \Rightarrow x^2 \neq 0.$$

Another way of interpreting this symbolic logic statement in English would be "Every nonzero real number has a nonzero square" or "A nonzero real number has a nonzero square." △

Example 4.19. The statement "The square of any rational number is a rational number" might be written

$$\forall x \in \mathbb{Q},\ x^2 \in \mathbb{Q}.$$ △

Example 4.20. The statement "There is a rational solution of the equation $x^2 + 2x + 1 = 0$" can be translated into symbolic logic as

$$\exists x \in \mathbb{Q},\ x^2 + 2x + 1 = 0.$$ △

Example 4.21. One way of translating "Every real solution of the equation $x^3 - x = 0$ is rational" into symbolic logic is as

$$\forall x \in \mathbb{R},\ x^3 - x = 0 \Rightarrow x \in \mathbb{Q}.$$ △

Example 4.22. We can translate the statement

$$\forall x \in \mathbb{R},\ x^2 = 2 \Rightarrow x \notin \mathbb{Q}$$

into English as "Every real number whose square is 2 is not rational." Alternatively, we could say "Given an arbitrary real number x, if $x^2 = 2$, then x is not rational." △

Example 4.23. The statement "A real number x has a real square root if it is positive" might be interpreted in symbolic logic as

$$\forall x \in \mathbb{R},\ x \text{ is positive} \Rightarrow x \text{ has a real square root.}$$

Or, we could be even more detailed, and write

$$\forall x \in \mathbb{R},\ (x > 0 \Rightarrow \exists y \in \mathbb{R},\ x = y^2),$$

which includes an interpretation of the phrase "has a real square root." △

Example 4.24. In order to write the statement "Every even integer greater than 3 can be written as a sum of two primes" in symbolic logic, we define the set $P = \{\text{prime numbers in } \mathbb{N}\}$. Then we can write

$$\forall x \in \mathbb{Z},\ (((x \text{ is even}) \wedge (x > 3)) \Rightarrow (\exists y \in P, \exists z \in P,\ x = y + z)).$$

Alternatively, we could let E be the set of even integers, and write the statement as

$$\forall x \in E,\ (x > 3 \Rightarrow \exists y \in P, \exists z \in P,\ x = y + z).$$

Another alternative would be to define $F = \{x \in \mathbb{Z} : x > 3 \text{ and } x \text{ is even}\}$ to be the set of even integers greater than 3, and write

$$\forall x \in F, \exists y \in P, \exists z \in P,\ x = y + z.$$ △

4.F Exercises

Exercise 4.1. Let x be a variable with domain $\mathbb{Z}$. Define the open sentences

$$P(x) : x > 1,$$
$$Q(x) : x^2 < 16, \text{ and}$$
$$R(x) : x + 1 \text{ is even.}$$

For each of the following compound open sentences, describe the subset of $\mathbb{Z}$ (by listing its elements or by using set-builder notation without mentioning the symbols P, Q, and R) where that open sentence is true.

(a) $P(x) \wedge Q(x)$.
(b) $Q(x) \wedge R(x)$.
(c) $(Q(x) \vee \neg P(x)) \wedge \neg R(x)$.
(d) $(P(x) \Rightarrow Q(x)) \Rightarrow R(x)$. (Hint: Simplify using logical equivalences.)

Exercise 4.2. For each $x \in \{1, 2, 3, 4, 5, 6\}$, write down the truth value of

$$P(x)\text{: If } x \text{ is an odd integer, then } \frac{3x+5}{2} \text{ is an odd integer,}$$

and then state whether you believe $\forall x \in \mathbb{Z},\ P(x)$ is true or false.

Exercise 4.3. For each $x \in \{1, 2, 3, 4, 5, 6\}$, write down the truth value of

$$Q(x)\text{: If } x \text{ is an even integer, then } 3x + 5 \text{ is an odd integer,}$$

and then state whether you believe $\forall x \in \mathbb{Z},\ Q(x)$ is true or false.

Exercise 4.4. Let A and B be two subsets of a universal set U. Write a symbolic logic interpretation of the statement $A = B$. Explicitly write out any quantifiers involved in the statement. (It is possible to do this with no reference to U.)

Exercise 4.5. Translate the following English sentences into symbolic logic. Explicitly write any quantifiers that are implied.

(a) There is an integer strictly between 4 and 6.
(b) The square of any odd integer is odd.
(c) If the square of an integer is odd, then the original integer is odd.
(d) If a real number is not rational, then it is not equal to 0.
(e) The sum of two rational numbers is rational.
(f) The square of any real number is a nonnegative real number.
(g) There is an integer solution to the equation $x^2 - 5x + 6 = 0$.
(h) Every real solution to $x^2 - 5x + 6 = 0$ is an integer.

Exercise 4.6. Translate the following symbolic logic statements into English.

(a) $\exists x \in \mathbb{R},\ x^2 = 2$.
(b) $\forall x \in \mathbb{Z},\ (x \text{ is even}) \Leftrightarrow (x^2 \text{ is even})$.
(c) $\forall x \in \mathbb{R},\ (x > 1) \Rightarrow (x^3 > 1)$.
(d) $\forall x \in \mathbb{R},\ (x^2 - 2x + 1 = 0) \Rightarrow (x = 1)$.
(e) $\exists x \in \mathbb{Q},\ 2x^3 - x^2 + 2x - 1 = 0$.
(f) $\forall x \in \mathbb{R}, \exists y \in \mathbb{Z}, \exists z \in [0, 1),\ x = y + z$.

5 Multiple quantifiers and negating sentences

5.A Statements with multiple quantifiers

Statements involving multiple quantifiers are quite common in mathematics. For such statements, it is important to understand how the quantifiers interact and to be able to analyze the truth values of the statements. We begin with two examples that demonstrate that the order in which quantifiers occur is important.

Example 5.1. Let $P(x, y)$ be the open sentence $x > y$ with the domain of both x and y being the real numbers. We examine the quantified statement

$$\forall x \in \mathbb{R}, \exists y \in \mathbb{R},\ P(x, y).$$

This statement means:

For each real number x, there is a real number y such that $x > y$,

or in other words, "For each real number x, there is a real number y (possibly depending on x) that is smaller than x." We claim that this statement is true. To see this, suppose that x is any real number. The statement asserts that no matter what x is, there is some real number y such that $x > y$. If we let $y = x - 1$, we see that regardless of what the value of x is, $x > y$. Hence, for each real number x, there is some y (for example $y = x - 1$) with $x > y$. △

Example 5.2. Let $P(x, y)$ again be the open sentence $x > y$ with the domain of both x and y being the real numbers. We next examine the quantified statement

$$\exists y \in \mathbb{R}, \forall x \in \mathbb{R},\ P(x, y).$$

Note that this statement is the same as the statement in Example 5.1, except for the order of the quantifiers. In words, this statement means

There is some real number y such that $x > y$ whenever x is a real number,

or in other words, "there is a real number y such that every real number is larger than y."

This statement is false, which we see as follows. If y is a real number, then $y - 1$ is not larger than y. Hence, there does not exist a real number y for which every real number is larger than y. △

Warning 5.3. Existentially quantified variables are always allowed to depend on any previously quantified variables. For instance, in Example 5.1 we are allowed to choose a value of y in terms of x, as we did by taking $y = x - 1$, since the variable x was quantified before y. However, a variable can never depend on another variable that is quantified later. For instance, in Example 5.2 we cannot define y in terms of x, since x is quantified later than y.

These two examples demonstrate that the order of quantifiers in a statement with multiple quantifiers can change the truth value of the statement. This only happens when the quantifiers are not the same type. Changing the order of existential quantifiers that are next to each other, or universal quantifiers that are next to each other, will not change the truth value of the statement. For example,

$$\forall x \in \mathbb{R}, \forall y \in \mathbb{R},\ x > y$$

has the same truth value as

$$\forall y \in \mathbb{R}, \forall x \in \mathbb{R},\ x > y$$

(in this case both are false). Since the order does not matter in this case we will abbreviate either of these statements as

$$\forall x, y \in \mathbb{R}, x > y,$$

which we can read as "for all x and y in $\mathbb{R}$, we have $x > y$."

Similarly,

$$\exists x \in \mathbb{R}, \exists y \in \mathbb{R},\ x > y$$

has the same truth value as

$$\exists y \in \mathbb{R}, \exists x \in \mathbb{R},\ x > y$$

(in this case both are true). Since the order does not matter in this case, we often abbreviate either of these statements as

$$\exists x, y \in \mathbb{R},\ x > y,$$

which we might read as "there are real numbers x and y such that $x > y$."

Remark 5.4. Sometimes, when the domain of a variable is well understood and we wish to quantify a statement over a subset of the domain that is defined by some easy mathematical condition, we might adapt the notation that we use for a quantifier. For instance, if x and y are variables with domain $\mathbb{R}$ and we wish to express the statement "Every positive real number has a positive real square root," we might write

$$\forall x > 0, \exists y > 0,\ x = y^2,$$

which is shorthand for

$$\forall x \in \mathbb{R}_{>0}, \exists y \in \mathbb{R}_{>0},\ x = y^2.$$

This modification of notation can only be used when the domain of a variable is explicitly stated, or completely clear from context. One common example of the use of this notation will occur in the definition of a limit in Chapter IX, where we will let ε, δ, and x be real variables, and define

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

to mean

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x \in \mathbb{R},\ 0 < |x - a| < \delta \Rightarrow |f(x) - L| < \varepsilon.$$

(Note that most authors remove $\forall x \in \mathbb{R}$, leaving it to the reader to fill in that gap.) ▲

5.B Negating statements

It is often necessary to understand both how to negate sentences and what those negations mean. For example, to qualify for President of the United States you must be a natural born citizen and be over the age of 35. Using these criteria, what disqualifies one from being president? One is disqualified either by not being a natural born citizen or being younger than 35. Notice that we have used the logical equivalence $\neg(P \wedge Q) \equiv \neg P \vee \neg Q$ (De Morgan's law).

Negating a statement can be done by placing the words "it is not the case that" in front of the statement. Alternatively, as we have seen, the negation of a statement (or open sentence) P can be interpreted as "P is false".

However, when negating compound sentences, it is often possible to make additional simplifications or modifications. For example, if we wish to negate the statement $P \wedge Q$ we could write the result as $\neg(P \wedge Q)$. By using De Morgan's law, we see that

$$\neg(P \wedge Q) \equiv (\neg P) \vee (\neg Q).$$

Hence, we can say that the negation of the statement $P \wedge Q$ is also (equivalent to) the statement $(\neg P) \vee (\neg Q)$, which is often easier to work with.

Example 5.5. If we wish to negate the statement $P \Rightarrow Q$ (where P and Q are statements), it is helpful to recall (see Exercise 3.3(b)) that

$$P \Rightarrow Q \equiv (\neg P) \vee Q.$$

Hence,

$$\neg(P \Rightarrow Q) \equiv \neg((\neg P) \vee Q) \equiv (\neg\neg P) \wedge (\neg Q) \equiv P \wedge (\neg Q),$$

where we have used De Morgan's law and double negation.

We can interpret this string of equivalences as follows. Saying "P implies Q is false" is the same as asserting "P is true and Q is false." △

Negating statements with quantifiers is quite common. Suppose that we wish to negate a statement of the form

$$R: \ \forall x \in S, \ P(x),$$

where $P(x)$ is an open sentence, and x has domain S. The negation is

$$\text{It is not the case that } P(x) \text{ is always true.}$$

What does this mean? Simply, that $P(x)$ must be false sometime (i.e., for at least one $x \in S$). More formally,

$$\neg R: \ \exists x \in S, \ \neg P(x).$$

Similar reasoning can be used to see that the negation of $\exists x \in S, \ P(x)$ is the statement $\forall x \in S, \ \neg P(x)$. We state our conclusions as an axiom.

Axiom 5.6. Let $P(x)$ be an open sentence, where x has domain S. Then the negation of the statement

$$\forall x \in S,\ P(x)$$

is equivalent to the statement

$$\exists x \in S,\ \neg P(x).$$

Similarly, the negation of the statement

$$\exists x \in S,\ P(x)$$

is equivalent to the statement

$$\forall x \in S,\ \neg P(x).$$

Example 5.7. Suppose that we wish to negate the statement "All integers are positive." We can write this statement with a quantifier as

$$P:\ \forall x \in \mathbb{Z},\ x > 0.$$

According to the axiom, the negation of this statement is (equivalent to)

$$\neg P:\ \exists x \in \mathbb{Z},\ \neg(x > 0).$$

In words, we could state this as "There exists an integer that is not positive," or "Some integer is not positive." We note that the negation of the open sentence $x > 0$ is easily seen to be $x \leq 0$. Hence, we could write $\neg P:\ \exists x \in \mathbb{Z},\ x \leq 0$ for the negation. (In this example which of the following is true: P or $\neg P$?) △

We can extend these methods to negate statements with multiple quantifiers.

Example 5.8. Let x, y, z be variables with domains S, T, and U, respectively, and let $P(x, y, z)$ be an open sentence. We will negate the statement

$$\forall x \in S, \forall y \in T, \exists z \in U,\ P(x, y, z).$$

In order to do this, we will use parentheses to make the order of the quantifiers clearer. Hence, the statement that we wish to negate is

$$\forall x \in S,\ (\forall y \in T,\ (\exists z \in U,\ P(x, y, z))).$$

Proceeding one level at a time, we see that

$$\begin{aligned} & \neg(\forall x \in S, \forall y \in T, \exists z \in U,\ P(x, y, z)) \\ \equiv\ & \exists x \in S,\ \neg(\forall y \in T, \exists z \in U,\ P(x, y, z)) \\ \equiv\ & \exists x \in S, \exists y \in T,\ \neg(\exists z \in U,\ P(x, y, z)) \\ \equiv\ & \exists x \in S, \exists y \in T, \forall z \in U,\ \neg P(x, y, z). \end{aligned}$$

Notice that negating this quantified statement was as simple as swapping all existential and universal quantifiers, and negating the open sentence at the end. △

Warning 5.9. One mistake that many students make is getting carried away with negating quantifiers. In particular, when negating the statement

$$\forall x \in S,\ P(x)$$

they might write, incorrectly,

$$\exists x \notin S,\ \neg P(x).$$

Notice that this sentence may be meaningless, because $P(x)$ may not have a truth value (or even make sense) when $x \notin S$. The thing to remember is to swap the quantifier **without** changing the domain of the variable.

Example 5.10. The problem mentioned in the previous warning most commonly happens when we use notation other than $x \in S$ in our quantifiers. For instance, if x is a real variable, and we wish to negate

$$\forall x > 0,\ P(x),$$

the correct negation is

$$\exists x > 0,\ \neg P(x).$$

Since "$x > 0$" is shorthand for $x \in \mathbb{R}_{>0}$, it would be incorrect to change it to $x \leq 0$. △

Example 5.11. Suppose that we wish to negate the statement

$$P:\ \forall x \in \mathbb{R}, \exists y \in \mathbb{R},\ x > y.$$

(Remember that this is the true statement from Example 5.1.) Then the negation is

$$\neg P:\ \exists x \in \mathbb{R}, \forall y \in \mathbb{R},\ x \leq y.$$

This new statement is false. △

5.C Greatest and least elements

We now use quantifiers to define some common terminology concerning sets of real numbers.

Definition 5.12. Let $S \subseteq \mathbb{R}$, and let $x \in \mathbb{R}$.

(1) We say that x is the *greatest element* of S if $x \in S$ and for each $y \in S$ we have $x \geq y$.
(2) We say that x is the *least element* of S if $x \in S$ and for each $y \in S$ we have $x \leq y$.

In symbols, x is the greatest element of S if

$$(x \in S) \wedge (\forall y \in S,\ x \geq y),$$

and x is the least element of S if

$$(x \in S) \wedge (\forall y \in S,\ x \leq y).$$

To say that a set S has a greatest element would be to say

$$\exists x \in \mathbb{R},\ (x \in S \wedge (\forall y \in S,\ x \geq y)).$$

However, this can be written in a shorter way as

$$\exists x \in S, \forall y \in S,\ x \geq y.$$

We summarize this information in the definition below.

Definition 5.13. Let $S \subseteq \mathbb{R}$.

(1) We say that S *has a greatest element* if there is some $x \in S$ that is the greatest element of S.
(2) We say that S *has a least element* if there is some $x \in S$ that is the least element of S.

Example 5.14. Let $S = \mathbb{N}$ be the set of natural numbers. Then $S \subseteq \mathbb{R}$. We note that S has no greatest element. To see this, we may negate the definition of "S has a greatest element," to obtain

$$\forall x \in S, \exists y \in S,\ x < y.$$

This statement is true since, for any $x \in S$, there is some $y \in S$ (namely $y = x + 1$) such that $x < y$. Since the negation of "S has a greatest element" is true, we see that "S has a greatest element" is false.

On the other hand, for $S = \mathbb{N}$, the statement "S has a least element" is true; the least element of the natural numbers is 1. △

Definition 5.15. Let $S \subseteq \mathbb{R}$.

(1) An *upper bound* for S is some $x \in \mathbb{R}$ such that for all $y \in S$, $y \leq x$.
(2) A *lower bound* for S is some $x \in \mathbb{R}$ such that for all $y \in S$, $y \geq x$.
(3) We say that S *has an upper bound* if there exists some $x \in \mathbb{R}$ such that for all $y \in S$, $y \leq x$.
(4) We say that S *has a lower bound* if there exists some $x \in \mathbb{R}$ such that for all $y \in S$, $y \geq x$.

Example 5.16. Let $S = (0, 1] = \{x \in \mathbb{R} : 0 < x \leq 1\}$. Then S has a lower bound; if we take $x = -1$, we see that every element $y \in S$ satisfies $y \geq x$. Note that we could have taken $x = 0$ or $x = -3$ as well. We remark that although S has a lower bound, it has no least element. △

We will investigate the ideas of greatest and least elements and upper and lower bounds more in the exercises.

5.D Chart of negation rules

The following chart summarizes how to simplify negated sentences.

$$\begin{aligned}
\neg(\neg P) &\longrightarrow P \\
\neg(P \vee Q) &\longrightarrow \neg P \wedge \neg Q \\
\neg(P \wedge Q) &\longrightarrow \neg P \vee \neg Q \\
\neg(P \Rightarrow Q) &\longrightarrow P \wedge \neg Q \\
\neg(\forall x \in S,\ P(x)) &\longrightarrow \exists x \in S,\ \neg P(x) \\
\neg(\exists x \in S,\ P(x)) &\longrightarrow \forall x \in S,\ \neg P(x)
\end{aligned}$$

5.E Exercises

Exercise 5.1. Write the negation of the following statements and open sentences. In each case, the domain of each variable x, ε, and δ is the set of real numbers. (Write any quantifiers and logical connectives using English.)

(a) $x > 2$ and $x < 3$.
(b) If $x > 3$, then $x > 2$.
(c) If $x > 3$ and $x \neq 4$, then $x^2 \neq 16$.
(d) If $3 < x < 4$, then $9 < x^2 < 16$.
(e) If $x = 2$ or $x = 3$, then $x^2 - 5x + 6 = 0$.
(f) For all $x \in \mathbb{R}$, it happens that $x^2 + 2x > 0$.
(g) There exists an $x \in \mathbb{R}$ such that $x^2 + 2x > 0$.
(h) For each $x \in \mathbb{R}$, there exists $y \in \mathbb{R}$ such that $y > x^2$.
(i) There exists an $x \in \mathbb{R}$, such that for all $y \in \mathbb{R}$, it holds that $y > x^2$.
(j) For each $\varepsilon > 0$, there exists some $\delta > 0$ such that for each $x \in \mathbb{R}$, if $0 < |x-2| < \delta$, then $|x^2 - 4| < \varepsilon$.

Exercise 5.2. For each pair of statements, decide if they have the same truth value.

(a) $\forall x \in \mathbb{R}, \exists y \in \mathbb{R},\ x + y = 0$ and $\exists y \in \mathbb{R}, \forall x \in \mathbb{R},\ x + y = 0$.
(b) $\forall x \in \mathbb{R}, \exists y \in \mathbb{R},\ xy = 0$ and $\exists y \in \mathbb{R}, \forall x \in \mathbb{R},\ xy = 0$.
(c) $\forall x \in \mathbb{R}, \exists y \in \mathbb{R},\ xy \neq 0$ and $\exists y \in \mathbb{R}, \forall x \in \mathbb{R},\ xy \neq 0$.
(d) $\forall x \in \mathbb{R}, \exists y \in \mathbb{R},\ y + x^2 > 0$ and $\exists y \in \mathbb{R}, \forall x \in \mathbb{R},\ y + x^2 > 0$.

Exercise 5.3. Do the following:

(a) Give an example of a set of real numbers that has an upper bound, but does not have a greatest element.
(b) Determine whether or not there can be a set that has a greatest element, but does not have an upper bound. Explain your answer.

Exercise 5.4. Let $S = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\right\} = \left\{\frac{1}{n} : n \in \mathbb{N}\right\}$.

(a) Does S have an upper bound? If so, give an upper bound.
(b) Does S have a greatest element? If so, what is it?
(c) Does S have a lower bound? If so, give a lower bound.
(d) Does S have a least element? If so, what is it?

Exercise 5.5. Let $S = (0, 1)$ be the open interval of real numbers between 0 and 1.

(a) Does S have an upper bound? If so, give an upper bound.
(b) Does S have a greatest element? If so, what is it?
(c) Does S have a lower bound? If so, give a lower bound.
(d) Does S have a least element? If so, what is it?

Exercise 5.6. Let S be a set of real numbers, and let $x \in \mathbb{R}$.

(a) Write (in symbolic logic) the negation of the statement "x is the greatest element of S."
(b) Write (in symbolic logic) the negation of the statement "x is an upper bound for S."
(c) Write (in symbolic logic) the negation of the statement "S has an upper bound."

(Hint: The answer in part (c) should not use the symbol "x", since x was already given a meaning earlier in the problem. Instead, use a different letter, like z, in place of x.)

Chapter III

Basic Proof Techniques

Proving the obvious has never been easy. Marty Rubin

The last two chapters were an introduction to the *language* of mathematics. Knowing the definitions and concepts of set theory and logic allow us to communicate thoughts more clearly and succinctly. In this chapter we will put our new knowledge to use in *proving* that statements are true.

A good proof is like a good painting. It opens the viewer's mind to deeper insights, connections, and beauties. The purpose of a proof is not only to convince the reader that something is true, but to do so in a way that aids in their understanding of *why* it is true.

6 Direct proofs

6.A Terminology

All mathematical arguments need a foundation on which to stand; we need truths which are taken to be self-evident. These truths are called *axioms*. Some axioms which are commonly used by mathematicians are the following:

- The empty set exists.
- For each real number, there is an integer greater than it.
- Any two lines either intersect in a single point or are parallel.

In this book we will not worry too deeply about which axioms we will use, trusting that readers will learn by example what sorts of statements they may use freely.

Once the language of mathematics is in place we also have *definitions*. We gave many examples in the previous two chapters of concepts we have defined, such as unions, intersections, logical connectives, and so forth. In this section, the following two definitions will be very important.

Definition 6.1. Let $n \in \mathbb{Z}$. We say n is an *even integer* if $n = 2k$ for some $k \in \mathbb{Z}$. The set of even integers is written

$$\textsc{Even} = 2\mathbb{Z} = \{\ldots, -4, -2, 0, 2, 4, 6, \ldots\}.$$

Definition 6.2. Let $n \in \mathbb{Z}$. We say that n is an *odd integer* if $n = 2k + 1$ for some $k \in \mathbb{Z}$. The set of odd integers is written

$$\textsc{Odd} = 2\mathbb{Z} + 1 = \{\ldots, -3, -1, 1, 3, 5, \ldots\}.$$

Example 6.3. The integer 3 is odd, since $3 = 2 \cdot 1 + 1$. The integer 16 is even, since $16 = 2 \cdot 8$. △

Warning 6.4. If n is an even integer, this does not mean that $n = 2k$ for all $k \in \mathbb{Z}$. It is impossible for a single number to equal *all* of the even integers at once.

The following three facts probably seem self-evident:

- If $x \in \mathbb{Z}$ is not even, then it is odd.
- If $x \in \mathbb{Z}$ is not odd, then it is even.
- There is no number which is both even and odd.

You may freely use these facts, treating them as axioms. (We will prove in a later chapter that they follow from a more basic axiom.)

Finally, there are special names given to statements we want to prove.

- A *theorem* is an important statement we prove. In this book we will only name results "Theorems" when they are extremely important.
- A *proposition* is a basic statement we prove.
- A *lemma* is a (usually minor) result that we prove in order to use it later to prove a bigger result. We use lemmas as a way of breaking up large proofs into smaller, more understandable pieces.
- A *corollary* is a (usually minor) result that follows easily from another result we have proved.

In the rest of Section 6 we will focus on techniques for proving statements of the form:

$$\forall x \in S,\ P(x) \Rightarrow Q(x).$$

6.B Trivial proofs

There are two situations where an implication is true for silly reasons. The following definition gives the first situation.

Definition 6.5. Let S be a set. We say that the statement

$$\forall x \in S,\ P(x) \Rightarrow Q(x)$$

is *trivially true* when the simpler statement $\forall x \in S,\ Q(x)$ is true. In other words, an implication is trivially true when the conclusion of the implication is always true.

Here are some easy examples, followed by proofs.

Proposition 6.6. *Let $x \in \mathbb{R}$. If $x^2 < 73$, then $0 < 1$.*

Proof. The conclusion $0 < 1$ is always true. So the implication is trivially true. □

Proposition 6.7. *Let $a \in \mathbb{Z}$. If a is odd, then $2a$ is even.*

Proof. The conclusion is always true; the number $2a$ is even since it is 2 times an integer. Hence the implication is trivially true. □

Notice that in both cases the premise of the implication was irrelevant. In the first proposition it didn't matter whether or not $x^2 < 73$, since $0 < 1$ is always true. Similarly, in the second proposition it didn't matter whether or not a is odd, because $2a$ is always even.

Also notice the little square box at the end of the proof. This tells the reader that you have finished the proof.

Warning 6.8. The word "trivial" should not be used in a proof to mean "this step is easy, so I will skip it."

Sometimes "trivial" proofs are not easy and take some work to prove.

Proposition 6.9. *Let $x \in \mathbb{R}$. If $x < 5$, then $x^2 - 2x \geq -1$.*

Proof. The inequality $x^2 - 2x \geq -1$ is equivalent to $x^2 - 2x + 1 \geq 0$. This is the same as $(x-1)^2 \geq 0$, which is always true. Thus, the implication is trivially true. □

Advice 6.10. It is always a good idea to understand the premise and conclusion of an implication separately before trying to prove the implication.

6.C Vacuous proofs

There is a second situation where implications are true for silly reasons. This happens when the premise is always false.

Definition 6.11. Let S be a set. We say that the statement

$$\forall x \in S,\ P(x) \Rightarrow Q(x)$$

is *vacuously true* when the simpler statement $\forall x \in S,\ \neg P(x)$ is true. In other words, an implication is vacuously true when the premise of the implication is never true.

We use the word "vacuous" when the premise of an implication is false everywhere in the domain because we think of the implication as an empty promise in that case. In other words, the statement is true because it isn't asserting anything!

The following are some examples of vacuously true statements and their proofs.

Proposition 6.12. *Let $x \in \mathbb{Z}$. If $3 < x$ and $x < 2$, then $x^2 + 4 = 7$.*

Proof. The premise is impossible, so the implication is vacuously true. □

Proposition 6.13. *Let $x \in \mathbb{Z}$. If $-x^2 > 2$, then $x = 5$.*

Proof. The premise is $-x^2 > 2$, which is equivalent to $0 > x^2 + 2$. The right side is the sum of a square and a positive number; such a sum can never be negative. Thus, the implication is vacuously true. □

Advice 6.14. To remember the difference between trivial and vacuous proofs, memorize the following phrase:
Trivial, the Q is true.
Vacuous, the premise is bogus.

Sometimes it is not obvious whether a statement is vacuously true or trivially true. The following is a list of some examples.

Example 6.15. Consider the following statements, and determine if they are trivially true, vacuously true, or neither trivial nor vacuous.
(1) Given $a \in \mathbb{R}$, if a^2 is a negative real number, then $a = 5$.
(2) Given $a \in \mathbb{C}$, if a^2 is a negative real number, then $a \in \mathbb{C} - \mathbb{R}$.
(3) Let $x \in \mathbb{N}$. If $x \in \emptyset$, then $x > 0$.
(4) Let $a \in \mathbb{N}$. Either a is even or a is odd.

The answers are as follows:
(1) The statement is vacuously true, since the premise is never true. It is not trivial, because the conclusion can be false for some $a \in \mathbb{R}$ (such as $a = 6$).
(2) This implication is not vacuous, since the premise is true for some $a \in \mathbb{C}$ (such as $a = i = \sqrt{-1}$). It is also not trivial since the conclusion is false for some $a \in \mathbb{C}$ (such as $a = 0$).
(3) This statement is both vacuous and trivial! The premise is never true, and the conclusion is always true (for $x \in \mathbb{N}$).
(4) This is not an implication, so it is not trivial or vacuous. △

6.D Outline of a direct proof

Now that we have dealt with the two *oddities* that can arise, we are ready to introduce the most important technique for proving implications.

Given a statement $\forall x \in S,\ P(x) \Rightarrow Q(x)$, we can prove it *directly* by assuming the premise $P(x)$ holds and then, using that information, we show that $Q(x)$ must also hold true. We begin with an example of such a proof.

Proposition 6.16. *For each $x \in \mathbb{Z}$, if x is even, then $5x + 3$ is odd.*

Proof. Let $x \in \mathbb{Z}$ be arbitrary. We will work directly. Assume x is even. This means $x = 2k$ for some $k \in \mathbb{Z}$. Thus

$$5x + 3 = 5(2k) + 3 = 10k + 3 = 2(5k + 1) + 1.$$

Since $5k + 1 \in \mathbb{Z}$, this means that $5x + 3$ is odd. □

In the first sentence, we deal with the $\forall x \in \mathbb{Z}$, by telling the reader we are letting x be an arbitrary integer. The next sentence tells the reader what type of proof technique we will use. In this case it is a direct proof. (We will have more options

available over the next few sections.) After telling the reader we are working directly, we *must* assume the premise of the implication. So the third sentence does exactly that; we assume x is even. We then do some work, and we finish by showing that the conclusion holds.

The basic outline of a direct proof is as follows.

Result to be proved. *Given $x \in S$, if $P(x)$ is true, then $Q(x)$ is true.*

Proof outline. Let $x \in S$.
We work directly.
Assume $P(x)$.
Do some work (to be filled in).
Thus, $Q(x)$ holds. □

Advice 6.17. When first writing proofs it can be useful to leave some space in the middle and write the last sentence on the bottom. That way, you can see where you need to end up.

Here are some more examples of direct proofs.

Proposition 6.18. *If n is an odd integer, then $n^2 - 2n + 3$ is even.*

Proof. Let $n \in \mathbb{Z}$. We work directly. Assume n is odd. Hence $n = 2k + 1$ for some $k \in \mathbb{Z}$. We then find

$$\begin{aligned} n^2 - 2n + 3 &= (2k+1)^2 - 2(2k+1) + 3 \\ &= 4k^2 + 4k + 1 - 4k - 2 + 3 = 4k^2 + 2 = 2(2k^2 + 1). \end{aligned}$$

Since $2k^2 + 1$ is an integer, $n^2 - 2n + 3$ is even. □

Proposition 6.19. *Let $n \in \mathbb{Z}$. If $3n$ is even, then $n + 7$ is odd.*

Proof. Let $n \in \mathbb{Z}$. We work directly. Assume $3n$ is even. Hence $3n = 2k$ for some $k \in \mathbb{Z}$. We find

$$\begin{aligned} n + 7 &= n + (2n - 2n) + 7 = 3n - 2n + 7 \\ &= 2k - 2n + 7 = 2(k - n + 3) + 1. \end{aligned}$$

Since $k - n + 3 \in \mathbb{Z}$, we have $n + 7$ is odd. □

In this proof, you might be tempted, after writing $3n = 2k$, to solve for n and get $n = 2k/3$. However, at that point you are no longer working with integers; instead, you are working with rational numbers. Try to fill in the rest of the proof and look for where you get an error. (Note: We will see in the next section an easier way to prove this proposition, which doesn't involve the trick of adding $0 = 2n - 2n$.)

We end this section with one final example.

Proposition 6.20. *If n is an odd integer, then $4n^2 - 1$ is odd.*

Proof. Let $n \in \mathbb{Z}$. We work directly. Assume n is odd. Thus, $n = 2k + 1$ for some $k \in \mathbb{Z}$. We find

$$\begin{aligned} 4n^2 - 1 &= 4(2k+1)^2 - 1 = 4(4k^2 + 4k + 1) - 1 \\ &= 16k^2 + 16k + 3 = 2(8k^2 + 8k + 1) + 1. \end{aligned}$$

Since $8k^2 + 8k + 1 \in \mathbb{Z}$, we have $4n^2 - 1$ is odd. □

Did you notice that there is another way to prove this proposition? (Hint: Is it trivial/vacuous?)

6.E Exercises

Exercise 6.1. Let $x \in \mathbb{R}$. Prove that if $x \neq 3$, then $x^2 - 2x + 3 \neq 0$. (Would this result be true if we took $x \in \mathbb{C}$?)

Exercise 6.2. Let $n \in \mathbb{N}$. Prove that if $2 < n < 3$, then $7n + 3$ is odd.

Exercise 6.3. Prove that if x is an odd integer, then x^2 is odd.

Exercise 6.4. Prove that if x is an even integer, then $7x - 5$ is odd.

Exercise 6.5. Let $a, b, c \in \mathbb{Z}$. Prove that if a and c are odd, then $ab + bc$ is even.

Exercise 6.6. Let $n \in \mathbb{Z}$. Prove that if $|n| < 1$, then $3n - 2$ is an even integer.

Exercise 6.7. Prove that every odd integer is a difference of two squares of integers. (Hint: Try small cases; write 1, 3, 5, and 7 as differences of squares. It might help to rephrase this statement as an implication, with a premise and conclusion.)

7 Contrapositive proof

While the technique of direct proof is a powerful tool, this section will introduce another method which is very similar in spirit. This new method is called "contrapositive proof." You may have experienced it in your own life. For instance, consider the following story:

Alice is at work on Friday and tells her coworker Bob: "If it rains on Monday, then I'm not coming in to work." Bob has an important deadline on Monday, and so works through the entire weekend. On Monday Bob sees Alice come into the office. He concludes it must not be raining.

Bob's logic is sound and will be explained (and exploited) in this section.

7.A What is the contrapositive?

Let P and Q be statements. We have been working with the implication

$$R: \ P \Rightarrow Q.$$

We say that the new sentence

$$\neg Q \Rightarrow \neg P$$

is the *contrapositive* of R. A truth table shows the amazing fact that the implication R has exactly the same truth table as the contrapositive! In other words:

Theorem 7.1. *Let P and Q be statements. Then*

$$P \Rightarrow Q \ \equiv \ \neg Q \Rightarrow \neg P.$$

Example 7.2. We will find the contrapositive of the statement: "If it rains on Monday, then I'm not coming in to work on Monday." The contrapositive (after an easy application of double negation) is exactly: "If I come in to work on Monday, then it is not raining on Monday." In the story at the beginning of this section, Bob used the contrapositive of Alice's sentence to conclude it was not raining. △

We can also take the contrapositive of an implication between two open sentences, as in the following example.

Example 7.3. The contrapositive of

$$\text{If } 3x - 7 \text{ is even, then } x \text{ is odd.}$$

is the new sentence

$$\text{If } x \text{ is even, then } 3x - 7 \text{ is odd.}$$

Which one looks easier to prove when universally quantified? △

To prove an implication $P \Rightarrow Q$, we can instead change to the contrapositive sentence $\neg Q \Rightarrow \neg P$ and work directly with this new sentence. We usually work this way when the direct proof (of the original sentence) is more difficult. The proof of the following proposition shows how this is to be done: we *assume* the negation of the conclusion and then *show* the negation of the premise.

Proposition 7.4. *Given $x \in \mathbb{Z}$, if $3x - 7$ is even, then x is odd.*

Proof. Let $x \in \mathbb{Z}$. We work contrapositively. Assume x is even. Thus $x = 2k$ for some $k \in \mathbb{Z}$. We find

$$3x - 7 = 3(2k) - 7 = 6k - 7 = 2(3k - 4) + 1$$

is odd. □

It is possible to prove this proposition directly; however, it will be more difficult. When we assume $3x - 7$ is even, we get $3x - 7 = 2k$. It is very tempting to solve for x by bringing 7 to the other side and dividing by 3. But if we do that, we get $x = (2k + 7)/3$ which is written as a rational fraction, so it is difficult to tell whether it is even or odd (or an integer!). However, there is a clever way to work around this fact:

Alternative proof. Let $x \in \mathbb{Z}$. We work directly. Assume $3x-7$ is even. So $3x-7 = 2k$ for some $k \in \mathbb{Z}$. Then

$$x = (3x - 7) - 2x + 7 = 2k - 2x + 7 = 2(k - x + 3) + 1$$

is odd. □

The moral of this story is:

Advice 7.5. To decide whether to do a direct proof or a contrapositive proof, choose the method which gives you the "best" information to start with.

For the next proposition, try to decide for yourself whether you should work directly or contrapositively.

Proposition 7.6. *Let $x \in \mathbb{Z}$. If $x^2 - 6x + 7$ is odd, then x is even.*

Do you want to work directly? If you do, you will assume $x^2 - 6x + 7$ is odd, and try to show that x is even.

Or do you want to work contrapositively? If you do, you will assume x is odd, and try to show that $x^2 - 6x + 7$ is even.

Proof. Let $x \in \mathbb{Z}$. We work contrapositively. Assume x is odd. So $x = 2k + 1$ for some $k \in \mathbb{Z}$. We find

$$\begin{aligned} x^2 - 6x + 7 &= (2k+1)^2 - 6(2k+1) + 7 = 4k^2 + 4k + 1 - 12k - 6 + 7 \\ &= 4k^2 - 8k + 2 = 2(2k^2 - 4k + 1) \end{aligned}$$

is even. □

For many people, to prove the previous proposition it is easier to work contrapositively than to work directly.

7.B Division

We have done many proofs using even and odd integers. We will now introduce a new definition which we can use to prove more statements.

Example 7.7. (1) Does 3 divide 6? Yes, dividing 6 by 3 yields the integer 2. So $6 = 3 \cdot 2$.

(2) Does 5 divide 9? No, dividing 9 by 5 yields a remainder of 4. △

Generalizing these examples we have the following:

Definition 7.8. Let $a, b \in \mathbb{Z}$. We say that a *divides* b if

$$\exists c \in \mathbb{Z},\ b = ac.$$

In other words, a is a *divisor* of b. Alternatively, b is a *multiple* of a.

We write $a \mid b$ to mean "a divides b."

Warning 7.9. The symbols $a \mid b$ represent a sentence that means a divides *into* b. It means that the fraction $\frac{b}{a}$ is an integer. It does **not** mean $a \div b$ (which is just the number $\frac{a}{b}$, not a sentence).

Example 7.10. (1) Is it true that $7 \mid 21$? Yes, $21 = 7 \cdot 3$.

(2) For any integer n, we have $1 \mid n$, since $n = 1 \cdot n$.

(3) What integers does 0 divide? If $0 \mid n$, then this means $n = 0c$ for some $c \in \mathbb{Z}$. Thus, $n = 0$. On the other hand, $0 \mid 0$ is true since $0 = 0 \cdot 1$.

(4) Does $8 \mid 4$? No, so we write $8 \nmid 4$. △

To prove a statement involving the condition $a \mid b$ we use its definition to turn it into the new sentence "$b = ac$ for some $c \in \mathbb{Z}$." (Notice that the term b which was on the right-hand side of the vertical bar is now alone on the left-hand side of the equality.) This is similar to what we did with even/odd proofs; when x is even, we rewrite this using the definition to say that $x = 2k$ for some $k \in \mathbb{Z}$.

Proposition 7.11. *Let $a, b \in \mathbb{Z}$. If $a \mid b$, then $a \mid 2b$.*

Proof. Let $a, b \in \mathbb{Z}$. We work directly. Assume $a \mid b$. Thus $b = ac$ for some $c \in \mathbb{Z}$.
Now, $2b = 2(ac) = a(2c)$. Since $2c \in \mathbb{Z}$, we have $a \mid 2b$. □

Question: In the sentence starting "Now, $2b = \ldots$," how did we know to look at $2b$?
Answer: We were working directly, so we knew we needed to show $a \mid 2b$. From the definition, we knew we needed to look at $2b$ and "pull out" $a \in \mathbb{Z}$ as a factor, with the other factor also an integer.

Can you figure out what is wrong with the following **FALSE** proposition and proof?

Proposition 7.12 (False!)**.** *Let $a, b \in \mathbb{Z}$. If $a \mid 2b$, then $a \mid b$.*

Proof. Let $a, b \in \mathbb{Z}$. We work directly. Assume $a \mid 2b$. Thus $2b = ac$ for some $c \in \mathbb{Z}$.
Now, $b = ac/2 = a(c/2)$. Since $c/2 \in \mathbb{Z}$, we have $a \mid b$. □

When proving statements about divisors, we usually do **NOT** actually divide!
Here are some more examples of proofs using divisors.

Proposition 7.13. *Let $a, b, c \in \mathbb{Z}$. If $a \mid b$ and $b \mid c$, then $a \mid c$.*

Proof. Let $a, b, c \in \mathbb{Z}$. We work directly. Assume $a \mid b$ and $b \mid c$. Thus $b = ax$ and $c = by$ for some $x, y \in \mathbb{Z}$.
Now, $c = by = (ax)y = a(xy)$. Since $xy \in \mathbb{Z}$, we have $a \mid c$. □

Question: Why did we use x and y instead of c?
Partial Answer: There are two reasons! First, because c was a symbol already in use. Can you figure out the reason why we needed *different* letters x and y?

Proposition 7.14. *Given $x \in \mathbb{Z}$, if $2 \nmid x$ then x is odd.*

Proof. Let $x \in \mathbb{Z}$. We work contrapositively. Assume x is even. Thus $x = 2k$ for some $k \in \mathbb{Z}$. Hence $2 \mid x$. □

Theorem 7.15. *Let $a, b, c, x, y \in \mathbb{Z}$. If $a \mid b$ and $a \mid c$, then $a \mid (bx + cy)$.*

Proof. Let $a, b, c, x, y \in \mathbb{Z}$. We work directly. Assume $a \mid b$ and $a \mid c$. Thus $b = au$ and $c = av$ for some $u, v \in \mathbb{Z}$.
Now,

$$bx + cy = (au)x + (av)y = a(ux + vy).$$

Since $ux + vy \in \mathbb{Z}$, we have $a \mid (bx + cy)$. □

7.C More terminology for implications

Suppose we are given an implication

$$R(x): \ P(x) \Rightarrow Q(x).$$

There are other interesting open sentences related to R. Consider the following:

- The *converse* of $R(x)$ is: $Q(x) \Rightarrow P(x)$.
- The *inverse* of $R(x)$ is: $\neg P(x) \Rightarrow \neg Q(x)$.
- The *contrapositive* of $R(x)$ is: $\neg Q(x) \Rightarrow \neg P(x)$.
- The *negation* of $R(x)$ is: $\neg R(x) \equiv \neg(P(x) \Rightarrow Q(x)) \equiv P(x) \wedge \neg Q(x)$.

The contrapositive has the same truth table as $R(x)$ (treating $P(x)$ and $Q(x)$ as components of a compound sentence). However, the converse has a different truth table. (Try it!) The negation has another, third, truth table. (Try it!) So the converse, the contrapositive, and the negation are (in general) all very different sentences.

What about the inverse? It does have the same truth table as one of the other three sentences! But which one? (Try it!)

7.D Biconditional

Over the past few sections, we have been focusing on proving (universally quantified) implications $P(x) \Rightarrow Q(x)$. Another common sentence to prove is the biconditional $P(x) \Leftrightarrow Q(x)$. To prove it, we show both $P(x) \Rightarrow Q(x)$ and the converse $Q(x) \Rightarrow P(x)$. Each direction can be proved directly or contrapositively! Consider the following example.

Proposition 7.16. *Let $n \in \mathbb{Z}$. The number n^2 is odd if and only if n is odd.*

Proof. We first prove that if n^2 is odd, then n is odd. We work contrapositively, so assume n is even. Then $n = 2k$ for some $k \in \mathbb{Z}$. Hence

$$n^2 = (2k)^2 = 4k^2 = 2(2k^2)$$

is an even integer.

We now prove conversely that if n is odd, then n^2 is odd. We work directly. Assume n is odd. Then $n = 2k + 1$ for some $k \in \mathbb{Z}$. Hence

$$n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$$

is an odd integer. □

To help the reader, sometimes a proof writer will use arrows to tell the reader which of the two directions is being proved. For instance, the previous proof could be rewritten as follows.

Proof. ($\Rightarrow$): We work contrapositively. Assume n is even, so $n = 2k$ for some $k \in \mathbb{Z}$. Then

$$n^2 = (2k)^2 = 4k^2 = 2(2k^2)$$

is an even integer.

($\Leftarrow$): We work directly. Assume n is odd, so $n = 2k + 1$ for some $k \in \mathbb{Z}$. Hence

$$n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$$

is an odd integer. □

We end this subsection with a proof of some statement that we haven't written down. Try to figure out **what** is being proved, from the proof itself.

Proof. Assume x is even and y is odd. Therefore $x = 2k$ and $y = 2\ell + 1$ for some $k, \ell \in \mathbb{Z}$. We find

$$(x + 1)y = (2k + 1)(2\ell + 1) = 4k\ell + 2k + 2\ell + 1 = 2(2k\ell + k + \ell) + 1$$

is an odd integer. □

Which of the following is being proved?

(1) If x is even and y is odd, then $(x + 1)y$ is odd.
(2) If $(x + 1)y$ is odd, then x is even and y is odd.
(3) If x is odd or y is even, then $(x + 1)y$ is even.
(4) If $(x + 1)y$ is even, then x is odd or y is even.

Answer: Did you see that **two** of the choices are actually correct? We could be working directly to prove (1), or we could be working contrapositively to prove (4).

7.E Exercises

Exercise 7.1. Let $a \in \mathbb{Z}$. Prove that if $a^2 + 3$ is odd, then a is even.

Exercise 7.2. Prove the following: Let $x, y \in \mathbb{Z}$. If $xy + y^2$ is even, then x is odd or y is even.

Exercise 7.3. Let $s \in \mathbb{Z}$. Prove that s is odd if and only if s^3 is odd.

Exercise 7.4. Consider the following situation. A student is asked to prove the statement: "Given $x \in \mathbb{Z}$, if $2 \mid x$, then x is even." The student writes: "Assume, contrapositively, that x is even. Then $x = 2k$ for some $k \in \mathbb{Z}$. Hence $2 \mid x$."

Identify what is wrong with this student's proof and write a correct proof.

Exercise 7.5. Let $a, b, c, d \in \mathbb{Z}$. Prove that if $a \mid c$ and $b \mid d$, then $ab \mid cd$.

Exercise 7.6. State the contrapositive of the implication in the previous exercise.

Exercise 7.7. Let $a \in \mathbb{Z}$. Prove that if $4 \nmid a^2$, then a is odd.

Exercise 7.8. Prove the following implication two ways (directly and contrapositively): Given $x \in \mathbb{Z}$, then $5x - 1$ is even only if x is odd. (Be careful to prove the correct implication. See Subsection 4.D for the meaning of "only if".)

8 Proof by cases

8.A Introductory examples

Some problems break down into natural cases. Here are some common examples:

- Integers are either even or odd.
- Real numbers (or integers) can be positive, negative, or zero.
- Numbers can be zero or nonzero.
- Real numbers can be rational or irrational.
- Sets can be empty or nonempty.
- Sets can be finite or infinite.

Some problems can be handled by considering all possible cases separately. The proof of the following proposition shows how this is to be done.

Proposition 8.1. *If $x \in \mathbb{Z}$, then $x^2 + x$ is even.*

Proof. We work directly. Assume $x \in \mathbb{Z}$. There are two natural cases.

Case 1: Suppose x is even. Then $x = 2k$ for some $k \in \mathbb{Z}$. Then

$$x^2 + x = (2k)^2 + 2k = 4k^2 + 2k = 2(2k^2 + k)$$

is even.

Case 2: Suppose x is odd. Then $x = 2k + 1$ for some $k \in \mathbb{Z}$. Then

$$x^2 + x = (2k + 1)^2 + (2k + 1) = 4k^2 + 6k + 2 = 2(2k^2 + 3k + 1)$$

is even.

In every case $x^2 + x$ is even. □

The key to working by cases is that we can break a problem into smaller "subproblems" that each can be handled separately. Sometimes it takes practice to recognize how to break a problem into smaller cases. On the other hand, sometimes a problem shouts "Do me by cases!" To demonstrate, let's introduce another definition.

Definition 8.2. Let $x, y \in \mathbb{Z}$. We say that x and y have the *same parity* if either they are both even or they are both odd. If this doesn't happen, we say that x and y have the *opposite parity*.

Proposition 8.3. *Let $x, y \in \mathbb{Z}$. If $x + y$ is even, then x and y have the same parity.*

Proof. Let $x, y \in \mathbb{Z}$. We work contrapositively. Thus, we assume x and y have opposite parity. There are two natural cases to consider.

Case 1: Suppose x is even and y is odd. Thus $x = 2k$ and $y = 2\ell + 1$ for some $k, \ell \in \mathbb{Z}$. Then

$$x + y = 2k + 2\ell + 1 = 2(k + \ell) + 1$$

is odd.

Case 2: Suppose x is odd and y is even. Thus $x = 2k + 1$ and $y = 2\ell$ for some $k, \ell \in \mathbb{Z}$. Then

$$x + y = 2k + 1 + 2\ell = 2(k + \ell) + 1$$

is odd.

In every case $x + y$ is odd. □

Notice that in this proof the two cases look almost exactly the same. When this happens sometimes mathematicians save time by writing "The other case is similar." or "There are two cases, but without loss of generality we may assume x is even and y is odd." Feel free to write this, but only if the two cases really are no different. Consider the following proposition, which is the inverse of the previous proposition.

Proposition 8.4. *Let $x, y \in \mathbb{Z}$. If $x + y$ is odd, then x and y have opposite parity.*

Proof. Let $x, y \in \mathbb{Z}$. We work contrapositively. Assume x and y have the same parity. There are two cases.

Case 1: Suppose x, y are both even. Then $x = 2k$ and $y = 2\ell$ for some $k, \ell \in \mathbb{Z}$. Then

$$x + y = 2k + 2\ell = 2(k + \ell)$$

is even.

Case 2: Suppose x, y are both odd. Then $x = 2k + 1$ and $y = 2\ell + 1$ for some $k, \ell \in \mathbb{Z}$. Then

$$x + y = 2k + 1 + 2\ell + 1 = 2(k + \ell + 1)$$

is even.

In every case $x + y$ is even. □

Notice that we didn't say that Case 2 is similar to Case 1. That's because they really are different!

In some situations, the cases we should consider come from one of our assumptions.

Proposition 8.5. *For any $x, y \in \mathbb{Z}$, if x is even or y is even, then xy is even.*

Proof. Let $x, y \in \mathbb{Z}$. We work directly. Assume that x is even or y is even.

Case 1: Suppose x is even. Hence $x = 2k$ for some $k \in \mathbb{Z}$. Hence $xy = 2ky$ is even.

Case 2: Suppose y is even. This is similar to Case 1.

Thus, xy is even in all cases. □

Why were those two cases enough to cover all possibilities? Aren't we missing the case where x is odd and y is odd? We *are* ignoring the case where x is odd and y is odd! But we can ignore that case because we *assumed* the fact that x is even or y is even (when working directly). Our assumption limited the number of cases we needed to consider. If we hadn't made that assumption, we would need to consider that last possibility.

Advice 8.6. If you *assume* $P \vee Q$, then you have two cases: case 1 is when P holds, and case 2 is when Q holds. You could also break this up into three separate cases: case 1 is when P and Q both hold, case 2 is when P holds but Q fails, and case 3 is when P fails but Q holds.

Warning 8.7. If you are trying to *show* $P \vee Q$, then you do **not** just consider the two cases P or Q. You do not yet know those are the only two options!

However, you do know that P or $\neg P$ happens. So, perhaps your two cases could be as follows. Case 1 is when P holds, and you are done. Case 2 is when P fails, and try to show Q now holds.

Here is another example of these ideas.

Proposition 8.8. *Let $x \in \mathbb{Z}$. Then $x^2 \mid x$ if and only if $x \in \{-1, 0, 1\}$.*

Proof. Let $x \in \mathbb{Z}$. We are proving a biconditional, so we need to prove both directions.

($\Rightarrow$): We will work directly. Assume $x^2 \mid x$. Thus $x = x^2 y$ for some $y \in \mathbb{Z}$. There are two cases we will consider.

Case 1: Suppose $x = 0$. Then we are done!

Case 2: Suppose $x \neq 0$. Then, we can divide by x, and get $1 = xy$. The only divisors of 1 are ± 1. Thus $x = 1$ or $x = -1$.

In every case $x \in \{-1, 0, 1\}$.

($\Leftarrow$) : We work directly again. Assume $x \in \{-1, 0, 1\}$. There are thus three cases: $x = -1$, $x = 0$, or $x = 1$. In all three cases, we can check immediately that $x^2 \mid x$. $\square$

Question: In the backwards direction, ($\Leftarrow$), we had three cases. Why didn't we have those three cases in the forward direction?
Answer: In the backwards direction we made the assumption $x \in \{-1, 0, 1\}$, which limited the possibilities for x to three cases. In the forward direction we didn't have such an assumption. So we had to consider every possibility.

The following theorem is very useful and also demonstrates these same ideas again.

Theorem 8.9. *Given $a, b \in \mathbb{C}$, we have $ab = 0$ if and only if $a = 0$ or $b = 0$.*

Proof. This is a biconditional, so we will prove both directions.

($\Rightarrow$): We work directly. Assume $ab = 0$. There are two possibilities for a, given by the following two cases.

Case 1: Suppose $a = 0$. Then we are done.

Case 2: Suppose $a \neq 0$. [Note: Why not suppose $b = 0$?] Since $a \neq 0$ we can divide the equality $ab = 0$ on both sides by a, to get $b = 0$.

Thus, in every case, either $a = 0$ or $b = 0$.

($\Leftarrow$): We now prove the converse. We again work directly, so we assume $a = 0$ or $b = 0$. There are two cases.

Case 1: Suppose $a = 0$. Then $ab = 0b = 0$.

Case 2: Suppose $b = 0$. Then $ab = a0 = 0$.

In every case $ab = 0$. □

What does this theorem mean? The following example shows one way to understand how to use it.

Example 8.10. We will show that the only complex solutions of $x^2 = 3x - 2$ are $x = 1$ or $x = 2$.

If $x^2 = 3x - 2$, then $x^2 - 3x + 2 = 0$. Factoring, this means $(x - 1)(x - 2) = 0$. By the previous theorem (take $a = x - 1$ and $b = x - 2$), we then must have $x - 1 = 0$ or $x - 2 = 0$. In other words, $x = 1$ or $x = 2$. △

8.B Congruence

If the clock on the wall says 9:00, and 37 hours pass, what time is it then? It isn't too difficult to figure out that the new time is 10:00. The way we figure this out is to notice that each 12 hour block keeps the clock fixed, and so 37 hours looks the same as 1 hour.

We do something similar when working with even and odd numbers. We know that an even number plus an odd number will always be odd. The reason is that adding any multiple of 2 doesn't change whether a number is odd or even.

These two situations are special cases of a much more general, and powerful, technique. In the first situation, we are looking at numbers and treating multiples of $n = 12$ as trivial. In the second case, we are treating multiples of $n = 2$ as trivial blocks. The following definition does this for an arbitrary integer n.

Definition 8.11. Let $a, b, n \in \mathbb{Z}$. We say that a *is congruent to* b *modulo* n if $n \mid (a - b)$. In other words, $a - b = nk$ for some $k \in \mathbb{Z}$. Or, in other words, a and b differ by a multiple of n. We write

$$a \equiv b \pmod{n}$$

to denote that a is congruent to b modulo n.

In the following example we work out some instances where this definition holds or does not hold.

Example 8.12. (1) Is $7 \equiv 3 \pmod 2$? Yes, we have $2 \mid (7-3)$.
(2) Is $3 \equiv 16 \pmod{31}$? This is asking, does $31 \mid (3-16)$? No, $31 \nmid (-13)$. Thus we write $3 \not\equiv 16 \pmod{31}$.
(3) Is $21 \equiv 9 \pmod{12}$? This is asking, does $12 \mid (21-9)$? Yes, $12 \mid 12$. Notice that this question is really asking whether 21 hours looks like 9 hours on a clock.
(4) Is $2 \equiv -2 \pmod 3$? This is asking does $3 \mid (2-(-2))$? No, $3 \nmid 4$. So we write $2 \not\equiv -2 \pmod 3$.
(5) Is $95482 \equiv 2892 \pmod 2$? We can see that the answer will be yes, since both numbers are even. However, we double-check that $2 \mid (95482 - 2892)$, which is true.
(6) If x is an integer, is $x \equiv x \pmod n$? We are asking whether $n \mid (x-x)$. The answer is yes, since $n \mid 0$. △

In many cases, we can work with congruences almost as if they were equations. We will later prove several theorems of this sort; the following proposition is an example.

Proposition 8.13. *Let $a, b, c, n \in \mathbb{Z}$. If $a \equiv b \pmod n$, then $ac \equiv bc \pmod n$.*

Proof. Let $a, b, c, n \in \mathbb{Z}$. Assume $a \equiv b \pmod n$. Thus $n \mid (a-b)$. In other words, $a - b = nk$ for some $k \in \mathbb{Z}$. Multiplying by c, we get $ac - bc = nkc$. Hence $n \mid (ac - bc)$, so $ac \equiv bc \pmod n$ as desired. □

This proposition says that when numbers are congruent, we can multiply by any integer and they stay congruent. For instance, we have $5 \equiv -9 \pmod 7$. Multiply by 5 to get $25 \equiv -45 \pmod 7$.

Example 8.14. Another way to think about congruence is that two numbers are congruent modulo n if they have the same remainder when we divide by n. When we divide by 2, there are only two remainders, so every $x \in \mathbb{Z}$ is either odd or even. In other words,

$$x \equiv 0 \pmod 2 \qquad \text{or} \qquad x \equiv 1 \pmod 2.$$

What happens if we work modulo 3? Now there are three remainders, and we get

$$x \equiv 0 \pmod 3, \qquad x \equiv 1 \pmod 3, \qquad \text{or } x \equiv 2 \pmod 3.$$

In other words, every integer x is of exactly one of the forms $3k$, $3k+1$, or $3k+2$ for some $k \in \mathbb{Z}$. (We will prove this later, but you can use it freely for now.) △

The previous example tells us that sometimes we can reduce questions about division into cases according to remainders!

Proposition 8.15. *If $x \in \mathbb{Z}$, then $x^3 \equiv x \pmod 3$.*

Proof. Let $x \in \mathbb{Z}$. We consider the following three cases.

Case 1: Suppose $x = 3k$ for some $k \in \mathbb{Z}$. Then

$$x^3 - x = (3k)^3 - 3k = 27k^3 - 3k = 3(9k^3 - k).$$

So $3 \mid (x^3 - x)$, hence $x^3 \equiv x \pmod 3$.

Case 2: Suppose $x = 3k + 1$ for some $k \in \mathbb{Z}$. Then

$$x^3 - x = (3k + 1)^3 - (3k + 1) = 27k^3 + 27k^2 + 9k + 1 - 3k - 1 = 3(9k^3 + 9k^2 + 2k).$$

So $3 \mid (x^3 - x)$, hence $x^3 \equiv x \pmod 3$.

Case 3: Suppose $x = 3k + 2$ for some $k \in \mathbb{Z}$. Then

$$x^3 - x = (3k+2)^3 - (3k+2) = 27k^3 + 54k^2 + 36k + 8 - 3k - 2 = 3(9k^3 + 18k^2 + 11k + 2).$$

So $3 \mid (x^3 - x)$, hence $x^3 \equiv x \pmod 3$. □

Example 8.16. The previous proposition asserts that $3 \mid (1037^3 - 1037)$. Check it! △

8.C Absolute values

One important place where the method of proof by cases arises is in proving statements about absolute values. Indeed, the very definition of the absolute value of a number is given in terms of cases.

Definition 8.17. Let $a \in \mathbb{R}$. Define

$$|a| = \begin{cases} a & \text{if } a \geq 0, \\ -a & \text{if } a < 0. \end{cases}$$

Example 8.18. We have $|2| = 2$, and $|-2| = 2$. If $x = -5$, then $|x| = -x = 5$. △

Warning 8.19. Many students have difficulty with the idea that $|x| = -x$, which is true anytime that $x < 0$. Part of this difficulty is that they think that an expression beginning with a negative sign (such as $-x$) must be negative. However, if x is negative, then $-x$ is positive.

Another instance where this problem can come up is if we compute $|-x|$. We note that this is not necessarily equal to x. In particular, if x is negative $|-x|$ *never* equals x.

We will now prove some statements involving absolute values. Several of these are important enough to be called Theorems (and one is even important enough to have a name).

Proposition 8.20. *Let $a, b, x \in \mathbb{R}$ and assume that $b \geq 0$. If*

$$|x - a| \leq b,$$

then $a - b \leq x \leq a + b$.

Proof. Suppose that $|x - a| \leq b$. We divide the proof into cases.

Case 1: Suppose that $x - a \geq 0$. Then $|x - a| = x - a$, so we have that $0 \leq x - a \leq b$. Adding a, we find that $a \leq x \leq b + a$. Since $a - b \leq a$, we see that

$$a - b \leq x \leq a + b.$$

Case 2: Suppose that $x - a < 0$. Then $0 < |x - a| = -(x - a) = a - x$. Hence, $0 < a - x \leq b$. Multiplying by -1, we get $-b \leq x - a < 0$. Now adding a, we get $a - b \leq x < a$. Since $a \leq a + b$, we have

$$a - b \leq x \leq a + b.$$

In both cases we obtain the desired inequalities. □

Theorem 8.21 (The Triangle Inequality)**.** *Let $x, y \in \mathbb{R}$. Then $|x+y| \leq |x|+|y|$.*

Proof. Without loss of generality, we may assume that $x \geq y$, so that if only one of x, y is nonnegative, it is x. We divide the remaining possibilities into four cases.

Case 1: Suppose that $x \geq 0$ and $y \geq 0$. Then $x + y \geq 0$, so

$$|x + y| = x + y = |x| + |y|.$$

Case 2: Suppose that $x < 0$ and $y < 0$. Then $x + y < 0$, so

$$|x + y| = -(x + y) = (-x) + (-y) = |x| + |y|.$$

Case 3: Suppose that $x \geq 0$, $y < 0$, and $x + y \geq 0$. Then we have

$$|x + y| = x + y = |x| - |y| < |x| + |y|$$

where the inequality arises because $-|y| < |y|$ (and one can just add $|x|$ to both sides).

Case 4: Suppose that $x \geq 0$, $y < 0$, and $x + y < 0$. Then

$$|x + y| = -(x + y) = (-x) + (-y) = -|x| + |y| \leq |x| + |y|.$$

Hence, in all four cases, the theorem is true. □

Theorem 8.22. *Let $x, y \in \mathbb{R}$. Then $|xy| = |x||y|$.*

Proof. This proof is left as Exercise 8.8. □

8.D Exercises

Exercise 8.1. Let $x, y \in \mathbb{Z}$. Prove that if x and y have the same parity, then $x^2 + xy$ is even.

Exercise 8.2. Let $a, b, c \in \mathbb{Z}$. Prove that if $a \nmid bc$, then $a \nmid b$ and $a \nmid c$. (The converse is not true. Can you see why?)

Exercise 8.3. Do the following:
(a) Prove that given $x \in \mathbb{Z}$, either $x^2 \equiv 0 \pmod 4$ or $x^2 \equiv 1 \pmod 4$.
(b) Prove that for any integer x we have $4 \mid (x^4 - x^2)$.

Exercise 8.4. Let $a, b, c, n \in \mathbb{Z}$. If $a \equiv b \pmod n$ and $b \equiv c \pmod n$, show that $a \equiv c \pmod n$.

If we know $11 \equiv -3 \pmod 7$ and $-3 \equiv 4 \pmod 7$, can we say that $11 \equiv 4 \pmod 7$?

Exercise 8.5. Prove, for any $n \in \mathbb{Z}$, that $3 \mid n$ if and only if $3 \mid n^2$. (Hint: Use the idea in Example 8.14 to divide the proof into cases.)

Exercise 8.6. Prove $3 \mid (2n^2 + 1)$ if and only if $3 \nmid n$, for $n \in \mathbb{Z}$.

Exercise 8.7. Let $a, b, c, d, n \in \mathbb{Z}$. If $a \equiv b \pmod n$ and $c \equiv d \pmod n$, prove that $ac \equiv bd \pmod n$.

What does this statement say if we take $c = a$ and $d = b$?

We know that $19 \equiv 5 \pmod 7$. Do we then know $19^2 \equiv 5^2 \pmod 7$? How about $19^3 \equiv 5^3 \pmod 7$?

Exercise 8.8. Prove Theorem 8.22; for any $x, y \in \mathbb{R}$, we have $|xy| = |x||y|$.

Exercise 8.9. Let $a \in \mathbb{R}$. Prove that $a^2 \leq 1$ if and only if $-1 \leq a \leq 1$. In the proof you may use the following two facts that are true for any $a, b, c \in \mathbb{R}$.
(1) If $a < b$ and $c > 0$, then $ac < bc$.
(2) If $a < b$ and $c < 0$, then $ac > bc$.

9 Proof by contradiction

9.A Basic technique and examples

In this section we explore a proof technique that can be applied not only to implications but to other statements as well; the technique is called "proof by contradiction." It is based upon the following simple idea:

Theorem 9.1. *Let R and S be statements. If $(\neg R) \Rightarrow S$ is true and S is false, then R must be true.*

Proof. If the conclusion of an implication is false, the only way for the implication to be true is if the premise is also false. Hence $\neg R$ is false. But this means R is true. (Alternatively, draw a truth table.) □

Another way to think about this theorem is that if by assuming $\neg R$ we can reach some false statement, then R must have been true after all. (If an assumption leads to nonsense, that assumption must have been false.) We will see this proof technique in action, by proving the following proposition.

Proposition 9.2. *There is no smallest positive rational number.*

Proof. Assume, by way of contradiction, that there is a smallest positive rational number; call it $r \in \mathbb{Q}$. Then $r/2$ is a rational number, still positive, and $r/2 < r$. This is impossible, because r was supposed to be the smallest element with these properties. This contradiction shows that our assumption must have been false, so there is no smallest positive rational number. □

Proof by contradiction is a very powerful technique, because it applies to many different types of statements. However, it is also limited in two ways:

(1) By assuming $\neg R$ you are working in an *imaginary* world, where you are pretending R is false (even though you believe R is true). Your ultimate goal is to find some false statement or contradiction that follows from your assumption. The only conclusion you can make after finding that false statement is that R must have been true after all (since you were working in an imaginary world).

(2) Sometimes it can be difficult to *find* the contradiction that arises from the assumption $\neg R$. Unlike direct proofs or contrapositive proofs, we do not know where we are headed. It can take practice to get a feeling for what to search for.

9.B Proof by contradiction for implications

When R is a statement of the form

$$R: \ \forall x, \ P(x) \Rightarrow Q(x)$$

then the negation is

$$\neg R: \ \exists x, \ P(x) \wedge \neg Q(x).$$

Thus, to do a proof by contradiction in the case of an implication, you assume the premise and the negation of the conclusion, and then search for a contradiction.

Advice 9.3. An easy way to remember what to assume when proving an implication by contradiction is that you have the same assumptions as in *both* a direct proof and a contrapositive proof.

Here is an example of how to do an "even/odd" proof by contradiction.

Proposition 9.4. *Let $x \in \mathbb{Z}$. If $2 \mid x$ then $2 \nmid (x^2+1)$.*

Proof. Assume, by way of contradiction, that $2 \mid x$ and $2 \mid (x^2+1)$ for some $x \in \mathbb{Z}$. We can then write $x = 2y$ and $x^2 + 1 = 2z$ for some $y, z \in \mathbb{Z}$. Plugging the first equality into the second, we get $4y^2 + 1 = 2z$. Therefore $1 = 2z - 4y^2 = 2(z - 2y^2)$ is even, which is false. Thus, the original implication is true. □

Here is another example where either a direct proof or a contrapositive proof would be very difficult to accomplish (unless you use a lemma to help).

Proposition 9.5. *Given $x \in \mathbb{Z}$, if $x^2 + 2x - 3$ is odd then $x^2 + 4x - 5$ is odd.*

Proof. Assume, by way of contradiction, that $x^2 + 2x - 3$ is odd and $x^2 + 4x - 5$ is even, for some $x \in \mathbb{Z}$. Then we can write $x^2+2x-3 = 2k+1$ and $x^2+4x-5 = 2\ell$ for some $k, \ell \in \mathbb{Z}$. Subtracting each side of the second equation from the corresponding side of the first equation, we obtain

$$(x^2 + 2x - 3) - (x^2 + 4x - 5) = 2k + 1 - 2\ell.$$

After simplifying, we obtain $1 = 2x + 2k - 2\ell$ is even, a contradiction. □

Contradiction proofs can also involve cases. We just need to check that every case ends in a contradiction, in order to show that that case could not have happened after all. The following result demonstrates this idea.

Proposition 9.6. *If $x \in \mathbb{Z}$ is even, then x is not the sum of three integers with an odd number of them being odd.*

Proof. Assume, by way of contradiction, that there is some even integer x that is the sum of three integers $a, b, c \in \mathbb{Z}$, an odd number of them being odd. There are two cases.

Case 1: Assume all three are odd. Then $a = 2k+1$, $b = 2\ell+1$, and $c = 2m+1$ for some $k, \ell, m \in \mathbb{Z}$. Adding, we get

$$x = a + b + c = 2k + 1 + 2\ell + 1 + 2m + 1 = 2(k + \ell + m + 1) + 1$$

is odd, contradicting the fact that x is even.

Case 2: Assume one of the three numbers is odd and that the other two are even. Without loss of generality, suppose c is the odd one. Then $a = 2k$, $b = 2\ell$, and $c = 2m + 1$ for some $k, \ell, m \in \mathbb{Z}$. Adding, we get that

$$x = a + b + c = 2k + 2\ell + 2m + 1 = 2(k + \ell + m) + 1$$

is odd, contradicting the fact that x is even.

Thus, in every case, we reached a contradiction. □

9.C Irrationality proofs

In this subsection we will focus on proofs which involve the following definition:

Definition 9.7. A real number r is *irrational* if it is not rational. In other words, $r \in \mathbb{R} - \mathbb{Q}$.

Do we know that there exist any irrational numbers? Yes, and this fact was proved thousands of years ago by the Pythagoreans. (Legend has it that the mathematician who originally proved this fact was either killed or exiled for the proof, since it ran counter to the doctrine of the times!) Here is the proof, essentially unchanged from that time.

Theorem 9.8. *The real number $\sqrt{2}$ is irrational.*

Proof. Assume, by way of contradiction, that $\sqrt{2} \in \mathbb{Q}$. We can then write $\sqrt{2} = a/b$ for some $a, b \in \mathbb{N}$, with a/b in lowest terms. By squaring and then clearing denominators, we have $a^2 = 2b^2$. Thus $2 \mid a^2$, and hence $2 \mid a$ by Proposition 7.16. Write $a = 2x$ for some $x \in \mathbb{Z}$.

Plugging $a = 2x$ into the equality $a^2 = 2b^2$ yields $4x^2 = 2b^2$, or in other words $b^2 = 2x^2$. Thus $2 \mid b^2$, and hence $2 \mid b$. However, now a and b are both even which contradicts the fact that a/b was assumed to be in lowest terms. Hence $\sqrt{2}$ is irrational. □

9.D Advice

We end this section with two pieces of advice.

First, sometimes one can tell that a result could be proved by contradiction because the statement R itself has some negative sounding words. For instance, in this section we proved the following statements R:

- There is **no** smallest positive rational number.

- The number $\sqrt{2}$ is **not** rational.
- If x is even, then 2 does **not** divide $x^2 + 1$.

It is usually easier to work with *positive* conditions, rather than negative conditions, which is why proofs by contradiction work so well in these cases. The negative conditions are turned positive after negating R.

Second, if a proof can be done without contradiction, then that is usually a better option, because you never enter an "imaginary" world where you assume something you are hoping to show is false.

9.E Exercises

Exercise 9.1. Let R and S be statements. Draw a truth table with columns labeled R, S, $\neg R$, and $(\neg R) \Rightarrow S$. Verify that the only row where S is false and $(\neg R) \Rightarrow S$ is true occurs when R is true.

Exercise 9.2. Prove the following statement directly, contrapositively, and by contradiction: Given $x \in \mathbb{Z}$, if $3x + 1$ is even, then $5x + 2$ is odd.

Exercise 9.3. Prove, by way of contradiction, the following statement: Given $a, b, c \in \mathbb{Z}$ with $a^2 + b^2 = c^2$, then a is even or b is even. (Hint: Consider Exercise 8.3.)

Exercise 9.4. Prove that $\sqrt{3}$ is irrational.

Exercise 9.5. Prove that $\sqrt[3]{2}$ is irrational.

Exercise 9.6. Prove: If $x \in \mathbb{Q}$ and $y \in \mathbb{R} - \mathbb{Q}$, then $x + y \in \mathbb{R} - \mathbb{Q}$.

Exercise 9.7. Prove: If we are given a nonzero rational number x and an irrational number y, then the number xy is irrational. (Hint: Your proof should, somewhere, use the fact that $x \neq 0$, because when $x = 0$ the conclusion is false.)

Exercise 9.8. Prove there is no smallest positive irrational number. (Hint: Use the result of the previous exercise.)

Exercise 9.9. Given $x, y \in \mathbb{Z}$, prove that $33x + 132y \neq 57$.

10 Proofs in set theory

A fundamental skill when doing proofs is handling sets. The methods developed in the previous few sections will be extremely useful in this regard. In this section we focus on three skills: (1) How to prove $x \in A$. (2) How to prove $A \subseteq B$. (3) How to prove $A = B$.

10.A Proving set membership

If we want to prove x is an element of some set A, the proof might depend on how we describe A. For instance, if A is given as a list of elements $\{1, 2, 3\}$, then we just have to check that x is one of those elements in the list. When A is described using set-builder notation, then we must verify that x satisfies all of the properties for elements in A. Here is an example.

Proposition 10.1. *Let $S = \{n : n \text{ is an odd integer}\}$. Then $3 \in S$.*

Proof. The number 3 is an odd integer, since $3 = 2 \cdot 1 + 1$. Thus $3 \in S$. □

Suppose $T = \{x \in \mathbb{N} : x \text{ is a squared integer}\}$. Is $0 \in T$? The answer is no. It is true that 0 is a squared integer, but there is a second requirement for elements of T, they must belong to $\mathbb{N}$. Thus $0 \notin T$ because $0 \notin \mathbb{N}$.

Here is one final example of proving that an element belongs to a set.

Proposition 10.2. $(28, 6) \in \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \equiv y \pmod{11}\}$.

Proof. First, we see that since 28 and 6 are integers, we have $(28, 6) \in \mathbb{Z} \times \mathbb{Z}$.

Second, we check directly that $28 - 6 = 22$ is divisible by 11. Hence $28 \equiv 6 \pmod{11}$.

Therefore $(28, 6)$ satisfies all of the conditions to belong to this set. □

10.B Proving inclusion of sets

Let A and B be sets. What does $A \subseteq B$ mean? It means that every element of A is an element of B. In other words

$$\forall x,\ x \in A \Rightarrow x \in B.$$

Most often we prove this implication using a direct proof. The steps are simple.

(1) Assume $x \in A$.
(2) Using that information, show $x \in B$.

We demonstrate how this is to be done with a few examples.

Proposition 10.3. *Let* $A = \{x \in \mathbb{Z} : 4 \,|\, x\}$ *and* $B = \{x \in \mathbb{Z} : 2 \,|\, x\}$. *It holds that* $A \subseteq B$.

Proof. Assume $a \in A$. Thus $a \in \mathbb{Z}$ and $4 \,|\, a$. We can write $a = 4m$ for some $m \in \mathbb{Z}$. Therefore $a = 2(2m)$, hence $2 \,|\, a$. We have demonstrated that a satisfies all the properties to belong to B, so $a \in B$.

As $a \in A$ was arbitrary, we have now shown $A \subseteq B$. □

In the next example we will prove that $\overline{S \cup T} \subseteq \overline{S} \cap \overline{T}$. (Drawing a Venn diagram helps us believe this is true, but we still need to give the formal proof.) The method we use is still the same: we will assume x is an element of the (hypothetically) smaller set $\overline{S \cup T}$ and then prove it belongs to the bigger set $\overline{S} \cap \overline{T}$.

Proposition 10.4. *Given sets* S *and* T, *then* $\overline{S \cup T} \subseteq \overline{S} \cap \overline{T}$.

Proof. Assume $x \in \overline{S \cup T}$. Thus $\neg(x \in S \cup T)$. In other words $\neg(x \in S$ or $x \in T)$. Using De Morgan's law, we have $\neg(x \in S)$ and $\neg(x \in T)$. In other words $x \in \overline{S}$ and $x \in \overline{T}$. Using the definition of intersection, we have $x \in \overline{S} \cap \overline{T}$. □

Warning 10.5. If instead of using $\neg(x \in A)$ you use the symbols $x \notin A$, be very careful! The statement $x \notin X \cup Y$ means that $x \notin X$ **and** $x \notin Y$, even though a union represents "or".

In the next example, we make use of the tautology $P \Rightarrow P \vee Q$. (If P is true, then P is true or Q is true.)

Proposition 10.6. *Given sets* S *and* T, *then* $S \subseteq S \cup T$.

Proof. Assume $x \in S$. By the tautology mentioned above we know that $x \in S$ or $x \in T$. Therefore, $x \in S \cup T$ by the definition of union. □

In summary, when proving a set inclusion, start with assuming x is an arbitrary element of the smaller set, and then prove it also must belong to the bigger set. Note: You do not need to always use the letter x, especially if some other letter makes more sense for the problem at hand.

10.C Proving equality

When are two sets equal? This happens when they have exactly the same elements. In other words $A = B$ means

$$\forall x,\ x \in A \Leftrightarrow x \in B.$$

This is a biconditional, which we often prove by doing each direction separately. Here is an example of how this can be done.

Proposition 10.7. *Let* $A = \{x \in \mathbb{R} : x^2 \leq 1\}$ *and* $B = \{x \in \mathbb{R} : -1 \leq x \leq 1\}$. *We have* $A = B$.

Proof. We first prove $A \subseteq B$. Assume $a \in A$. Thus $a \in \mathbb{R}$ and $a^2 \leq 1$. The second condition is equivalent to $-1 \leq a \leq 1$, by Exercise 8.9. Thus $a \in B$.

Conversely, we now show $B \subseteq A$. Assume $b \in B$. Thus $b \in \mathbb{R}$ and $-1 \leq b \leq 1$. As mentioned above, the second condition is equivalent to $b^2 \leq 1$. Hence $b \in A$. □

In some situations it takes more work to show that two sets are equal. For instance we have:

Proposition 10.8. $\{x \in \mathbb{Z} : 6 \,|\, x\} = \{x \in \mathbb{Z} : 2 \,|\, x\} \cap \{x \in \mathbb{Z} : 3 \,|\, x\}$.

Proof. Let $A = \{x \in \mathbb{Z} : 6 \,|\, x\}$, $B = \{x \in \mathbb{Z} : 2 \,|\, x\}$, and $C = \{x \in \mathbb{Z} : 3 \,|\, x\}$. We first prove $A \subseteq B \cap C$. Let $a \in A$. Thus $a \in \mathbb{Z}$ and $6 \,|\, a$. The second condition means $a = 6k$ for some $k \in \mathbb{Z}$. Hence $a = 2(3k)$ and $a = 3(2k)$, so $2 \,|\, a$ and $3 \,|\, a$. We have shown that $a \in B$ and $a \in C$ (since it satisfies all the necessary conditions for set membership). Therefore $a \in B \cap C$.

Conversely, we now show that $B \cap C \subseteq A$. Fix $z \in B \cap C$. Thus $z \in B$ and $z \in C$. Hence $z \in \mathbb{Z}$, and $2 \,|\, z$, and $3 \,|\, z$. Thus $z = 2k$ and $z = 3\ell$ for some $k, \ell \in \mathbb{Z}$. Since z is even, ℓ must be even. (You can prove this, contrapositively.) Hence $\ell = 2m$ for some $m \in \mathbb{Z}$. Thus $z = 3\ell = 3(2m) = 6m$. So $6 \,|\, z$. Therefore $z \in A$. □

We finish with one more example.

Proposition 10.9. *Given sets* X, Y, *and* Z, *we have* $X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$.

Proof. We first prove the inclusion $X \cup (Y \cap Z) \subseteq (X \cup Y) \cap (X \cup Z)$. Assume $a \in X \cup (Y \cap Z)$. Therefore $a \in X$ or $a \in Y \cap Z$. We deal with those cases separately.

Case 1: Assume $a \in X$. Then, by tautology, $a \in X$ or $a \in Y$. Hence $a \in X \cup Y$, by the definition of union. By similar reasoning $a \in X \cup Z$. Thus $a \in (X \cup Y) \cap (X \cup Z)$ by the definition of intersection.

Case 2: Assume $a \in Y \cap Z$. Thus $a \in Y$ and $a \in Z$. By tautology we know $a \in X$ or $a \in Y$, hence $a \in X \cup Y$ from the definition of union. Similarly, $a \in X \cup Z$. Therefore, $a \in (X \cup Y) \cap (X \cup Z)$ by definition of intersection.

In both cases, we have shown $a \in (X \cup Y) \cap (X \cup Z)$. This proves the needed inclusion.

We now show, conversely, that $(X \cup Y) \cap (X \cup Z) \subseteq X \cup (Y \cap Z)$. Assume $b \in (X \cup Y) \cap (X \cup Z)$. We have $b \in X \cup Y$ and $b \in X \cup Z$, from the definition of intersection. Hence, $b \in X$ or $b \in Y$, and we also have $b \in X$ or $b \in Z$. There are two cases to consider.

Case 1: Assume $b \in X$. Thus (by a tautology) $b \in X$ or $b \in (Y \cap Z)$. Hence $b \in X \cup (Y \cap Z)$.

Case 2: Assume $b \notin X$. Then, from our work above we know that $b \in Y$ and $b \in Z$. Hence $b \in Y \cap Z$. By tautology, $b \in X$ or $b \in Y \cap Z$. Therefore $b \in X \cup (Y \cap Z)$, by definition of union.

In every case we proved $b \in X \cup (Y \cap Z)$, so we have proved the needed inclusion. □

10.D Laws for sets

The following theorem lists some of the most useful set equalities. It can be useful to try to prove for oneself a few of these equalities. (Some are silly to prove, like the commutative laws.)

Theorem 10.10. *Let A, B, and C be sets. Assume that they all are subsets of some universal set U. The following properties hold:*

- *Commutative laws.*
  $A \cap B = B \cap A$.
  $A \cup B = B \cup A$.
- *Associative laws.*
  $(A \cap B) \cap C = A \cap (B \cap C)$.
  $(A \cup B) \cup C = A \cup (B \cup C)$.
- *Distributive laws.*
  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
- *Identity laws.*
  $A \cup \emptyset = A$.
  $A \cap U = A$.
- *Complement laws.*
  $A \cup \overline{A} = U$.
  $A \cap \overline{A} = \emptyset$.

It is a deep fact that every other set equality involving *only* unions, intersections, and complements, can be derived (algebraically) from these laws. However, there are a few other common properties that can be proved quite easily just using the methods of this section. (In fact, we proved part of one of them earlier!)

Theorem 10.11. *Let A and B be sets. Assume that they all are subsets of some universal set U. The following properties hold:*

- *De Morgan's laws.*
  $\overline{A \cup B} = \overline{A} \cap \overline{B}$.
  $\overline{A \cap B} = \overline{A} \cup \overline{B}$.
- *Double negation.*
  $\overline{\overline{A}} = A$.

10.E Exercises

Exercise 10.1. For each element and set listed below, explain why the element does or does not belong to the set.
(a) Is $3 \in \{1, 2, 3, 4, 5, 6, 7\}$?
(b) Is $\pi \in \{1, 2, 3, 4, 5, 6, 7\}$?
(c) Is $\pi \in \mathbb{R}$?
(d) Is $2/3 \in \{x \in \mathbb{R} : x < 1\}$?
(e) Is $2/3 \in \{x \in \mathbb{Z} : x < 1\}$?

Exercise 10.2. Fix

$$A = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : 4 \,|\, (x - y)\}$$

and fix

$$B = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \text{ and } y \text{ have the same parity}\}.$$

Prove $A \subseteq B$.

Exercise 10.3. Let X be the set of integers which are congruent to -1 modulo 6 and let Y be the set of integers which are congruent to 2 modulo 3. Prove $X \subseteq Y$.

Exercise 10.4. Let A and B be sets inside some universal set U.
(a) Prove that $\overline{A \cap B} \subseteq \overline{A} \cup \overline{B}$.
(b) Prove that $\overline{A} \cup \overline{B} \subseteq \overline{A \cap B}$.
(c) Putting those two previous parts together, what have you proved?

Exercise 10.5. Let X and Y be sets. Prove $X - (X - Y) \subseteq X \cap Y$. (Hint: Remember that $s \in S - T$ means $s \in S$ and $s \notin T$. Thus, $s \notin S - T$ means $s \notin S$ or $s \in T$.)

Exercise 10.6. Given a set X, prove that $X \cup \emptyset = X$. (Hint: If you have a case where $x \in \emptyset$, then you know that case doesn't actually happen.)

Exercise 10.7. Let $n \in \mathbb{Z}$. Prove that

$$\{x \in \mathbb{Z} : n \,|\, x\} = \{x \in \mathbb{Z} : x \equiv 0 \pmod{n}\}.$$

Exercise 10.8. Let A, B, and C be sets. Prove that

$$A - (B \cap C) \subseteq (A - B) \cup (A - C).$$

Is the other inclusion true?

Exercise 10.9. For each $n \in \mathbb{N}$, define $S_n = \{m \in \mathbb{Z} : m \leq n\}$. Prove that

$$\bigcup_{n \in \mathbb{N}} S_n = \mathbb{Z}.$$

(Recall that, by Definition 2.9,

$$x \in \bigcup_{i \in I} S_i$$

means that $x \in S_i$ for some $i \in I$.)

11 Existence proofs and counterexamples

In the previous sections we have mainly proved universal statements:

$$\forall x \in S,\ R(x).$$

In this section we focus our attention on proving *existential* statements:

$$\exists x \in S,\ R(x).$$

Often these statements are significantly easier to prove than universal statements, because you just need to find a single example to demonstrate their truth.

11.A Constructive existence proofs

When attempting to prove $\exists x \in S,\ R(x)$, it is sometimes possible to find an element x of the set S satisfying the property $R(x)$. Here are some instances where we can do this.

Example 11.1. (1) The statement $\exists x \in \mathbb{R},\ x > 1$ is true, because $x = 3$ is an example. (There are many more examples, we just needed to find one.)

(2) The statement $\exists n \in \mathbb{N},\ n^3 < 2$ is true because $n = 1$ is an example. (In this case, the number $n = 1$ is the only example.)

(3) The statement $\exists x \in \mathbb{R},\ x^2 < x$ is true because $x = 0.5$ is an example. △

When proving a statement with universal quantifiers $\forall x \in S,\ R(x)$, we have in the past started the proof with the sentence: "Let $x \in S$." This is shorthand for saying "Let x be an *arbitrary* element of the set S." When working with existence quantifiers, we will try to avoid the word "let" and use words like "fix," "put," or "set" as in the following example.

Proposition 11.2. *There exists an integer n such that $n^3 = n^2$.*

Proof. Fix $n = 0$. We see that $n^3 = n^2 = 0$, as needed. □

Warning 11.3. Some mathematicians use the word "let" when handling existential statements. For instance, they might have started the previous proof with the sentence "Let $n = 0$."

In some cases a statement asks for the existence of more than one element.

Proposition 11.4. *There exist $a, b \in \mathbb{N}$ such that $a^2 \mid b^3$ but $a \nmid b$.*

Proof. Fix $a = 8$ and $b = 4$. Since $64 \mid 64$ we have $a^2 \mid b^3$. However $8 \nmid 4$, so $a \nmid b$. □

You might wonder what to do when a statement has more than one quantifier. In that case, you deal with each quantifier as it arises. The following proposition and proof show how this is to be done.

Proposition 11.5. *Every odd integer is the sum of two consecutive integers.*

Proof. We are proving the following statement:

$$\forall a \in \mathbb{Z},\ a \text{ is odd} \Rightarrow\ (\exists b \in \mathbb{Z},\ a = b + (b+1))\,.$$

Let $a \in \mathbb{Z}$. Assume a is odd. Hence $a = 2k + 1$ for some $k \in \mathbb{Z}$.

Fix $b = k$. Then $b + (b + 1) = k + (k + 1) = 2k + 1 = a$. Hence a is the sum of two consecutive integers. □

11.B Nonconstructive existence proofs

Every existence problem in the previous section was solved by finding an explicit example. This is also called *constructing* an example, or giving a *constructive proof*. Sometimes it is possible to prove an object exists without constructing an example. Here is one instance.

Proposition 11.6. *One of the digits of* $\sqrt{2} = 1.414213562\ldots$ *occurs infinitely many times.*

Proof. There are only finitely many possible digits

$$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

but there are infinitely many decimal places. □

Notice that in this example we did not actually find out which digit occurs infinitely often, which is why we say the proof is *nonconstructive*. We only proved that at least one of the digits does show up over and over. It is not much more difficult to show that in fact two digits must occur infinitely many times. (*Sketch:* If, by way of contradiction, we assume that only one digit showed up infinitely many times, then the decimal expansion for $\sqrt{2}$ would eventually just repeat that digit. This would show that $\sqrt{2}$ is rational, which we previously proved it is not.) Quite surprisingly, it is an open problem in mathematics whether three digits must occur infinitely often in the decimal expansion of $\sqrt{2}$!

Here is another example, where we *almost* construct an example.

Proposition 11.7. *There exist irrational numbers* a, b *such that* a^b *is rational.*

Proof. We will consider two possible cases.

Case 1: Assume $\sqrt{2}^{\sqrt{2}}$ is rational. In this case we fix $a = b = \sqrt{2}$ and have that a^b is rational.

Case 2: Assume $\sqrt{2}^{\sqrt{2}}$ is irrational. In this case we fix $a = \sqrt{2}^{\sqrt{2}}$ and $b = \sqrt{2}$, and calculate

$$a^b = \left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} = \sqrt{2}^2 = 2$$

which is rational. □

In this proof, we do not know (or care) which case is true. We just show that we can solve the problem in either case. (If you do care, it is known that Case 2 is the true case, but this is not easy to prove.) There does exist a constructive proof of the previous proposition (take $a = \sqrt{2}$ and $b = \log_2(9)$ and prove b is irrational).

We finish with one more example of how to prove existence (in a nonconstructive way). In the proof below we use some standard theorems from calculus which we will not prove in this textbook.

Proposition 11.8. *The equation $x^5 + 2x - 1 = 0$ has a solution in the interval $(0, 1)$.*

Proof. The function given by the polynomial equation $f(x) = x^5 + 2x - 1$ is continuous everywhere. We find that $f(0) = -1$ and $f(1) = 2$. By the Intermediate Value Theorem, we know that f must take the intermediate value 0 for some input $c \in (0, 1)$. This number c is a solution. □

11.C Uniqueness

Some problems ask for *more* than mere existence; they want uniqueness as well. This means that you are asked to prove two things: first that there is an element satisfying the given condition, and second that there are no other solutions. For example, we can improve the previous proposition to the following:

Proposition 11.9. *The equation $x^5 + 2x - 1 = 0$ has exactly one real solution.*

Proof. (Existence): We already proved, above, that the equation has at least one real solution.

(Uniqueness): We now show that the equation can have at most one solution. Letting $f(x) = x^5 + 2x - 1$ we compute the derivative $f'(x) = 5x^4 + 2 > 0$. Thus, by the first derivative test, the function f is strictly increasing. Hence, it can equal 0 only once. □

Here is another example.

Proposition 11.10. *For each integer $x \in \mathbb{Z}$, there exists a unique integer y, such that $x < y < x + 2$.*

Proof. Let $x \in \mathbb{Z}$.

(Existence): Fixing $y = x + 1 \in \mathbb{Z}$ works.

(Uniqueness): Since $(x + 2) - x = 2$, these integers are exactly distance 2 apart. Thus, there is at most one integer between them. □

One technique for showing uniqueness is to assume there are two solutions (not necessarily distinct), and then show that those two solutions are in fact equal. We will demonstrate this technique in the proof of the following proposition.

Proposition 11.11. *Every odd integer is uniquely the difference of two consecutive squares.*

Proof. Let $n \in \mathbb{Z}$ be odd.

(Existence): Write $n = 2k + 1$ for some $k \in \mathbb{Z}$. We have

$$(k + 1)^2 - k^2 = k^2 + 2k + 1 - k^2 = 2k + 1 = n$$

so n is a difference of consecutive squares.

(Uniqueness): Assume that $n = (x + 1)^2 - x^2$ and $n = (y + 1)^2 - y^2$ for some $x, y \in \mathbb{Z}$. Thus, we have

$$(x + 1)^2 - x^2 = (y + 1)^2 - y^2.$$

Expanding the squares we get

$$x^2 + 2x + 1 - x^2 = y^2 + 2y + 1 - y^2.$$

In other words $2x + 1 = 2y + 1$. Subtracting 1 from both sides, and cancelling the 2, we get $x = y$. Thus, n can be written as a difference of consecutive squares in at most one way. □

It is common to write

$$\exists!\, x \in S,\ R(x)$$

to denote the statement "There exists a unique $x \in S$ satisfying $R(x)$." The exclamation mark after $\exists$ denotes uniqueness.

11.D Counterexamples and disproof

Mathematicians often encounter statements for which the truth value is unknown. After working on the problem, they might discover that the statement they were trying to prove is actually *false*. Thus, they must *disprove* the statement by proving its negation. Consider the statement:

$$\forall x \in \mathbb{R},\ x > \frac{1}{2}x^2.$$

Is this statement true or false?

After a moment's reflection, we realize it is false. So we wish to give a disproof. In other words, we want to prove its negation:

$$\exists x \in \mathbb{R},\ x \leq \frac{1}{2}x^2.$$

Disproof of original statement. Fix $x = 0 \in \mathbb{R}$. We easily see that $x \leq \frac{1}{2}x^2$. □

This type of disproof is called *finding a counterexample*. We showed that the original universal statement was false by finding an example where it failed. Here is another instance of this idea.

Example 11.12. Disprove the statement: Every $x \in \mathbb{Z}$ is odd. △

Disproof. The integer $x = 2$ is a counterexample. □

The negation of an existential statement becomes a universal statement. Thus, when disproving an existential statement you must show that its negation is true everywhere in the domain. For example, the statement

$$\exists x \in \mathbb{R},\ x^2 < -1$$

is false. Here is a disproof.

Disproof. Let $x \in \mathbb{R}$ be arbitrary. We know that $x^2 \geq 0 > -1$. □

Try to decide whether the following statement is true or false, and then give a proof or disproof.

$$\text{Given } a, b \in \mathbb{Z}, \text{ if } a \mid b \text{ and } b \mid a, \text{ then } a = b.$$

Answer: This statement is false. Its negation is

$$\exists a, b \in \mathbb{Z},\ (a \mid b) \wedge (b \mid a) \wedge (a \neq b).$$

Disproof. Fix $a = 2$ and $b = -2$, which are integers. We have $a \mid b$, $b \mid a$ and $a \neq b$. □

Advice 11.13. When asked to either prove or disprove a statement, clearly tell the reader which of the two you have decided to try.

11.E Exercises

Exercise 11.1. Prove the following:

(a) There exist $a, b \in \mathbb{Q}$ such that $a^b \in \mathbb{Q}$.
(b) There exist $a, b \in \mathbb{Q}$ such that $a^b \in \mathbb{R} - \mathbb{Q}$.
(c) There exist $a, b \in \mathbb{R} - \mathbb{Q}$ such that $a^b \in \mathbb{R} - \mathbb{Q}$.
(d) There exist $a \in \mathbb{Q}$ and $b \in \mathbb{R} - \mathbb{Q}$ such that $a^b \in \mathbb{Q}$.
(e) There exist $a \in \mathbb{Q}$ and $b \in \mathbb{R} - \mathbb{Q}$ such that $a^b \in \mathbb{R} - \mathbb{Q}$.

(f) There exist $a \in \mathbb{R} - \mathbb{Q}$ and $b \in \mathbb{Q}$ such that $a^b \in \mathbb{Q}$.
(g) There exist $a \in \mathbb{R} - \mathbb{Q}$ and $b \in \mathbb{Q}$ such that $a^b \in \mathbb{R} - \mathbb{Q}$.
(Hint: For part (c), note that

$$\sqrt{2}^{\sqrt{2}} \cdot \sqrt{2}^{1-\sqrt{2}} = \sqrt{2},$$

and hence at least one of the two factors on the left-hand side is irrational. Alternatively, one can solve part (c) constructively, by proving the fact that $\log_2(3)$ is irrational.

For part (e), note that

$$2^{\sqrt{2}} \cdot 2^{\frac{1}{2}-\sqrt{2}} = \sqrt{2},$$

and hence at least one of the two factors on the left-hand side is irrational. Alternatively, we have the equality

$$\left(2^{\frac{1}{\sqrt{2}}}\right)^{\frac{1}{\sqrt{2}}} = \sqrt{2}.$$

Break into cases, according to whether $2^{\frac{1}{\sqrt{2}}}$ is rational or irrational.)

Exercise 11.2. Prove or disprove: Given $x \in \mathbb{Q}$ and $y \in \mathbb{R} - \mathbb{Q}$, then $xy \in \mathbb{R} - \mathbb{Q}$.

Exercise 11.3. Prove or disprove: Let $s \in \mathbb{Z}$. If $6s - 3$ is odd, then s is odd.

Exercise 11.4. Prove or disprove: There exists an integer x such that $x^2 + x$ is odd.

Exercise 11.5. Prove or disprove: Given any positive rational number a, there is an irrational number $x \in (0, a)$.

Exercise 11.6. Prove that for any two real numbers $x < y$, there exists a rational number in the interval (x, y). In this proof you may freely use the fact that if two real numbers are more than 1 apart, then an integer lies between them.

Idea 1: To help motivate the general proof, first consider the specific case when $y = 0.60100\ldots$ and $x = 0.59922\ldots$. They are not more than 1 apart, so you cannot find an integer between them. However, if you multiply them both by 10^3 you get

$$10^3 y = 601 \text{ and } 10^3 x = 599.22\ldots,$$

which are more than one apart. The integer 600 lives between $10^3 y$ and $10^3 x$, and so the rational number $10^{-3} \cdot 600$ belongs to the interval (x, y).

Idea 2: In the previous example, how did we know that we needed to multiply by 10^3? The decimal expansion of $y - x$ is $0.001888\ldots > 0.00100\ldots = 10^{-3}$.

Idea 3: Write the decimal expansion of $y - x$ as $d_k d_{k-1} \ldots d_1 d_0 . d_{-1} d_{-2} d_{-3} \ldots$. Since $y - x > 0$, at least one of the decimal digits is nonzero; call it d_ℓ. Prove that $y - x > 10^{\ell-1}$, and so $10^{-\ell+1} y$ is more than 1 away from $10^{-\ell+1} x$.

12 Set proofs in logic

In Section 10 we focused on performing three standard tasks involving sets:

- Prove $x \in S$.
- Prove $S \subseteq T$.
- Prove $S = T$.

In this section we work with compound statements involving sets.

12.A Implications involving set statements

Consider the statement

$$\forall S, T, \text{ if } S \subseteq T, \text{ then } S \subseteq S \cap T.$$

It looks like a standard implication. We will approach it directly by assuming $S \subseteq T$, and then using that assumption we will prove $S \subseteq S \cap T$. Now, remember that $S \subseteq S \cap T$ is *also* an implication; namely, $\forall x,\ x \in S \Rightarrow x \in S \cap T$. So we approach proving this implication directly as well! We make the second assumption $x \in S$, and using it (and the previous assumption) we prove $x \in S \cap T$. The proof will look something like the following:

Proof outline. Let S and T be sets.
Assume $S \subseteq T$.
    Assume $x \in S$.
        $\vdots$
    Conclude $x \in S \cap T$.
Conclude $S \subseteq S \cap T$. □

Now that we have mapped out the proof, here it is in its entirety.

Proposition 12.1. *Given sets S and T, if $S \subseteq T$ then $S \subseteq S \cap T$.*

Proof. Let S and T be sets. Assume $S \subseteq T$. We now show $S \subseteq S \cap T$.

Assume $x \in S$. From our assumption $S \subseteq T$, we have $x \in T$. Thus $x \in S$ and $x \in T$. Hence, from the definition of intersection, $x \in S \cap T$. As x was an arbitrary element of S, we have shown $S \subseteq S \cap T$. □

Advice 12.2. When proving compound statements with multiple implications, work step by step. For instance, suppose you want to prove a statement of the form

$$A \Rightarrow (B \Rightarrow C).$$

Working directly, you would assume A and show $B \Rightarrow C$. But to show $B \Rightarrow C$, you can again work directly by assuming B and then showing that C holds. So you have two assumptions, which can be used together to show C.

Warning 12.3. Suppose you want to prove

$$(A \Rightarrow B) \Rightarrow (C \Rightarrow D).$$

Working directly (twice) you have the assumptions $A \Rightarrow B$ and C. Your goal is to prove D. Note that you do not know yet that either A or B is true!

Proposition 12.1 can be strengthened into a biconditional statement as follows.

Proposition 12.4. *Given sets S and T, we have $S \subseteq T$ if and only if $S = S \cap T$.*

Before reading further, attempt to outline the proof of this proposition.

Proof Outline. Let S and T be sets.
($\Rightarrow$): Assume $S \subseteq T$.
    ($\subseteq$): We first show $S \subseteq S \cap T$.
        Assume $x \in S$.
        $\vdots$
        Conclude $x \in S \cap T$.
    ($\supseteq$): We now show $S \supseteq S \cap T$.
        Assume $y \in S \cap T$.
        $\vdots$
        Conclude $y \in S$.
    Conclude that $S = S \cap T$.
($\Leftarrow$): Assume $S = S \cap T$.
    ($\subseteq$): We show $S \subseteq T$.
        Assume $z \in S$.
        $\vdots$
        Conclude $z \in T$.
    Conclude that $S \subseteq T$. □

We leave the details for the full proof of Proposition 12.4 as an exercise.

12.B Do we always work directly?

Some compound statements involving sets should not be proved directly. For instance, for fixed sets A and B, consider the statement:

$$B \neq A \cup B \Rightarrow A \neq B. \tag{12.5}$$

It is usually difficult to work with unequal sets, since $A \neq B$ means

$$\exists x,\ x \in A - B \text{ or } x \in B - A.$$

So, instead we should try to prove (12.5) contrapositively. Here is a proof outline.

Proof outline. Let A and B be sets. We work contrapositively.
Assume $A = B$.
($\subseteq$): We first show $B \subseteq A \cup B$.
Assume $x \in B$.
$\vdots$
Conclude $x \in A \cup B$.
($\supseteq$): We now show $B \supseteq A \cup B$.
Assume $y \in A \cup B$.
$\vdots$
Conclude $y \in B$.
Conclude $B = A \cup B$. □

We stated above that it is *usually* difficult to work with unequal sets. However, there is one situation where this piece of advice fails. Which of the following two statements is easier to work with?

- $A = \emptyset$.
- $A \neq \emptyset$.

The second statement is easier, because it tells us that A actually has an element. To illustrate this fact, we will prove the following theorem.

Proposition 12.6. *Let A and B be sets. If $A = \emptyset$ and $B = \emptyset$, then $A \cup B = \emptyset$.*

Proof. Let A and B be sets. We work contrapositively. Assume $A \cup B \neq \emptyset$. Hence, we can fix some element $x \in A \cup B$. Thus $x \in A$ or $x \in B$. Therefore $A \neq \emptyset$ or $B \neq \emptyset$. □

Now, compare this with a direct proof!

Proof. Let A and B be sets. We work directly. Assume $A = \emptyset$ and $B = \emptyset$. We will show $A \cup B = \emptyset$.

($\subseteq$): We first show $A \cup B \subseteq \emptyset$. Assume $x \in A \cup B$. Then $x \in A$ or $x \in B$. In either case, this contradicts the fact that $A = \emptyset$ and $B = \emptyset$. Hence our assumption was false, so $A \cup B$ has no elements.

($\supseteq$): We now show the reverse inclusion $A \cup B \supseteq \emptyset$. This holds vacuously.

Thus, we have shown $A \cup B = \emptyset$, as desired. □

12.C Set proofs with Cartesian products

Let S and T be sets. What does $x \in S \times T$ mean? It means that $x = (s, t)$ for some $s \in S$ and some $t \in T$. Just as with intersections, unions, and complements, we use the *definition* of the Cartesian product to simplify a statement. We illustrate how to prove statements about products with the following theorem.

Theorem 12.7. *Given four sets A, B, C, D, if $A \subseteq C$ and $B \subseteq D$, then $A \times B \subseteq C \times D$.*

Proof. Let A, B, C, and D be sets. Assume $A \subseteq C$ and $B \subseteq D$. We will show $A \times B \subseteq C \times D$.

Assume $x \in A \times B$. Thus $x = (a, b)$ for some $a \in A$ and $b \in B$. Since $A \subseteq C$, we know $a \in C$. Also since $B \subseteq D$ we know $b \in D$. Thus $x = (a, b) \in C \times D$, as desired. □

12.D Exercises

Exercise 12.1. For sets A, B, C, prove that if $A \subseteq B \cap C$, then $A \subseteq B$. Also give an example to show that the converse fails.

Exercise 12.2. Give a complete proof for Proposition 12.4, using the sketched outline.

Exercise 12.3. Consider the statement:

Let S and T be sets. Then $S \subseteq T$ if and only if $T = S \cup T$.

Outline a proof of the statement. (Give as much detail as in the outline after Proposition 12.4. You do not need to prove the statement.)

Exercise 12.4. Let S and T be sets. Prove the following.
(a) If $S \cap T = T \cup S$, then $S = T$.
(b) If $S \times T = T \times S$ and both S and T are nonempty, then $S = T$.

Exercise 12.5. Let S and T be sets. Prove that $S = T$ if and only if $S - T = T - S$. (Hint: For the backwards direction, work contrapositively. For any sets U and V, note that if $U \neq V$, then either there is some element $x \in U$ with $x \notin V$, or vice versa.)

Exercise 12.6. Let S and T be sets. Prove or disprove: $S = T$ if and only if $S - T \subseteq T$.

Exercise 12.7. Let S be a set. Prove that $\emptyset \times S = \emptyset$.
(Hint: It suffices to show that the assumption $x \in \emptyset \times S$ leads to a contradiction.)

Exercise 12.8. For sets S and T, show that $S \times T = \emptyset$ if and only if $S = \emptyset$ or $T = \emptyset$.

Exercise 12.9. Consider the statement: Given sets A, B, C, if $A \times B \subseteq B \times C$ and $B \neq \emptyset$, then $A \subseteq C$.
Write an outline of a proof, and then (separately) give a complete proof. Is the conclusion true if we remove the hypothesis that $B \neq \emptyset$?

Exercise 12.10. Prove or disprove the converse of Theorem 12.7.

Chapter IV

Proof by Induction

Without continual growth and progress, such words as improvement, achievement, and success have no meaning. Benjamin Franklin

Mathematical induction is a proof technique that is designed to prove statements about all natural numbers. It should not be confused with inductive reasoning in the sciences, which claims that if repeated observations support a hypothesis, then the hypothesis is probably true; mathematical induction gives a definitive proof.

The basic idea of mathematical induction is to use smaller cases to prove larger ones. For instance, if one wished to prove that the open sentence

$$P(n): \ n < 2^n$$

is true for each positive integer n, one might first check that it is true when $n = 1$. In fact, it is easy to check it for many different values of n.

Suppose we could prove that whenever $P(k)$ is true for some positive integer k, then $P(k+1)$ is true. We could use this to finish the problem as follows:

Since $P(1)$ is true, $P(2)$ must be true; since $P(2)$ is true, $P(3)$ must be true; since $P(3)$ is true, $P(4)$ must be true; and so on, forever.

Induction is a technique for making clear what the phrase "and so on, forever" means in the previous sentence. Anytime we find ourselves wanting to repeat a process infinitely often in a proof, it is a sign that we should think about using induction.

13 Mathematical induction

13.A The principle of mathematical induction

An important property of the natural numbers is the principle of mathematical induction. It is a basic axiom that is used in the definition of the natural numbers, and as such it has no proof. It is as basic a fact about the natural numbers as the fact that if we add 1 to any natural number, we obtain a natural number (although its statement is more complicated).

Axiom 13.1 (The Principle of Mathematical Induction)**.** Let $P(n)$ be an open sentence, where the domain of n is $\mathbb{N}$. Suppose that
(i) $P(1)$ is true and
(ii) $\forall k \in \mathbb{N},\ P(k) \Rightarrow P(k+1)$.
Then $P(n)$ is true for all $n \in \mathbb{N}$.

A proof by mathematical induction proceeds by verifying that (i) and (ii) are true, and then concluding that $P(n)$ is true for all $n \in \mathbb{N}$. We call the verification that (i) is true the *base case* of the induction and the proof of (ii) the *inductive step*. Typically, the inductive step will involve a direct proof; in other words, we will let $k \in \mathbb{N}$, assume that $P(k)$ is true, and then prove that $P(k+1)$ follows. If we are using a direct proof we call $P(k)$ the *inductive hypothesis*.

A proof by induction thus has the following four steps.

Identify $P(n)$: Clearly identify the open sentence $P(n)$. If $P(n)$ is obvious, then this identification need not be a written part of the proof.

Base Case: Verify that $P(1)$ is true. This will typically be done by direct computation or by giving an example.

Inductive Step: Prove the implication $P(k) \Rightarrow P(k+1)$ for any $k \in \mathbb{N}$. Typically this will be done by a direct proof; assume $P(k)$ and show $P(k+1)$. (Occasionally it may be done contrapositively or by contradiction.)

Conclusion: Conclude that the theorem is true by induction. As with identifying $P(n)$, this may not need to be a written part of the proof.

Remark 13.2. An intuitive way to think of mathematical induction is as a ladder with infinitely many rungs, numbered from the bottom. Stepping on rung number n corresponds to confirming that $P(n)$ is true. Hypothesis (i) of mathematical induction says that we *can* step onto the first rung. Hypothesis (ii) of mathematical induction says that *if* we can reach rung number k, then we can reach rung number $k+1$. Together, these two hypotheses allow us to reach any rung of the ladder. ▲

The following diagram gives a visualization of the preceding remark.

We begin at rung 1, by (i).

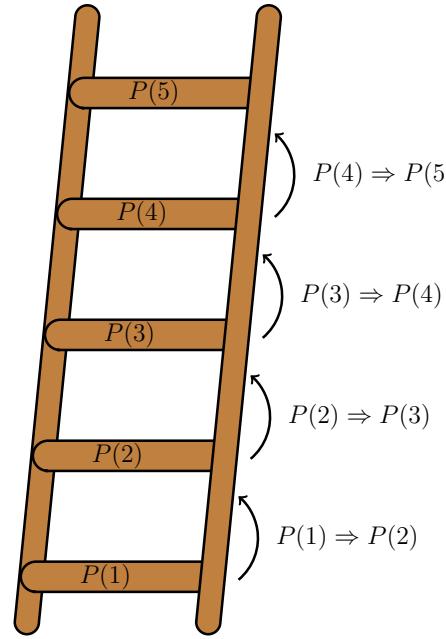
Since we can reach rung 1, we can reach rung 2, by (ii).

Since we can reach rung 2, we can reach rung 3, by (ii).

Since we can reach rung 3, we can reach rung 4, by (ii).

Since we can reach rung 4, we can reach rung 5, by (ii).

And so on... For any given rung, we see that we can reach it.



Warning 13.3. Note the importance of the base case. Without it, the inductive step shows that we can move from one rung of the ladder to the next higher one, but there is no evidence that we can reach the bottom of the ladder at all. Perhaps the ladder is suspended high above the ground; the base case shows that we can actually reach the bottom rung.

Remark 13.4. Summation notation comes up often in induction proofs. It is important to be familiar with it. The notation

$$\sum_{i=1}^{n} f(i)$$

means $f(1)+f(2)+\cdots+f(n-1)+f(n)$, where we evaluate f at each possible integer i between 1 and n, and add these values together. Hence,

$$\begin{aligned}\sum_{i=1}^{n+1} f(i) &= f(1) + \ldots + f(n) + f(n+1) \\ &= (f(1) + \ldots + f(n)) + f(n+1) \\ &= \left(\sum_{i=1}^{n} f(i)\right) + f(n+1).\end{aligned}$$

This fact, and variations of it, are often used in induction proofs involving summation.

▲

We now proceed to give an example of proof by induction in which we prove a formula for the sum of the first n natural numbers. We will first sketch the *strategy* of the proof and afterwards write the formal proof.

Proposition 13.5. *For each $n \in \mathbb{N}$,*

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

Proof Strategy. We begin by identifying the open sentence $P(n)$. In this case, $P(n)$ is the equality

$$P(n)\text{: } \sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

The base case, verifying that $P(1)$ holds, is done by a simple computation (plugging 1 in for n).

For the inductive step, we assume $P(k)$ and show $P(k+1)$. Hence, we are assuming for some $k \in \mathbb{N}$ that $P(k)$ is true, so

$$\sum_{i=1}^{k} i = \frac{k(k+1)}{2},$$

and we wish to show

$$P(k+1)\text{: } \sum_{i=1}^{k+1} i = \frac{(k+1)(k+2)}{2}.$$

Examining the statement of $P(k+1)$, we find that the left-hand side transforms as follows:

$$\begin{aligned}\sum_{i=1}^{k+1} i &= 1 + 2 + \cdots + k + (k+1) \\ &= (1 + 2 + \cdots + k) + (k+1) = \left(\sum_{i=1}^{k} i\right) + (k+1),\end{aligned}$$

where the sum on the right is the sum involved in $P(k)$. We now use our assumption of $P(k)$ to simplify the sum, and complete the proof. □

Now that we have sketched the proof method, let's write a full and formal proof.

Proof. Let $P(n)$ be the open sentence

$$P(n)\text{: } \sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

We work by induction to prove that $P(n)$ is true for each $n \in \mathbb{N}$.

Base case: $P(1)$ is true, since we have

$$\sum_{i=1}^{1} i = 1 = \frac{1(1+1)}{2}.$$

Inductive step: Let $k \in \mathbb{N}$ and assume that

$$P(k): \sum_{i=1}^{k} i = \frac{k(k+1)}{2}$$

is true. We want to show that

$$P(k+1): \sum_{i=1}^{k+1} i = \frac{(k+1)(k+2)}{2}$$

is true. Starting with the left-hand side, and simplifying with the right-hand side as a target, we find that

$$\begin{aligned}
\sum_{i=1}^{k+1} i &= 1 + 2 + 3 + \cdots + k + (k+1) \\
&= (1 + 2 + 3 + \cdots + k) + (k+1) \\
&= \left(\sum_{i=1}^{k} i\right) + (k+1) \\
&= \frac{k(k+1)}{2} + (k+1) && \text{(by the inductive hypothesis)} \\
&= \frac{k(k+1) + 2(k+1)}{2} && \text{(getting a common denominator)} \\
&= \frac{(k+1)(k+2)}{2} && \text{(factoring out } k+1\text{).}
\end{aligned}$$

So $P(k+1)$ is true.

Hence, by induction, $P(n)$ is true for all $n \in \mathbb{N}$. □

Remark 13.6. It can be helpful to point out to the reader of your proofs where you use the inductive hypothesis, as done above. Note that if you do not use the inductive hypothesis, then you could have just proved the theorem without induction. ▲

Remark 13.7. With practice you will become better at seeing how $P(k)$ and $P(k+1)$ are related (especially with sums like the one above), and these proofs will go more smoothly for you. For instance, with practice we could have gone directly to the equality

$$\sum_{i=1}^{k+1} i = \left(\sum_{i=1}^{k} i\right) + (k+1)$$

in the proof above. ▲

Warning 13.8. A common mistake that students make is to consider $P(k)$ as a number. It is a statement, not a number. For example, in the previous proof students might mistakenly write

$$P(k) = \sum_{i=1}^{k} i = \frac{k(k+1)}{2},$$

which is incorrect as it says that $P(k)$ is equal to the *number* $\frac{k(k+1)}{2}$. Another incorrect use of $P(k)$ is the following

$$\begin{aligned}\sum_{i=1}^{k+1} i &= \sum_{i=1}^{k} i + (k+1)\\ &= P(k) + (k+1)\\ &= \frac{k(k+1)}{2} + (k+1)\end{aligned}$$

Note that this also arises from thinking, incorrectly, of $P(k)$ as equal to *part* of the statement that it represents.

Remark 13.9. It might appear that in an induction proof we are assuming what we are attempting to prove. For instance, if we are trying to prove

$$\forall n \in \mathbb{N},\ P(n)$$

by induction, then in the inductive step of the proof we will need to assume $P(k)$. It would indeed be a logical mistake to assume $P(k)$ *if* our immediate goal is to prove $P(k)$.

However, that is not the case. The goal of the inductive step is not to prove $P(k)$, but to prove that $P(k+1)$ follows from $P(k)$. Hence, in fact, we are *not* assuming what we wish to prove (namely that $P(n)$ is true for each $n \in \mathbb{N}$). Note also that proving

$$\forall k \in \mathbb{N},\ P(k) \Rightarrow P(k+1)$$

by itself does not prove that $P(k)$ is true for any natural number; it just proves that *if* $P(k)$ is true for some k, then $P(k+1)$ must be true as well (which is why we also need the base case to start the induction). ▲

Here is another result we can prove by induction.

Proposition 13.10. *Given $n \in \mathbb{N}$, it happens that $2^n > n$.*

Proof. We work by induction on $n \in \mathbb{N}$.

Base case: We see that $2^1 = 2 > 1$.

Inductive step: Let $k \in \mathbb{N}$ and assume $2^k > k$. We want to prove $2^{k+1} > k+1$. We find

$$\begin{aligned} 2^{k+1} &= 2 \cdot 2^k \\ &> 2 \cdot k && \text{(by the inductive assumption)} \\ &= k + k \\ &\geq k + 1. && \text{(since } k \geq 1\text{)} \end{aligned}$$

This finishes the inductive step, so by induction we know that $2^n > n$ for each $n \in \mathbb{N}$. □

Induction can often be used to prove facts about finite sets. In this case, the general technique is to induct on the size of the sets. Typically, a proposition will be easy to prove for the empty set, or for sets with a single element. We may assume the proposition holds for sets of size k, and let A be a set of size $k+1$. Removing one element from A yields a set of size k, to which the inductive hypothesis applies. Then, we only need to extend the proposition to A; how we do it depends on what exactly we wish to prove. The following theorem is a typical example.

Proposition 13.11. *Let A be a finite nonempty set of real numbers. Then A has a least element.*

Proof. Let $P(n)$ be the open sentence

$$P(n)\text{: Every set of } n \text{ real numbers has a least element.}$$

We work by induction to show that $P(n)$ is true for each $n \in \mathbb{N}$.

Base case: It is clear that any set consisting of only 1 real number has a least element, so $P(1)$ is true.

Inductive step: Let $k \in \mathbb{N}$ and assume $P(k)$. In other words, assume that every set of k real numbers has a least element. This is our inductive hypothesis; we want to use it to prove $P(k+1)$.

Let A be any set consisting of $k+1$ real numbers. Choose one of them, and call it a. Let $B = A - \{a\}$. We note that B has k elements, so by the inductive hypothesis, B has a least element. Call this element b. By the definition of a least element, we have $b \leq x$ for each $x \in B$. Note that since $b \in B = A - \{a\}$, we must have $b \neq a$. Therefore, either $b < a$ or $b > a$.

Case 1. If $b < a$, then $b \leq x$ for each $x \in B \cup \{a\} = A$. Hence, b is the least element of A.

Case 2. If $a < b$, then $a < b \leq x$ for each $x \in B$, and $a \leq a$. Hence, $a \leq x$ for each $x \in B \cup \{a\} = A$, and so a is the least element of A.

In either case, A has a least element. Since A was an arbitrary set with $k+1$ elements, $P(k+1)$ is true. This completes the inductive step.

Hence, by induction, $P(n)$ is true for each $n \in \mathbb{N}$. Therefore, any finite set of real numbers has a least element. □

We now give an application of induction by proving a very important counting principle in mathematics; the pigeonhole principle. This principle may seem like common sense, hence all the more reason to prove it.

Theorem 13.12 (The Pigeonhole Principle)**.** *Let m and n be natural numbers, with $m > n$. If m objects are placed in n bins, then two (or more) objects must share a bin.*

Proof. We prove this by induction on $n \in \mathbb{N}$.

Let $P(n)$ be the open sentence

$P(n)$: For each $m \in \mathbb{N}$, if $m > n$ and m objects are placed in n bins, then two (or more) objects must share a bin.

Base case: We verify that $P(1)$ is true. If we have more than one object, and we place them all in one bin, then all the objects must clearly share a bin.

Inductive step: Let $k \in \mathbb{N}$ and assume $P(k)$.

We now prove $P(k+1)$. Let $m \in \mathbb{N}$. Assume $m > k+1$ and m objects are placed into $k+1$ bins. We need to show that two objects share a bin. Choose one of the objects and call it x. We divide the proof into two cases.

Case 1. Suppose that the object x shares a bin with at least one other object. In this case two objects clearly share a bin, so we are finished.

Case 2. The object x is in a bin by itself; no other object shares the bin with x. In this case there are $m-1$ remaining objects, none of which are in the same bin as object x. Hence, these $m-1$ objects must all be placed into the k remaining bins. By the inductive hypothesis (that $P(k)$ is true), we know that two of these objects must share a bin, since $m - 1 > k$.

In both cases two objects must share a bin, which completes the inductive step.

Hence, by the principle of mathematical induction, $P(n)$ is true for all $n \in \mathbb{N}$. □

Remark 13.13. The pigeonhole principle can be applied to many situations. For instance, if we choose three integers, then two of them must have the same parity. Here there are two bins; even and odd. If we choose three numbers, two of them (possibly all three) must end up in the same bin, or in other words they have the same parity.

As another example, in a class of 30 people, if each person scores between 80 and 100 percent on an exam (with no fractional scores allowed), then two people must have received the same score since there are 21 possible scores (bins) which must contain the 30 people. ▲

A variation on the pigeonhole principle occurs if we are assigning objects to bins and we have fewer objects than bins. In this case, common sense tells us that some bin will remain empty. You will be asked to prove this "common sense" statement in Exercise 13.8.

Warning 13.14. It is important to note that induction cannot be used to prove "infinite" statements. It does prove infinitely many statements. For instance, we can prove that

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

for each $n \in \mathbb{N}$. However, since $\infty \notin \mathbb{N}$, induction cannot be used to prove anything about

$$\sum_{i=1}^{\infty} i.$$

In terms of the ladder analogy, induction proves that we can reach every rung of the ladder, but it cannot be used to prove that we can reach the top of the ladder (since the ladder actually has no top).

13.B Exercises

Exercise 13.1. Prove that for each $n \in \mathbb{N}$,

$$\sum_{i=1}^{n} (2i-1) = n^2.$$

Exercise 13.2. Prove that for each $n \in \mathbb{N}$,

$$\sum_{i=1}^{n} \frac{1}{(2i-1)(2i+1)} = \frac{n}{2n+1}.$$

Exercise 13.3. Prove that for each $n \in \mathbb{N}$,

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

Exercise 13.4. (a) Prove that for each $n \in \mathbb{N}$,

$$n < 3^n.$$

(b) Prove that for each $n \in \mathbb{Z}$, $n < 3^n$. (Hint: With part (a) in hand, you might not need induction for part (b).)

Exercise 13.5. Let $x \in \mathbb{R} - \{1\}$. Prove that for each $n \in \mathbb{N}$,

$$\sum_{i=0}^{n} x^i = \frac{1 - x^{n+1}}{1 - x}.$$

Exercise 13.6. Let $x \in \mathbb{R}$ and assume $x > -1$. Prove that for each $n \in \mathbb{N}$,

$$(1+x)^n \geq 1 + nx.$$

Exercise 13.7. Let S be *any* nonempty set of natural numbers. Prove that S has a least element. (Hint: Use Proposition 13.11 and the fact that for any $n \in \mathbb{N}$, any subset of $\{1, \ldots, n\}$ is finite. You will not need to use induction in your proof, since the induction is done in the proof of Proposition 13.11.)

The fact that any nonempty subset of the natural numbers has a least element is called the *well-ordering principle*.

Exercise 13.8. Prove the following variation of the pigeonhole principle.

Let $m \in \mathbb{N} \cup \{0\}$, let $n \in \mathbb{N}$, and assume $m < n$. If we suppose m objects are placed in n bins, conclude that some bin does not contain any object.

(Hint: Use induction on n.)

14 More examples of induction

In this section we will discuss two tricks related to induction.

14.A Starting induction somewhere else

Often, we wish to prove a statement of the form

$$P(n) \text{ is true for all integers } n \geq a$$

where a is a fixed integer. Note that if $a = 1$, this is just a proof of a statement for all natural numbers.

Induction can be used to prove such statements. The only change is that our base case starts at a instead of 1. We will give a proof of this fact at the end of this section, but for now we demonstrate how this changes proofs by giving some examples.

Proposition 14.1. *For all integers $n \geq 10$ we have $2^n > n^3$.*

Before we start the proof, we make a few remarks. First, why are we restricting to integers $n \geq 10$? The reason is because the claim is *false* for some smaller integers. The inequality is false when $n = 9$. (Try it!) Second, what is the open sentence $P(n)$? It is just $P(n)$: $2^n > n^3$. When we plug in $k + 1$ for n, we have

$$P(k+1) : \ 2^{k+1} > (k+1)^3.$$

The right-hand side can be simplified a bit. We note that

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

In the computation in the proof below we will slowly try to "peel off" each of the terms k^3, $3k^2$, $3k$, and 1, one at a time, so that eventually we can end up with $(k+1)^3$.

We are now ready for the formal proof.

Proof. We wish to prove that the open sentence

$$P(n)\text{: } 2^n > n^3$$

is true for each $n \geq 10$ with $n \in \mathbb{N}$. We work by induction.

Base case: We verify that $P(10)$ is true, as follows:

$$2^{10} = 1024 > 1000 = 10^3.$$

Inductive step: Let $k \in \mathbb{N}$ with $k \geq 10$. Assume that $P(k)$ is true. So we now know that $2^k > k^3$. We wish to prove $P(k + 1)$, which states that

$$2^{k+1} > (k+1)^3.$$

In order to do this, we examine 2^{k+1} closely.

$$\begin{aligned}
2^{k+1} &= 2 \cdot 2^k \\
&= 2^k + 2^k \\
&> k^3 + k^3 && (2^k > k^3, \text{ by inductive hypothesis}) \\
&\geq k^3 + 10k^2 && (\text{since } k \geq 10) \\
&= k^3 + 3k^2 + 7k^2 && (\text{peeling off } 3k^2) \\
&\geq k^3 + 3k^2 + 70k && (\text{since } k \geq 10) \\
&= k^3 + 3k^2 + 3k + 67k && (\text{peeling off } 3k) \\
&> k^3 + 3k^2 + 3k + 1 && (\text{since } 67k > 1) \\
&= (k+1)^3.
\end{aligned}$$

Hence, $P(k+1)$ is true.

Therefore, by mathematical induction, $P(n)$ is true for each $n \geq 10$. □

Remark 14.2. When trying to prove the inductive step, it can sometimes be difficult to verify that $P(k+1)$ follows from $P(k)$. Notice that in the previous example we wanted to show that $2^{k+1} > (k+1)^3$. In order to do this, we wrote down one side of the inequality (the left-hand side) and manipulated it in order to reach the other side.

In the previous example we were aided by the knowledge that the right-hand side is

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

This gave us a target to shoot for. Moreover, the right-hand side cannot be simplified much further, which is why we started with the left-hand side in the proof above. It is often useful to manipulate both sides of an equation or inequality in order to work out (on scratch paper) how to get from one side to the other. However, in the proof we must be careful that our inequalities all go the same direction. ▲

We now introduce an important mathematical object called the factorial. Its definition is reminiscent of induction because the factorial of an integer n is defined in terms of the factorial of $(n-1)$. It is no surprise that many theorems about factorials are proved by induction.

Definition 14.3. Given $n \in \mathbb{Z}_{\geq 0}$, we define the *factorial of* n, written as $n!$ and read as "n factorial," to be

$$n! = \begin{cases} 1 & \text{if } n = 0, \\ n \cdot (n-1)! & \text{if } n > 0. \end{cases}$$

Example 14.4. If we wish to compute 5!, we use the formula repeatedly as follows:

$$\begin{aligned} 5! &= 5 \cdot 4! \\ &= 5 \cdot 4 \cdot 3! \\ &= 5 \cdot 4 \cdot 3 \cdot 2! \\ &= 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1! \\ &= 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 0! \\ &= 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 1 \\ &= 120. \end{aligned}$$

In general, for $n > 0$ we see that

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1. \qquad \triangle$$

Proposition 14.5. *For each integer $n \geq 4$ we have $n! > 2^n$.*

Proof. We wish to prove that the open sentence

$$P(n): \ n! > 2^n$$

is true for each $n \geq 4$. We work by induction.

Base case: Note that $4! = 24 > 16 = 2^4$. Hence, $P(4)$ is true.

Inductive step: Let $k \geq 4$ be an integer, and assume $P(k)$ is true. Then we know that $k! > 2^k$. Now we wish to show that $(k+1)! > 2^{k+1}$. We find

$$\begin{aligned} (k+1)! &= (k+1)k! \\ &> (k+1)2^k && \text{(since } k! > 2^k\text{, by the inductive hypothesis)} \\ &> 2 \cdot 2^k && \text{(since } k+1 > 4 > 2\text{)} \\ &= 2^{k+1}. \end{aligned}$$

Hence, $P(k+1)$ is true. Therefore, by induction $P(n)$ is true for each $n \geq 4$. □

If we wish to prove facts about finite sets, it will often be convenient to start our induction with the base case being any set of size 0 (namely, the empty set).

Proposition 14.6. *If A is a finite set, then $|\mathscr{P}(A)| = 2^{|A|}$.*

Proof. Let $P(n)$ be the open sentence

$$P(n): \text{ If } A \text{ is a set with } n \text{ elements, then } |\mathscr{P}(A)| = 2^n.$$

We will prove that $P(n)$ is true for each integer $n \geq 0$, by induction.

Base case: $P(0)$ is true; the only set with 0 elements is the empty set, and its only subset is itself, so $|\mathscr{P}(\emptyset)| = 1 = 2^0$.

Inductive step: Assume that $P(k)$ is true for some $k \geq 0$; namely, that for any set A with k elements, $|\mathscr{P}(A)| = 2^k$.

We want to prove $P(k+1)$. Let B be a set with $k+1$ elements. Choose an element of B and call it b. We divide the power set of B into two collections of subsets. Let

$$S = \{X \in \mathscr{P}(B) \,:\, b \in X\},$$

and let

$$T = \{X \in \mathscr{P}(B) \,:\, b \notin X\}.$$

We note that T consists of the subsets of $B - \{b\}$; hence, T is just the power set of $B - \{b\}$. Since $B - \{b\}$ has k elements (one element less than B), our inductive hypothesis tells us that $|T| = 2^k$.

On the other hand, each element of S is uniquely the union of an element of T with the set $\{b\}$. Hence, $|S| = |T| = 2^k$. Since S and T have no elements in common, the number of elements in $\mathscr{P}(B) = S \cup T$ is $|S| + |T| = 2^k + 2^k = 2 \cdot 2^k = 2^{k+1}$. Hence, $P(k+1)$ is true.

Therefore, by induction we see that $P(n)$ is true for each $n \geq 0$. □

We can often use induction to extend statements concerning two objects to statements concerning any finite number of objects. For instance, the following proposition is an extension of De Morgan's law, from two terms to an arbitrary (finite) number of terms.

Proposition 14.7. *For any $n \in \mathbb{N}$, if $P_1, \ldots, P_n$ are arbitrary statements, then*

$$\neg(P_1 \vee \cdots \vee P_n) \equiv (\neg P_1) \wedge \cdots \wedge (\neg P_n).$$

Proof. Let $Q(n)$ be the open sentence

For any n statements $P_1, \ldots, P_n$, we have $\neg(P_1 \vee \cdots \vee P_n) \equiv (\neg P_1) \wedge \cdots \wedge (\neg P_n)$.

We will now work by induction on $n \geq 1$.

Base case: $Q(1)$ is just the statement $\neg P_1 \equiv \neg P_1$, which is true.

Inductive step: Let $k \in \mathbb{N}$ and assume that $Q(k)$ is true; i.e.,

$$\neg(P_1 \vee \cdots \vee P_k) \equiv (\neg P_1) \wedge \cdots \wedge (\neg P_k).$$

Then we have

$$\begin{aligned}
\neg(P_1 \vee \cdots \vee P_{k+1}) &\equiv \neg((P_1 \vee \cdots \vee P_k) \vee P_{k+1}) && \text{(associativity of } \vee) \\
&\equiv \neg(P_1 \vee \cdots \vee P_k) \wedge (\neg P_{k+1}) && \text{(De Morgan's law)} \\
&\equiv (\neg P_1) \wedge \cdots \wedge (\neg P_k) \wedge (\neg P_{k+1}) && \text{(inductive hypothesis).}
\end{aligned}$$

Hence, $Q(k+1)$ is true.

Therefore, $Q(n)$ is true for all $n \in \mathbb{N}$. □

14.B Many base cases

When proving a statement by induction, sometimes the proof of the inductive step needs special cases when k is small. To avoid overcomplicating the inductive step, instead we first prove $P(n)$ for several small values of n, and then give a proof of the inductive step for values of k starting at the largest special case. We demonstrate with an example.

Theorem 14.8. *For each $n \in \mathbb{N}$, we have $2^{n+1} > n^2$.*

Proof. We take $P(n)$ to be

$$P(n): \ 2^{n+1} > n^2.$$

When $n = 1$, we have $2^{1+1} = 4 > 1 = 1^2$. When $n = 2$, we have $2^{2+1} = 8 > 4 = 2^2$. When $n = 3$, we have $2^{3+1} = 16 > 9 = 3^2$. We will now use induction to prove that $P(n)$ is true for all $n \geq 3$.

Base case: $P(3)$ has already been shown to be true.

Inductive step: Now assume $P(k)$, for some integer $k \geq 3$. Hence, we know that $2^{k+1} > k^2$. Then

$$\begin{aligned}
2^{(k+1)+1} &= 2 \cdot 2^{k+1} \\
&> 2 \cdot k^2 && \text{(by the inductive assumption)} \\
&= k^2 + k^2 \\
&\geq k^2 + 3k && \text{(since } k \geq 3\text{)} \\
&= k^2 + 2k + k && \text{(peeling off } 2k\text{)} \\
&> k^2 + 2k + 1 && \text{(since } k > 1\text{)} \\
&= (k+1)^2.
\end{aligned}$$

Hence, $P(k+1)$ is true.

Therefore, by induction $P(n)$ is true for each $n \geq 3$. Since we have already demonstrated $P(1)$ and $P(2)$, we see that $P(n)$ is true for each $n \in \mathbb{N}$. □

Advice 14.9. To decide whether or not to do extra cases, try the inductive step first (perhaps on scratch paper). If you need extra information (as we did above, to replace k^2 with $3k$) this could be a reason to do extra base cases.

Another reason to use extra cases is if you are working with a piecewise defined function. Doing small cases might help handle places where the piecewise function is different.

14.C Proof of generalized induction

Here is the promised proof that induction can start at any integer.

Theorem 14.10. *Let $a \in \mathbb{Z}$, and let $P(n)$ be an open sentence whose domain includes the set $S = \{n \in \mathbb{Z} : n \geq a\}$. If*
(i) *$P(a)$ is true and*
(ii) *$P(k) \Rightarrow P(k+1)$ for all $k \in S$,*
then $P(n)$ is true for all $n \in S$.

Proof. For $n \in \mathbb{N}$, define $P'(n) = P(n + a - 1)$. Then we have a correspondence between P and P':

$$\begin{array}{ccccccccc} P(a) & P(a+1) & P(a+2) & P(a+3) & P(a+4) & P(a+5) & P(a+6) & \cdots \\ \| & \| & \| & \| & \| & \| & \| & \\ P'(1) & P'(2) & P'(3) & P'(4) & P'(5) & P'(6) & P'(7) & \cdots \end{array}$$

This correspondence makes it clear that if we can prove $P'(n)$ for each $n \in \mathbb{N}$, then we will have proved $P(n)$ for each $n \in S$.

Now, $P'(1) = P(a)$ is true by (i).

Further, for each $k \in \mathbb{N}$, we see that $P'(k) \Rightarrow P'(k+1)$ holds since $P(k+a-1) \Rightarrow P(k+1+a-1)$, by (ii).

Hence, by the principle of mathematical induction, $P'(n)$ is true for each $n \in \mathbb{N}$, so $P(n)$ is true for each $n \in S$. □

Remark 14.11. In order to illustrate the connection between $P(n)$ and $P'(n)$, we describe the corresponding open sentences for Proposition 14.5. In that proposition, it is asserted that the open sentence

$$P(n): \ n! > 2^n$$

is true for each $n \geq 4$. Thus $a = 4$ and

$$P'(n) = P(n + a - 1) = P(n + 3): \ (n + 3)! > 2^{n+3}.$$

Proving that $n! > 2^n$ is true for each $n \geq 4$ is the same as proving that $(n+3)! > 2^{n+3}$ is true for each $n \geq 1$. ▲

14.D Exercises

Exercise 14.1. Prove that $n! > 3^n$ for each natural number $n > 6$.

Exercise 14.2. Prove that if n is any natural number greater than 5, then $n! > n^3$.

Exercise 14.3. Prove that for each $n \in \mathbb{N}$, we have $3^n \geq n^3$.

(Hint: Demonstrate this by direct calculation for $n = 1, 2, 3$. Then use induction to complete the proof for $n \geq 3$, with $n = 3$ as your base case.)

Exercise 14.4. Prove that for any $n \in \mathbb{N}$ with $n \geq 2$, if $P_1, \ldots, P_n$ are statements, then

$$\neg(P_1 \wedge \cdots \wedge P_n) \equiv (\neg P_1) \vee \cdots \vee (\neg P_n).$$

Exercise 14.5. Prove that for any $n \in \mathbb{N}$, if $x_1, \ldots, x_n \in \mathbb{R}$, then

$$\left| \sum_{i=1}^{n} x_i \right| \leq \sum_{i=1}^{n} |x_i|.$$

(Note that for $n = 2$ this is just Theorem 8.21, the triangle inequality.)

Exercise 14.6. The *Fibonacci numbers* are a collection of natural numbers labeled $F_1, F_2, F_3, \ldots$ and defined by the rule

$$F_1 = F_2 = 1,$$

and for $n > 2$,

$$F_n = F_{n-1} + F_{n-2}.$$

For instance, $F_3 = F_2 + F_1 = 2$ and $F_4 = F_3 + F_2 = 2 + 1 = 3$.

(a) Write down the first fifteen Fibonacci numbers.

(b) Prove by induction that for each $n \geq 1$,

$$\sum_{i=1}^{n} F_i = F_{n+2} - 1.$$

(c) Prove by induction that for each $n \geq 1$,

$$\sum_{i=1}^{n} F_i^2 = F_n F_{n+1}.$$

Exercise 14.7. Using the definition of the Fibonacci numbers from the previous problem, prove by induction that for any integer $n > 12$ that $F_n > n^2$.

(Hint: One possible idea for a proof is to let $P(n)$ be the open sentence

$$P(n) : \ F_n > n^2 \text{ and } F_{n-1} > (n-1)^2.$$

Use induction to prove that $P(n)$ is true for all $n \geq 14$. This then implies that $F_n > n^2$ for all $n \geq 13$.)

15 Strong induction

15.A The definition of strong induction

Sometimes, when trying to do a proof by induction, the inductive step is not feasible because $P(k)$ does not provide enough information to conclude $P(k+1)$. In this case, a variation on induction called "strong induction" is often useful.

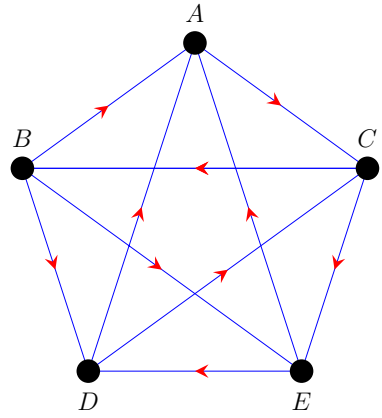
The idea of strong induction is very intuitive. Recall the ladder analogy. If we can climb to the kth rung, we just need to know that we can climb to the $k+1$st rung. However, we have more information available. If we have climbed up to the kth rung, then we have also climbed all the steps below! It is possible to make use of this extra information by making a stronger inductive hypothesis. In the inductive step, instead of merely assuming $P(k)$ we instead assume the stronger statement $Q(k) = P(1) \wedge P(2) \wedge \ldots \wedge P(k)$. In other words, we assume that we climbed each of the steps from the first to the kth.

The only difference between a proof by "normal" induction and "strong" induction is that, in the inductive step, we make the stronger assumption $Q(k)$ above. Everything else is precisely the same—we still need a base case, and in the inductive step we still want to conclude by showing $P(k+1)$.

Because these two proof techniques are so similar it is not necessary to use the word "strong" in such a proof, unless you want to emphasize this fact to the reader.

15.B Strong induction by example

Consider the following situation. There are five cities that are (conveniently) named A, B, C, D, and E. Each pair of distinct cities is connected by a single one-way road. For instance, one possible collection of such roads is given by the following diagram, where each city is represented by a dot and each road is represented by a line with an arrow indicating the direction of the road.



In this diagram, we can travel directly from A to C, but not from C to A.

A natural question one might ask is whether or not it is possible to find a path through each of the cities, without ever revisiting a city. In this diagram, one such path passing through all the cities is $B \to A \to C \to E \to D$. Others are $A \to C \to B \to E \to D$, and $B \to D \to C \to E \to A$, and $E \to D \to A \to C \to B$. This shows that for the specific combination of one-way roads given in the diagram above there are several paths.

Does the answer to our question change if we change the directions of the one-way roads? What if we change the number of cities?

To answer these new questions we need to set up some notation. Let S be a finite set of cities. We will call a collection of one-way roads, with a single road connecting each pair of distinct cities in S, a *system of one-way roads* for S. If there is some path through the cities, which follows that system of one-way roads and visits each city exactly once, we will call it a *valid path* through the cities.

We are now ready to answer our questions!

Proposition 15.1. *If S is any finite nonempty set of cities with a system of one-way roads, then there is a valid path through the cities.*

Proof. Let $P(n)$ be the open sentence: "If S is a set of n cities with a system of one-way roads, then there is a valid path through those cities." We work by (strong) induction to show that $P(n)$ is true for each $n \geq 1$.

Base Case: Let $n = 1$. In this case, starting in the single city, we don't need to go anywhere to say that we have visited all the cities. Hence, $P(1)$ is true.

Inductive Step: Assume that for some $k \in \mathbb{N}$, we know that $P(1), P(2), \ldots, P(k)$ are each true. (This is the only place where our proof would look different from a standard induction. Instead of merely assuming $P(k)$ is true, we have assumed all the steps from the first to the kth are true.) From this information, we will try to prove $P(k+1)$.

To that end, let U be a set of $k+1$ cities with a system of one-way roads connecting them. As U is nonempty, we may fix one of the cities in U, call it X. There are a total of k cities remaining that are not equal to X. We divide these k cities into two disjoint sets:

$$S = \{Y \in U - \{X\} \text{ : the road between } X \text{ and } Y \text{ runs toward } X\}$$

and

$$T = \{Y \in U - \{X\} \text{ : the road between } X \text{ and } Y \text{ runs away from } X\}.$$

Let $k_1 = |S|$ and let $k_2 = |T|$. We note that $0 \leq k_1 \leq k$ and $0 \leq k_2 \leq k$, and that U is the disjoint union of the three sets S, T, and $\{X\}$. In particular, $k_1 + k_2 = k$.

Note that the cities in S are connected by a system of one-way roads; namely, the roads in U that run between cities in S. A similar remark holds for T. We now examine three cases.

Case 1. Assume $k_1 = 0$. In this case, we have $k_2 = k$. Since $P(k)$ is true, there is a valid path for the cities in T. Starting at X, then travelling to the first city in this path, and continuing along the path, we obtain a valid path in U.

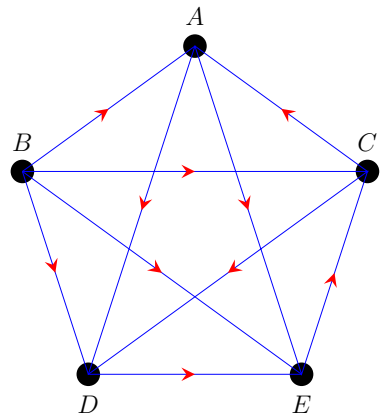
Case 2. Assume $k_2 = 0$. This case is similar to case 1, except that we put X at the end of the path instead of the beginning.

Case 3. Assume $k_1, k_2 \neq 0$. Then since $1 \leq k_1 < k$ and $1 \leq k_2 < k$, we see that both $P(k_1)$ and $P(k_2)$ are true by the (strong) inductive hypothesis. Hence, there is a valid path for S and a valid path for T. Traversing the first path we end at a city in S. From there we may travel to X. We then travel to the first city in the valid path for T. We finish by traversing this path for T. This yields a valid path for U.

By (strong) induction, $P(n)$ is true for each $n \in \mathbb{N}$. □

Note that in the first two cases of the proof above, we only needed to know that $P(k)$ is true. Thus, in those cases, normal induction would work. However, in the third case we needed to use the fact that the theorem was true for networks of cities and roads with an *arbitrary* number of cities smaller than $k+1$, not just for networks with exactly k cities.

Example 15.2. Let $U = \{A, B, C, D, E\}$ and consider the following system of one-way roads for U (which is different from the system we considered previously).



By choosing a city, say $X = C$, we divide the remaining cities, as in the previous proof, into two subsets $S = \{B, E\}$ and $T = \{A, D\}$. Note that there is a valid path for S (namely, $B \to E$), and there is a valid path for T (namely, $A \to D$). So, the valid path our proof would construct is the path $B \to E \to C \to A \to D$. In this case, our inductive step would be using $P(2)$ twice to get us to $P(5)$.

Alternatively, if we choose $X = E$, then $S = \{A, B, D\}$ and $T = \{C\}$. Between the cities of S we find the valid path $B \to A \to D$, and the valid path between the cities of T is the singleton path C. The total valid path is then $B \to A \to D \to E \to C$. For this choice of X, our inductive step uses $P(3)$ and $P(1)$ to get us to $P(5)$. △

It may be a useful exercise for the student to go back to the system of roads introduced at the beginning of Subsection 15.B and, using the proof method above, see which valid paths are constructed for each X. Also, it may be useful to note which values of $P(i)$ are being used to conclude $P(5)$, for each choice of X. Can you find a valid path which does not arise from the inductive proof?

15.C More examples of strong induction

Our next example of strong induction will be given in the proof of the following:

Proposition 15.3. *Every natural number can be written as a sum of distinct integers, each of which is a power of* 2.

Before we begin the proof, we want to make a few remarks which will help explain what we are trying to prove.

Remark 15.4 (Sums of one object)**.** When mathematicians say that a number can be written as a sum they allow the possibility of adding only one object. Hence, the number $8 = 2^3$ can be considered as a sum of a single power of two. ▲

Remark 15.5 (Meaning of distinct)**.** In English, the word "distinct" is often used to mean "special" or "distinguished." In mathematics, the word has a very precise meaning, which is quite different; a list of objects is called *distinct* if no two of the objects are equal.

For instance, there are several ways to write the number 6 as a sum of powers of two. We have

$$\begin{aligned} 6 &= 2^0 + 2^0 + 2^0 + 2^0 + 2^0 + 2^0 = 2^1 + 2^0 + 2^0 + 2^0 + 2^0 \\ &= 2^1 + 2^1 + 2^0 + 2^0 = 2^1 + 2^1 + 2^1 = 2^2 + 2^0 + 2^0 = 2^2 + 2^1. \end{aligned}$$

Note that only the last way ($6 = 4 + 2$) gives 6 as a sum of **distinct** powers of 2 (all the rest have repetition). ▲

Remark 15.6 (Base 2)**.** We illustrate the theorem for the first few natural numbers.

$$1 = 2^0, \quad 2 = 2^1, \quad 3 = 2^1{+}2^0, \quad 4 = 2^2, \quad 5 = 2^2{+}2^0, \quad 6 = 2^2{+}2^1, \quad 7 = 2^2{+}2^1{+}2^0.$$

In each case, we have written n as a sum of distinct powers of two. Mathematicians call this representing a number in *binary* (or *base* 2). Most current cultures write numbers in base 10, but other bases can be very important, such as base 16 (or hexadecimal), for computing. ▲

We are now ready to begin the proof. Try to figure out why the proof will fail if, in the inductive step, we only assume $P(k)$.

Proof. Let $P(n)$ be the open sentence

$$P(n)\text{: } n \text{ can be written as a sum of distinct integers, each a power of 2.}$$

Let $Q(n)$ be the open sentence

$$Q(n)\text{: } P(1) \wedge \cdots \wedge P(n).$$

We work by (strong) induction to show $P(n)$ is true for each $n \in \mathbb{N}$.

Base case: $P(1)$ is true, since $1 = 2^0$.

Inductive step: Let $k \in \mathbb{N}$, and assume $Q(k)$. In other words, assume that every integer from 1 to k can be written as a sum of distinct powers of 2. We wish to use this assumption to prove $P(k+1)$, i.e., that $k+1$ can be written as a sum of distinct powers of two.

We will examine two cases.

Case 1. If $k+1$ is odd, then k is even. By our inductive hypothesis, k can be written as a sum of distinct powers of 2. Since only one power of two is odd (namely $2^0 = 1$) and k is even, all of these powers of two must be even. Adding $2^0 = 1$ to the collection, we still have a collection of distinct powers of two, and they now add up to $k+1$.

Case 2. If $k+1$ is even, then $k+1 = 2a$ for some $a \in \mathbb{N}$. Now $1 \leq a \leq k$, so a can be written as a sum of distinct powers of two, by our inductive hypothesis. Increasing each of these exponents by one gives us a collection of powers of two, all distinct, that add to $2a = k+1$.

We have thus proved that $Q(k) \Rightarrow P(k+1)$. Hence, by the principle of mathematical induction $P(n)$ is true for each $n \in \mathbb{N}$. □

In order to further illustrate how the proof works, we give an example of the proof in action. In the proof, if we have an odd number, we subtract 1. If we have an even number, we divide by 2. We stop when we hit the base case. Thus we have:

$$11 \longrightarrow 10 \longrightarrow 5 \longrightarrow 4 \longrightarrow 2 \longrightarrow 1.$$

Now, writing $1 = 2^0$ we work backwards through this chain, either adding 1 or multiplying by 2 (to undo what we originally did in the chain above). Eventually we end up with a binary expansion for 11, as follows:

$$2^0 \longrightarrow 2^1 \longrightarrow 2^2 \longrightarrow 2^2 + 2^0 \longrightarrow 2^3 + 2^1 \longrightarrow 2^3 + 2^1 + 2^0.$$

Thus, we have $\boxed{11 = 2^3 + 2^1 + 2^0}$.

Try this yourself, by writing the number 14 in binary. Note that our inductive step uses $P(7)$ to help us obtain $P(14)$.

15.D Formalizing strong induction

One might wonder if, just as for induction, we need to assume strong induction as an axiom. The answer is no, because strong induction is really standard induction in disguise. In the following theorem we will prove strong induction using induction on a new sentence $Q(n)$.

Theorem 15.7. *Let $P(n)$ be an open sentence, where the domain of n is $\mathbb{N}$. If*
(i) *$P(1)$ is true and*
(ii) *for each $k \in \mathbb{N}$, $(P(1) \wedge P(2) \wedge \ldots \wedge P(k)) \Rightarrow P(k+1)$,*
then $P(n)$ is true for all $n \in \mathbb{N}$.

Proof. Assume (i) and (ii) above. Let $Q(n)$ be the open sentence

$$Q(n)\text{: } P(1) \wedge P(2) \wedge \ldots \wedge P(n).$$

Note that $Q(k+1) \equiv Q(k) \wedge P(k+1)$. Additionally, we see easily that since $Q(k) \Rightarrow P(k+1)$ for each $k \in \mathbb{N}$ by (ii), we have $Q(k) \Rightarrow Q(k) \wedge P(k+1)$, so that $Q(k) \Rightarrow Q(k+1)$. Applying the principle of mathematical induction to $Q(n)$, we see that $Q(n)$ is true for all natural numbers. This immediately implies that $P(n)$ is true for all natural numbers. □

Remark 15.8. When using strong induction, typically you will not explicitly write out what $Q(n)$ is. It is much more common when proving the inductive step to say something like "Assume $P(i)$ for all integers i in the range $1 \leq i \leq k$." Sometimes it will be convenient to say "Assume that $P(i)$ is true for all natural numbers less than or equal to k." Or even just "Assume $P(1)$, $P(2), \ldots, P(k)$ are each true." ▲

15.E Where to start?

Strong induction can start at any integer, just as induction can. The formal statement merges Theorems 14.10 and 15.7, as below. (The proof is similar to that of Theorem 14.10 and will be omitted.)

Theorem 15.9. *Let $a \in \mathbb{Z}$ and let $P(n)$ be an open sentence whose domain includes the set $S = \{n \in \mathbb{Z} : n \geq a\}$. For $n \in S$, let $Q(n)$ be the open sentence*

$$Q(n) : P(a) \wedge \cdots \wedge P(n).$$

If
(i) *$P(a)$ is true and*
(ii) *$Q(k) \Rightarrow P(k+1)$ for each $k \in S$,*
then $P(n)$ is true for each $n \in S$.

We demonstrate how such proofs work with two examples.

Proposition 15.10. *Let n be any integer greater than 5. Any square can be subdivided into n squares.*

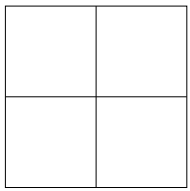
Proof. Let $P(n)$ be the open sentence

$$P(n)\text{: A square can be subdivided into } n \text{ squares.}$$

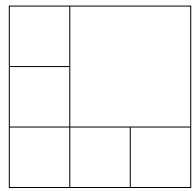
We work by induction on $n \geq 6$.

Base case: We can verify that $P(6)$ is true. We also verify that $P(7)$ and $P(8)$ are true in the diagrams below. (We do these extra base cases to help with the inductive step.)

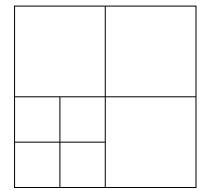
In addition, we have also given a picture showing how to subdivide a square into 4 smaller squares; we will use this in our proof.



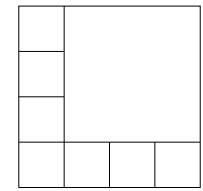
$n = 4$



$n = 6$



$n = 7$



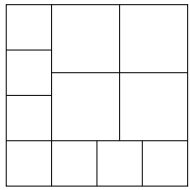
$n = 8$

Inductive step: Let $k \geq 6$, and assume that $P(\ell)$ is true for $6 \leq \ell \leq k$. We wish to prove that $P(k+1)$ is true. If $k = 6$ or $k = 7$, we have already seen that $P(k+1)$ is true, so we may assume that $k \geq 8$.

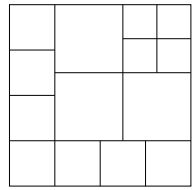
Since $k \geq 8$, we have that $k - 2 \geq 6$. Hence, by our inductive assumption, since $6 \leq k - 2 \leq k$, we know that $P(k-2)$ is true. In other words, we know that we can subdivide a square into $k - 2$ squares. Starting with this subdivision, we further subdivide the upper-rightmost square into 4 squares. This adds three squares to the subdivision. Thus, we have subdivided the square into $k - 2 + 3 = k + 1$ squares. Hence, $P(k+1)$ is true.

Therefore, by mathematical induction, $P(n)$ is true for all $n \geq 6$. □

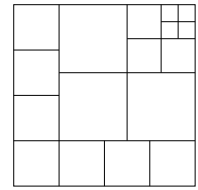
To see this proof in action, we demonstrate how to subdivide a square into 20 squares. Start from the subdivision into 8 squares, and repeatedly divide the upper-rightmost square into 4 smaller squares at each stage.



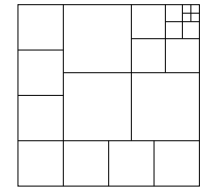
$n = 11$



$n = 14$



$n = 17$



$n = 20$

We see easily how we could extend these diagrams to demonstrate the result for $n = 23, 26, 29, 32, \ldots$ (although the upper-right square would quickly become too small to see).

To motivate our last example of a statement that can be proved by strong induction, consider the following problem.

Postage Stamp Problem: Given several denominations of postage stamps, what possible postage can be paid precisely?

Note that only a *nonnegative* number of each stamp can be used. (You can't use a negative number of stamps!) We will demonstrate this idea with a specific example. Suppose that your local post office has stamps with two denominations: 5 cents and 7 cents. What other denominations can you get using these two types of stamps?

You can pay 21 cent postage using three 7 cent stamps. You can also pay 22 cent postage, with three 5 cent stamps and one 7 cent stamp. However, 23 cent postage is not possible using those two denominations. To see this, first note that we cannot use more than three 7 cent stamps. This leaves four other options. However, none of $23 - 0(7) = 23$, $23 - 1(7) = 16$, $23 - 2(7) = 9$, or $23 - 3(7) = 2$ is a multiple of 5.

In the following proposition, we prove that every postage above 23 cents *is* possible. The main idea will be that if we can get five consecutive denominations, then by adding enough 5 cent stamps we can reach all other higher denominations.

Proposition 15.11. *Prove that every integer $n > 23$ can be written as*

$$n = 5x + 7y$$

for some integers $x, y \geq 0$.

Proof. We will let

$$P(n): \text{ For some nonnegative integers } x_n \text{ and } y_n, \text{ we have } n = 5x_n + 7y_n.$$

Here, we made the dependence of the integers x and y on n explicit, because we will later be handling multiple different values of n, at the same time.

We have the following equations:

$$\begin{aligned} 24 &= 5 \cdot 2 + 7 \cdot 2 & 25 &= 5 \cdot 5 + 7 \cdot 0 \\ 26 &= 5 \cdot 1 + 7 \cdot 3 & 27 &= 5 \cdot 4 + 7 \cdot 1 \\ 28 &= 5 \cdot 0 + 7 \cdot 4. \end{aligned}$$

The first equation shows us that $P(24)$ is true, by taking $x_{24} = 2$ and $y_{24} = 2$. Similarly, $P(25)$, $P(26)$, $P(27)$, and $P(28)$ are true. (The reason we did so many cases will become apparent in the inductive step below.) We proceed by (strong) induction on $n \geq 24$.

Base cases: We have seen that $P(24), \ldots, P(28)$ are all true.

Inductive step: Assume that for some $k \geq 28$, all of $P(24), \ldots, P(k)$ are true.

Since $k \geq 28$, we have $k - 4 \geq 24$. Hence, $P(k-4)$ is true; in other words, we can write $k - 4 = 5x_{k-4} + 7y_{k-4}$ for some integers $x_{k-4}, y_{k-4} \geq 0$. Then adding a 5 cent stamp yields

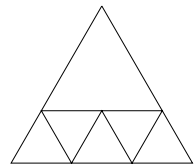
$$k + 1 = k - 4 + 5 = 5x_{k-4} + 7y_{k-4} + 5 = 5(x_{k-4} + 1) + 7y_{k-4}.$$

Taking $x_{k+1} = x_{k-4} + 1 > x_{k-4} \geq 0$ and taking $y_{k+1} = y_{k-4}$, we see that $P(k+1)$ is true.

Hence, by induction $P(n)$ is true for each $n \geq 24$. □

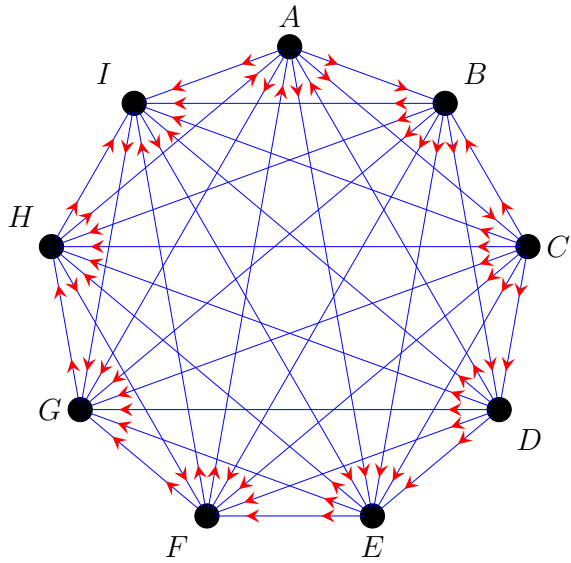
15.F Exercises

Exercise 15.1. Prove by induction that for each integer $n > 5$, it is possible to subdivide an equilateral triangle into n equilateral triangles. (For example, a subdivision into 6 equilateral triangles is given below.)



Exercise 15.2. For the network of nine cities with one-way roads below, find a route that visits all nine cities. Do this using the method found in the proof of Proposition 15.1, letting X be the city denoted by A.

(Note that there are many routes that solve the problem, but only one that arises from letting $X = A$ in the proof of Proposition 15.1.)



Exercise 15.3. (a) Prove that every integer $n > 13$ can be written as $n = 3x_n + 8y_n$ for some integers $x_n, y_n \geq 0$ (where x_n and y_n depend on n).

(b) Prove that 13 cannot be written as $3x + 8y$ for any integers $x, y \geq 0$.

Exercise 15.4. Let $n \in \mathbb{N}$. Prove (by induction) that $n = 2^{k_n} m_n$ for some nonnegative $k_n \in \mathbb{Z}$ and some odd $m_n \in \mathbb{N}$. (Again, k_n and m_n may depend on n.)

Exercise 15.5. Prove that for each natural number $n > 43$, we can write

$$n = 6x_n + 9y_n + 20z_n$$

for some nonnegative integers x_n, y_n, z_n. Then prove that 43 cannot be written in this form.

(Hint: Write 44, 45, 46, 47, 48, and 49 in the given form. Use induction to prove that any larger number can be written in the given form.)

Exercise 15.6. Find the largest postage that cannot be paid exactly with 4, 10, and 15 cent stamps. Prove that your answer is correct. (This proof will include showing not only that the postage that you find cannot be achieved, but also that every larger postage can be achieved. The correct solution is smaller than 30.)

Exercise 15.7. Recall the definition of the Fibonacci numbers from Exercise 14.6. Prove that every positive integer is a sum of one or more distinct Fibonacci numbers. (Hint: Given a positive integer $k+1$, we can find some m so that $F_m \leq k+1 < F_{m+1}$. Write $k+1 = F_m + (k+1-F_m)$, and show that $k+1-F_m$ is either 0 or (by the inductive hypothesis) a sum of distinct Fibonacci numbers, each smaller than F_m.)

16 The Binomial Theorem

16.A Binomial coefficients and Pascal's triangle

Binomial coefficients show up throughout mathematics. As we will see, they are the coefficients of x^k in the expansion of $(x+1)^n$. They also allow us to count certain collections of objects in a finite set. These binomial coefficients have many amazing properties, some of which we will prove by use of mathematical induction.

Definition 16.1. Let $n, k \in \mathbb{Z}$. We define the *binomial coefficient* as

$$\binom{n}{k} = \begin{cases} \dfrac{n!}{k!(n-k)!} & \text{if } 0 \le k \le n, \\ 0 & \text{otherwise.} \end{cases}$$

We read the symbol

$$\binom{n}{k}$$

as "n choose k."

Example 16.2. We may compute

$$\begin{aligned} \binom{7}{3} = \frac{7!}{3!(7-3)!} &= \frac{7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}{(3\cdot 2\cdot 1)(4\cdot 3\cdot 2\cdot 1)} \\ &= \frac{7\cdot 6\cdot 5}{3\cdot 2\cdot 1} = 7\cdot 5 = 35. \end{aligned}$$

In Exercise 16.1 you will prove that for any $n \ge 0$,

$$\binom{n}{0} = \binom{n}{n} = 1 \qquad \text{and} \qquad \binom{n}{1} = \binom{n}{n-1} = n. \qquad \triangle$$

Remark 16.3. It is not obvious from Definition 16.1 that the binomial coefficient is an integer, but it will follow from the properties that we describe below.

We read $\binom{n}{k}$ as "n choose k" because the binomial coefficient $\binom{n}{k}$ counts the number of ways to **choose** k objects from among n objects. See Theorem 16.6 for a proof of this fact. ▲

We now state a fundamental property of the binomial coefficients.

Theorem 16.4. *Let $n, k \in \mathbb{Z}$, with $n \ge 0$. Then*

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}.$$

Before proving Theorem 16.4, we demonstrate how to use it to build Pascal's triangle, which is a useful computational mnemonic for the binomial coefficients. See Figure 16.5. Here, the top row is row 0, and row n corresponds to the values of $\binom{n}{k}$, with $\binom{n}{0}$ being the leftmost dark entry in each row. The arrows indicate how $\binom{5}{1}$ and $\binom{5}{2}$ add to give $\binom{6}{2}$. Note that each entry (except on the top row) is the sum of the two closest entries in the row above it.

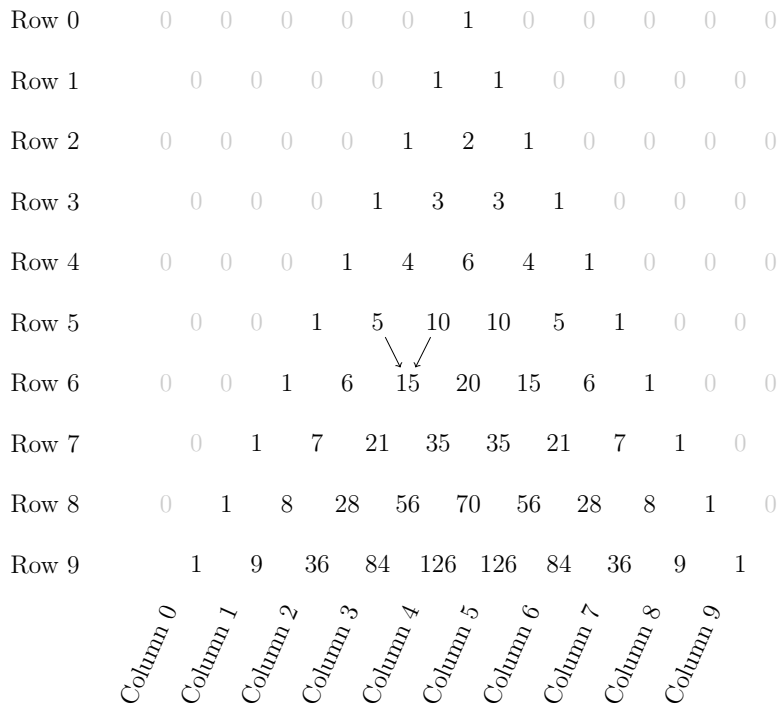


Figure 16.5: Pascal's triangle

We now prove Theorem 16.4.

Proof. We will use properties of the factorial function here. In particular, we note that $(k+1)! = (k+1) \cdot k!$, and $(n-k)! = (n-k) \cdot (n-k-1)!$ (when $0 \leq k < n$). We will use these facts to obtain a common denominator in the fractions defining the binomial coefficients $\binom{n}{k}$ and $\binom{n}{k+1}$.

We break the proof into five cases, doing the easiest cases first.

Case 1. Assume $k = n$. Then both $\binom{n}{k}$ and $\binom{n+1}{k+1}$ are 1, and $\binom{n}{k+1} = 0$.

Case 2. Assume $k > n$. Then all three binomial coefficients in the formula are 0.

Case 3. Assume $k = -1$. Then $\binom{n}{k} = 0$ but $\binom{n}{k+1} = \binom{n+1}{k+1} = 1$.

Case 4. Assume $k < -1$. Then all three binomial coefficients in the formula are 0.

Case 5. Assume $0 \leq k < n$. Then $0 < k+1 \leq n$, and all three binomial coefficients in the formula are defined by the rule involving factorials. Hence,

$$\begin{aligned}\binom{n}{k} + \binom{n}{k+1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k+1)!(n-k-1)!} \\ &= \frac{n!(k+1)}{(k+1)!(n-k)!} + \frac{n!(n-k)}{(k+1)!(n-k)!} \\ &= \frac{n!(k+1)+n!(n-k)}{(k+1)!(n-k)!} \\ &= \frac{n!(n+1)}{(k+1)!(n-k)!} \\ &= \frac{(n+1)!}{(k+1)!((n+1)-(k+1))!} \\ &= \binom{n+1}{k+1}.\end{aligned}$$

□

We can now use mathematical induction to give an interpretation of the binomial coefficients in terms of counting subsets.

Theorem 16.6. *If $n, k \in \mathbb{Z}$, then the binomial coefficient $\binom{n}{k}$ counts, for any fixed set of cardinality n, the number of subsets of cardinality k.*

Proof. If $n < 0$, there are no sets of cardinality n, so the theorem holds in that case. We will now deal with the case $n \geq 0$ by induction. Let $P(n)$ be the open sentence

$P(n)$: For each $k \in \mathbb{Z}$, the binomial coefficient $\binom{n}{k}$ counts, for any fixed set of cardinality n, the number of subsets of cardinality k.

Base case: We verify that $P(0)$ is true. For $k \neq 0$, there are no subsets of cardinality k of the empty set, matching the value $\binom{0}{k} = 0$. For $k = 0$ there is one subset of cardinality k of the empty set, matching the value $\binom{0}{0} = 1$. Hence, $P(0)$ is true.

Inductive step: Let $m \geq 0$ be an integer and assume that $P(m)$ is true. (We use m here because k already has a meaning.) Thus, for each $k \in \mathbb{Z}$, any set of cardinality m has $\binom{m}{k}$ subsets of cardinality k.

Let S be a set consisting of $m+1$ elements. Choose one of them and call it x. Let $k \in \mathbb{Z}$. If k is negative, then $\binom{m+1}{k} = 0$ is the number of k element subsets of S. Similarly, if $k = 0$, then $\binom{m+1}{k} = 1$ is the number of k element subsets of S. Therefore, in what follows, we may assume that $k > 0$.

We will count subsets T of cardinality k in S by counting the subsets with $x \in T$ separately from those with $x \notin T$ and then adding the two counts together.

Any subset T of cardinality k with $x \in T$ corresponds to the set $T - \{x\}$ which has exactly $k-1$ elements. We have $T - \{x\} \subseteq S - \{x\}$. Since $|S - \{x\}| = m$ the inductive hypothesis says that there are $\binom{m}{k-1}$ such subsets.

Any subset T of cardinality k with $x \notin T$ is a k-element subset of $S - \{x\}$. Again, the inductive hypothesis applies, and says that there are $\binom{m}{k}$ such subsets.

Adding these, the number of subsets of cardinality k in S is

$$\binom{m}{k-1} + \binom{m}{k} = \binom{m+1}{k}$$

by Theorem 16.4. This completes the inductive step.

Therefore, by mathematical induction $P(n)$ is true for each $n \geq 0$. □

Theorem 16.6 gives an easy way to prove that the binomial coefficients are all integers, a fact that is not at all obvious from the definition.

Theorem 16.7. *Let $n, k \in \mathbb{Z}$. Then the binomial coefficient $\binom{n}{k}$ is an integer.*

Proof. By Theorem 16.6, $\binom{n}{k}$ counts the number of k-element subsets in a set of size n. Hence, it must be an integer. □

16.B Proof of the Binomial Theorem

With basic facts about the binomial coefficients established, we are now ready to prove the Binomial Theorem.

Theorem 16.8 (Binomial Theorem)**.** *Let x, y be variables and let $n \geq 0$ be an integer. Then*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k.$$

Proof. We prove the theorem by induction on $n \geq 0$. Let $P(n)$ be the open sentence

$$P(n)\text{: } (x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k.$$

We note that $P(0)$ is true, since $(x+y)^0 = 1$ and

$$\sum_{k=0}^{0} \binom{0}{k} x^{0-k} y^k = \binom{0}{0} x^0 y^0 = 1,$$

Now suppose that $P(m)$ is true for some $m \geq 0$. Then we know that

$$(x+y)^m = \sum_{k=0}^{m} \binom{m}{k} x^{m-k} y^k.$$

Multiplying this first by x we obtain

$$\begin{aligned} x(x+y)^m &= \sum_{k=0}^{m} \binom{m}{k} x^{m-k+1} y^k \\ &= \binom{m}{0} x^{m+1} y^0 + \binom{m}{1} x^m y^1 + \ldots + \binom{m}{m} x y^m \end{aligned} \tag{16.9}$$

and multiplying by y we obtain

$$\begin{aligned} y(x+y)^m &= \sum_{k=0}^{m} \binom{m}{k} x^{m-k} y^{k+1} \\ &= \binom{m}{0} x^m y^1 + \ldots + \binom{m}{m-1} x y^m + \binom{m}{m} x^0 y^{m+1}. \end{aligned} \tag{16.10}$$

Adding equations (16.9) and (16.10), we obtain

$$\begin{aligned} (x+y)^{m+1} &= (x+y)(x+y)^m = x(x+y)^m + y(x+y)^m \\ &= \binom{m}{0} x^{m+1} y^0 + \binom{m}{1} x^m y^1 + \ldots + \binom{m}{m} x y^m \\ &\qquad + \binom{m}{0} x^m y^1 + \ldots + \binom{m}{m-1} x y^m + \binom{m}{m} x^0 y^{m+1} \\ &= \binom{m}{0} x^{m+1} y^0 + \left(\binom{m}{1} + \binom{m}{0} \right) x^m y + \cdots \\ &\qquad + \left(\binom{m}{m} + \binom{m}{m-1} \right) x y^m + \binom{m}{m} x^0 y^{m+1} \\ &= \binom{m+1}{0} x^{m+1} y^0 + \binom{m+1}{1} x^m y + \cdots \\ &\qquad + \binom{m+1}{m} x y^m + \binom{m+1}{m+1} x^0 y^{m+1} \\ &= \sum_{k=0}^{m+1} \binom{m+1}{k} x^{m+1-k} y^k \end{aligned}$$

This finishes the inductive step. □

16.C Exercises

Unless otherwise noted, exercises in this section should **not** be done using induction.

Exercise 16.1. Use the definition of the binomial coefficient to prove that for each integer $n \geq 0$,

$$\binom{n}{0} = \binom{n}{n} = 1 \qquad \text{and} \qquad \binom{n}{1} = \binom{n}{n-1} = n.$$

Exercise 16.2. Prove that for any $n, k \in \mathbb{Z}$,

$$\binom{n}{k} = \binom{n}{n-k}.$$

Exercise 16.3. Let $n, j, k \in \mathbb{Z}$. Using the definition of the binomial coefficient, prove that

$$\binom{n}{j}\binom{n-j}{k} = \binom{n}{k}\binom{n-k}{j}.$$

(Hint: You must deal with the three cases where $j < 0$, $k < 0$, and $j + k > n$, as well as the remaining case.)

Exercise 16.4. Prove that for any integer $n \geq 0$,

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

Exercise 16.5. Prove that for any $n \in \mathbb{N}$,

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0.$$

Exercise 16.6. Determine the coefficient of x^5y^3 in the expansion of $(2x + 3y)^8$. (*Warning*: It is not just $\binom{8}{5}$.)

Exercise 16.7. Use the definition of the binomial coefficient to prove that for any $n, k \in \mathbb{Z}$,

$$k\binom{n}{k} = n\binom{n-1}{k-1}.$$

Exercise 16.8. Do the following:

(a) Find the values of $\binom{2n}{n}$ for each $n \in \{0, 1, 2, 3, 4\}$. (Notice that these are the binomial coefficients that appear in the "middle" of Pascal's triangle.)

(b) Prove that for $n \in \mathbb{N}$, the "middle" binomial coefficient

$$\binom{2n}{n}$$

is an even integer. (Hint: Use Theorem 16.4 and Exercise 16.2.)

Exercise 16.9. Let $n, k \in \mathbb{Z}$.

(a) Use induction to prove that for $n > 8$,

$$\binom{n}{k} < 2^{n-2} \quad \text{for each } k \in \mathbb{Z}.$$

(b) Use induction to prove that for $n > 7$,

$$\binom{n}{k} < (n-3)! \quad \text{for each } k \in \mathbb{Z}.$$

Chapter V

Theory of the Integers

Mathematics is the queen of the sciences and number theory is the queen of mathematics. Carl Friedrich Gauss

One of the oldest surviving mathematical texts is Euclid's *Elements*, a collection of 13 books. This work, dating back to several hundred years BC, is one of the earliest examples of logical reasoning in mathematics still available for us to read. Although most of the books are devoted to theorems concerning geometry (many of which you may have seen in some form in a high school geometry class), books seven and nine deal with the arithmetic of the integers. In this chapter we will study some of the material found in these two books.

In particular, Euclid dealt with the topics of divisibility and greatest common divisors. It is remarkable that, thousands of years before the advent of electronic computers, Euclid wrote down a very efficient algorithm for computing GCDs that is still used essentially without change in modern computer systems.

In addition, Euclid defines prime numbers, proves that there are infinitely many primes, and proves the *Fundamental Theorem of Arithmetic*, which states that every natural number greater than 1 has a unique factorization into prime numbers.

The results that we present in this chapter have thus stood the test of time and have been studied by mathematicians over millennia. Besides being important and useful results on their own, they form a significant part of the common heritage of mathematics.

17 Divisibility

17.A Divisibility and common divisors

We now prove several facts about divisibility, some of which we took for granted in previous sections (often treating them as axioms).

Theorem 17.1. *Let a and b be nonzero integers. If $a \mid b$, then $|a| \leq |b|$.*

Proof. Assume that $a \mid b$. Then $b = ak$ for some $k \in \mathbb{Z}$. Note that $k \neq 0$, so $|k| \geq 1$. We conclude that

$$|b| = |ak| = |a||k| \geq |a|. \qquad \square$$

Corollary 17.2. *Let $a, b \in \mathbb{Z}$ be nonzero. If $a \mid b$ and $b \mid a$ then $a = \pm b$.*

Proof. Assume that $a \mid b$ and $b \mid a$. Then by Theorem 17.1, $|a| \leq |b|$ and $|b| \leq |a|$. Hence, $|a| = |b|$, so $a = \pm b$. $\square$

Theorem 17.3. *Let $b \in \mathbb{Z}$ with $b \neq 0$. There are finitely many integers that divide b.*

Proof. If $a \in \mathbb{Z}$ divides b, then $|a| \leq |b|$ by Theorem 17.1. Hence, $a \in \{-|b|, \ldots, |b|\}$. This set is finite, so there are only finitely many possibilities for a. $\square$

Example 17.4. If $b = 12$, the divisors of b are

$$\{-12, -6, -4, -3, -2, -1, 1, 2, 3, 4, 6, 12\}. \qquad \triangle$$

Definition 17.5. A *common divisor* of two integers a and b is an integer c such that $c \mid a$ and $c \mid b$.

Example 17.6. If $a = 12$ and $b = 18$, then the common divisors of a and b are ± 1, ± 2, ± 3, and ± 6. $\triangle$

Theorem 17.7. *Let a and b be integers, not both 0. The set of common divisors of a and b has a largest element.*

Proof. Without loss of generality, let $a \neq 0$. The set of divisors of a is finite and includes the set of common divisors of a and b, so the set of common divisors is finite. Since it is finite and nonempty (as 1 is an element), this set has a largest element. $\square$

Definition 17.8. The *greatest common divisor*, or GCD, of two integers a and b (not both zero) is the largest common divisor of a and b. We will write the greatest common divisor of a and b as $\text{GCD}(a, b)$.

Example 17.9. If $a = 12$ and $b = 18$, the list of divisors of a is

$$\{-12, -6, -4, -3, -2, -1, 1, 2, 3, 4, 6, 12\}$$

and the list of divisors of b is

$$\{-18, -9, -6, -3, -2, -1, 1, 2, 3, 6, 9, 18\}.$$

The set of numbers common to both of these sets (their intersection) is

$$\{-6, -3, -2, -1, 1, 2, 3, 6\}.$$

Hence, the greatest common divisor of 12 and 18 is 6. △

Example 17.10. Let $a \in \mathbb{Z}$ be nonzero. Then every divisor of a is less than or equal to $|a|$, and in fact $|a|$ is a divisor of a. In addition, $|a|$ is a divisor of 0. Hence, $\text{GCD}(a, 0) = |a|$. △

Remark 17.11. Many mathematicians write (a, b) for $\text{GCD}(a, b)$. We will avoid that notation in this book since it already has two other meanings (as an open interval and as an ordered pair).

Some take $\text{GCD}(0, 0) = 0$. We will also do so when convenient. ▲

We finish this subsection by stating a standard result about the GCD (the proof is left to the motivated reader).

Lemma 17.12. *Let $a, b \in \mathbb{Z}$. We have both*

- $\text{GCD}(a, b) = \text{GCD}(b, a)$ *and*
- $\text{GCD}(a, b) = \text{GCD}(|a|, |b|)$.

17.B The division algorithm

A fundamental property of the integers that relates addition and multiplication is the division algorithm. The fact that we can divide integers and get a unique quotient and remainder is the key to understanding divisibility, congruence, and modular arithmetic.

Theorem 17.13 (The Division Algorithm)**.** *Let $n, d \in \mathbb{Z}$ with $d \neq 0$. Then there are unique integers q, r such that*

$$n = qd + r$$

and $0 \leq r < |d|$.

In order to organize the proof of this theorem we first prove uniqueness of the quotient and remainder as a lemma before proceeding with the remainder of the proof. A portion of the proof is left for the reader in Exercise 17.8.

Lemma 17.14. *Let $n, d, q, r, q', r' \in \mathbb{Z}$ with $d \neq 0$. If $n = qd + r = q'd + r'$ with $0 \leq r, r' < |d|$, then $q = q'$ and $r = r'$.*

Proof of Lemma. Suppose that $n = qd+r = q'd+r'$ with $0 \leq r < |d|$ and $0 \leq r' < |d|$. Without loss of generality, we may assume that $r \leq r'$. Then we have

$$(q - q')d = r' - r. \tag{17.15}$$

Note that since $r' < |d|$ and $r \geq 0$ we have $0 \leq r' - r \leq r' < |d|$, so that $|r' - r| < |d|$. However, by (17.15), we know $d \,|\, (r' - r)$. Hence, by Theorem 17.1, we see that it must be the case that $r' - r = 0$, so that $r' = r$. Since $d \neq 0$, (17.15) now implies that $q = q'$. □

We will prove the existence of q and r only in the case when $n \geq 0$ and $d > 0$. The other cases of the proof (when n is negative, or when d is negative) will be left to the exercises (see Exercise 17.8).

Partial proof of Theorem 17.13. Fix $d > 0$. We work by induction to prove that

$$P(n)\text{: There are integers } q_n, r_n \text{ such that } n = q_n d + r_n \text{ and } 0 \leq r_n < d$$

is true for each $n \geq 0$.

Base Case: We note that taking $q_0 = r_0 = 0$, we have $0 = q_0 d + r_0$, and $0 \leq r_0 < d$. Hence, $P(0)$ is true.

Inductive Step: Assume that $P(k)$ is true for some $k \geq 0$; in other words, there are integers q_k and r_k such that $k = q_k d + r_k$, with $0 \leq r_k < d$. Then we have that $k + 1 = q_k d + r_k + 1$. Note that $0 \leq r_k < r_k + 1 \leq d$. We now break the proof up into cases, depending on whether $r_k + 1 = d$ or not.

Case 1: If $r_k + 1 < d$, then we see that $P(k + 1)$ is true (with $q_{k+1} = q_k$ and $r_{k+1} = r_k + 1$).

Case 2: If $r_k + 1 = d$, then $k + 1 = q_k d + d = (q_k + 1)d + 0$, so $P(k + 1)$ is true (with $q_{k+1} = q_k + 1$ and $r_{k+1} = 0$).

Hence, by induction, $P(n)$ is true for all $n \geq 0$ and $d > 0$. □

Advice 17.16. This proof of the division algorithm does not immediately give us an easy way to find the quotient and remainder. However, finding q and r is a simple task using standard long division with remainder, as taught in many elementary schools. Although we will not review long division, we demonstrate the work to compute q and r for $n = 978$ and $d = 13$.

$$\begin{array}{r} 75 \\ 13\overline{)978} \\ \underline{91} \\ 68 \\ \underline{65} \\ 3 \end{array}$$

Hence, we find that $q = 75$ and $r = 3$, so that $978 = 75 \cdot 13 + 3$.

17.C Computing the GCD

Listing all the divisors of a and b is a very inefficient way of computing the GCD. We will now give a very efficient algorithm to compute $\text{GCD}(a, b)$. It is based on the following theorem.

Theorem 17.17 (The GCD-switching Theorem)**.** *Let $a, b, c, x \in \mathbb{Z}$ and assume that $a = xb + c$. Then* $\text{GCD}(a, b) = \text{GCD}(b, c)$.

Proof. If $b = 0$, the theorem is obviously true. So assume $b \neq 0$. Let S be the set of common divisors of a and b. Let T be the set of common divisors of b and c. We will show that $S = T$. Once this is shown, the largest element of S must be the same as the largest element of T, and the theorem will be proved.

($S \subseteq T$): Assume that $d \in S$. Then $d \,|\, a$ and $d \,|\, b$. Now $c = a - xb$, so we must have $d \,|\, c$. Hence, d is a common divisor of b and c, so $d \in T$. Thus, $S \subseteq T$.

($T \subseteq S$): Now assume that $d \in T$. Then $d \,|\, b$ and $d \,|\, c$. Since $a = xb + c$, we see that $d \,|\, a$. Hence, d is a common divisor of a and b, so $d \in S$. Therefore, $T \subseteq S$.

Hence, $S = T$. □

Advice 17.18. The previous theorem does not require that $c < |b|$, so it applies in situations which can be more general than the division algorithm. The following example gives just one instance of how useful this theorem can be.

Example 17.19. Let $n \in \mathbb{Z}$. We will compute the possible GCDs for the numbers $3n + 1$ and $n - 2$. Notice that

$$3n + 1 = 3(n - 2) + 7.$$

Thus, by the GCD-switching theorem, we have $\text{GCD}(3n+1, n-2) = \text{GCD}(n-2, 7)$. The only possibilities are $1, 7$.

Can both possibilities happen? Yes, but it depends on the value of n. If $n = 1$ then $\text{GCD}(3n+1, n-2) = \text{GCD}(4, -1) = 1$. If $n = 2$ then $\text{GCD}(3n+1, n-2) = \text{GCD}(7, 0) = 7$. △

17.D The Euclidean algorithm

We now describe an algorithm that very efficiently computes $\text{GCD}(a, b)$. This algorithm will involve nothing more than repeated applications of the division algorithm; in particular, it does not involve computing divisors of a and b. After describing the algorithm, we will prove that it gives the correct answer.

Algorithm 17.20. Given two integers a, b not both 0, assume that $a \neq 0$ and that $|a| \geq |b|$ (if either of these does not hold, swap a and b so that both hold).

If $b = 0$, then the $\text{GCD}(a, b) = |a|$, and we are finished.

Otherwise, apply the division algorithm multiple times, as follows.

Divide a by b	$a = q_1 b + r_1$	with $0 \leq r_1 < \lvert b \rvert$.
Divide b by r_1	$b = q_2 r_1 + r_2$	with $0 \leq r_2 < r_1$
Divide r_1 by r_2	$r_1 = q_3 r_2 + r_3$	with $0 \leq r_3 < r_2$
⋮	⋮	⋮
Divide r_{n-1} by r_n	$r_{n-1} = q_{n+1} r_n + r_{n+1}$	with $0 \leq r_{n+1} < r_n$.

Continue to divide until we get a remainder $r_{n+1} = 0$ (we can't go any further, since we can't divide by 0).

If $r_1 = 0$, then $\text{GCD}(a, b) = |b|$ and we are finished.

If $r_{n+1} = 0$ for $n \geq 1$, then $\text{GCD}(a, b) = r_n$ and we are finished.

To show that an algorithm works correctly there are two things that need to be demonstrated. First, the answer that the algorithm computes must be correct. Second, the algorithm must terminate after finitely many steps; it does us no good if the algorithm takes forever to compute an answer. We will demonstrate that both of these facts hold true.

First, the algorithm must terminate since we have a strictly decreasing sequence of nonnegative integers $|b| > r_1 > r_2 > r_3 > r_4 > \cdots \geq 0$. This sequence can certainly not have length more than $|b| + 1$.

Now we show that the output is correct. Notice that if $b = 0$, then the algorithm completes by asserting $\text{GCD}(a, b) = |a|$. By Example 17.10, this is the correct answer.

Next, consider the case when $r_1 = 0$. The algorithm asserts that the GCD is $|b|$. By Theorem 17.17, we have

$$\text{GCD}(a, b) = \text{GCD}(b, r_1) = \text{GCD}(b, 0) = |b|.$$

Finally, choose $n \in \mathbb{N}$ so that r_{n+1} is 0. Then we have the following sequence of equalities, from Theorem 17.17.

$$\text{GCD}(a, b) = \text{GCD}(b, r_1) = \text{GCD}(r_1, r_2) = \cdots = \text{GCD}(r_n, r_{n+1}).$$

Using Example 17.10 once again, we obtain

$$\mathrm{GCD}(r_n, r_{n+1}) = \mathrm{GCD}(r_n, 0) = r_n.$$

The algorithm asserts the same answer.

For experienced computer programmers, you may recognize that the algorithm can be written recursively; we give a recursive version here.

Algorithm 17.21. Given two integers a, b, assume that $a \neq 0$. We will also assume that $|a| \geq |b|$ (if not, switch a and b). Perform the following steps.

(1) If $b = 0$, then $\mathrm{GCD}(a, b) = |a|$, and we are done.
(2) Use the division algorithm to find $a = qb + r$ with $q, r \in \mathbb{Z}$ and $0 \leq r < |b|$.
(3) At this point, we know that $\mathrm{GCD}(a, b) = \mathrm{GCD}(b, r)$. Use the algorithm to compute $\mathrm{GCD}(b, r)$.

Example 17.22. Suppose that we wish to compute the GCD of 39 and 57. We perform our divisions as follows

$$\begin{aligned} 57 &= 1 \cdot 39 + 18 \\ 39 &= 2 \cdot 18 + 3 \\ 18 &= 6 \cdot 3 + 0 \end{aligned}$$

The last nonzero remainder is 3, so the GCD of 39 and 57 is 3. △

Example 17.23. We will find $\mathrm{GCD}(1073, 1537)$.

$$\begin{aligned} 1537 &= 1 \cdot 1073 + 464 \\ 1073 &= 2 \cdot 464 + 145 \\ 464 &= 3 \cdot 145 + 29 \\ 145 &= 5 \cdot 29 + 0 \end{aligned}$$

The last nonzero remainder is 29, so the GCD of 1073 and 1537 is 29. △

Remark 17.24. Notice that the number of divisions is actually significantly less than $|b|$. In fact, it can be shown (although we will not prove it) that the number of divisions required is always less than $5 \log_{10} |b|$, which is actually slightly less than 5 times the number of digits in $|b|$. Hence, for example, if b is a four digit number, no more than 20 divisions will ever be needed. ▲

17.E Exercises

Exercise 17.1. For the given values of n and d, compute the values of q and r guaranteed by the division algorithm.

(a) Let $n = 17$, $d = 5$.
(b) Let $n = 17$, $d = -5$.
(c) Let $n = -17$, $d = 5$.
(d) Let $n = -17$, $d = -5$.
(e) Let $n = 256$, $d = 25$.
(f) Let $n = 256$, $d = -25$.
(g) Let $n = -256$, $d = 25$.
(h) Let $n = -256$, $d = -25$.

Exercise 17.2. Let a be an integer. Recall that a is even if there is some $k \in \mathbb{Z}$ such that $a = 2k$, and a is odd if there is some $\ell \in \mathbb{Z}$ such that $a = 2\ell + 1$. Prove the following statements, which we took for granted previously. (Hint: Use the division algorithm with $d = 2$.)

(a) Every integer is even or odd.
(b) No integer is both even and odd.

Exercise 17.3. Write out all the divisors of 60 in a list, and then all the divisors of 42 in a separate list. Write the common divisors in a third list, and find the GCD. (All the lists should be ordered from least to greatest.)

Exercise 17.4. Use the Euclidean algorithm to compute the following GCDs.

(a) $\text{GCD}(60, 42)$.
(b) $\text{GCD}(667, 851)$.
(c) $\text{GCD}(1855, 2345)$.
(d) $\text{GCD}(589, 437)$.

Exercise 17.5. Recall that the Fibonacci numbers are defined by the relations $F_1 = 1$, $F_2 = 1$, and for $n > 2$ the recursion $F_n = F_{n-1} + F_{n-2}$.

Prove by induction that for each $n \in \mathbb{N}$ we have $\text{GCD}(F_{n+1}, F_n) = 1$.

Exercise 17.6. Let $n \in \mathbb{Z}$. Prove that $\text{GCD}(2n + 1, 4n + 3) = 1$.

Exercise 17.7. Let $n \in \mathbb{Z}$. Prove that $\text{GCD}(6n + 2, 12n + 6) = 2$.

Exercise 17.8. Complete the proof of Theorem 17.13 as follows.

(a) Using the fact that the theorem is true for nonnegative n and positive d, prove the theorem for arbitrary n and positive d. (Hint: If $n < 0$, then $-n > 0$. Use the proven case of the division algorithm to write $-n = qd + r$. Then $n = (-q)d - r$. If $r = 0$, we are done; otherwise, we need to make an adjustment to get the remainder between 0 and d.)
(b) Using the fact that the theorem is true for positive d, prove the theorem for negative d.

18 The extended Euclidean algorithm

18.A The GCD as a linear combination

We now recall the result of Exercise 13.7.

Theorem 18.1. *Every nonempty subset of the natural numbers has a least element.*

We will use this theorem to prove an important and useful statement about $\mathrm{GCD}(a, b)$. The following definition will help us to state the result.

Definition 18.2. An (*integral*) *linear combination* of two integers a and b is a number of the form $ax + by$ where $x, y \in \mathbb{Z}$.

Example 18.3. Let $a = 16$ and $b = 21$. We will list some of the linear combinations of a and b.

We see that 37 is a linear combination of 16 and 21, since

$$37 = 16 + 21 = a + b = a \cdot 1 + b \cdot 1.$$

We see that $0 = a \cdot 0 + b \cdot 0$ is a linear combination of 16 and 21. (Will 0 be an integral linear combination of any two integers a and b?)

What is the smallest linear combination of a and b? (By smallest, we will mean in the ordering on the integers.) There isn't a smallest combination! For instance

$$a \cdot (-100) + b \cdot (-100) = -3700.$$

We can form negative numbers which can be as "small" as we like.

However, there is a smallest *positive* linear combination of a and b . We see that

$$1 = 16 \cdot 4 + 21 \cdot (-3).$$

There are no positive integers smaller than 1, so this is indeed the smallest. △

Example 18.4. When performing the division algorithm, the remainder is a linear combination of the numerator and denominator. Indeed,

$$r = n - qd = n \cdot 1 + d \cdot (-q).$$ △

Theorem 18.5. *Let a and b be integers, not both equal to 0. The smallest positive integral linear combination of a and b is* $\mathrm{GCD}(a, b)$.

Proof. Let S be the set of positive integral linear combinations of a and b. In other words,

$$S = \{ax + by \,:\, x, y \in \mathbb{Z}, ax + by > 0\}.$$

It is clear that S is a subset of the natural numbers, since its elements are positive integers. In addition, S is nonempty since it contains at least one of the following:

$$a = a \cdot 1 + b \cdot 0, \quad -a = a \cdot (-1) + b \cdot 0, \quad b = a \cdot 0 + b \cdot 1, \quad -b = a \cdot 0 + b \cdot (-1).$$

Hence S has a least element, which we call s. Fix some $x, y \in \mathbb{Z}$ so that $s = ax + by$. Note that $s > 0$.

Let $d = \text{GCD}(a, b)$. Then $d \,|\, a$ and $d \,|\, b$, so by Theorem 7.15, $d \,|\, ax + by$, and we have that $d \,|\, s$. Hence, $d \leq s$.

Now we use the division algorithm to write $a = qs + r$ with $0 \leq r < s$. Then

$$r = a - qs = a - q(ax + by) = a(1 - qx) + b(-qy)$$

is an integral linear combination of a and b. If r were positive, then r would be an element of S that is smaller than s (which would contradict the minimality of s). Hence, r must be 0. Therefore $a = qs$ and we see that $s \,|\, a$. A similar argument shows that $s \,|\, b$. Since s is a common divisor of a and b, it cannot be larger than the greatest common divisor d. Hence, $s \leq d$.

Combining the facts that $d \leq s$ and $s \leq d$, we see that $d = s$. □

Using Theorem 18.5, one can show that the linear combinations of a and b are exactly the integer multiples of $\text{GCD}(a, b)$, see Exercise 18.4.

Example 18.6. Let $a = 6$ and $b = 9$. The theorem asserts that $3 = \text{GCD}(6, 9)$ should be the smallest positive linear combination of a and b.

We see that $3 = 6(-1)+9(1)$ is indeed a linear combination. If we had $2 = 6x+9y$, then since $3 \,|\, 6$ and $3 \,|\, 9$, we would have $3 \,|\, 2$, a contradiction. Similarly, 1 cannot be a linear combination of a and b. Therefore 3 is indeed the smallest positive linear combination and is the GCD.

Note that $-15 = 3 \cdot (-5)$ is a multiple of $3 = \text{GCD}(6, 9)$. Hence, as indicated in the sentence preceding this example, we expect -15 to be a linear combination of 6 and 9. Note that

$$-15 = 3 \cdot (-5) = (6(-1) + 9(1)) \cdot (-5) = 6(5) + 9(-5)$$

is indeed a linear combination of 6 and 9. △

18.B Calculating the GCD as a linear combination

Now that we know that $\text{GCD}(a, b)$ can be written as an integral linear combination of a and b, the natural question is how to compute x and y so that

$$\text{GCD}(a, b) = ax + by.$$

We begin by performing the Euclidean algorithm for a and b, and solving each equation for the remainder.

$$\begin{aligned} a &= q_1 b + r_1 & \qquad r_1 &= a - q_1 b \\ b &= q_2 r_1 + r_2 & \qquad r_2 &= b - q_2 r_1 \\ &\vdots & &\vdots \\ r_{n-3} &= q_{n-1} r_{n-2} + r_{n-1} & \qquad r_{n-1} &= r_{n-3} - q_{n-1} r_{n-2} \\ r_{n-2} &= q_n r_{n-1} + r_n & \qquad r_n &= r_{n-2} - q_n r_{n-1} \end{aligned}$$

The bottom right equation then expresses r_n as a linear combination of the previous two remainders, r_{n-1} and r_{n-2}. We replace r_{n-1} in this equation by the integral linear combination expressed in the equation on the preceding line, so

$$\begin{aligned} r_n &= r_{n-2} - q_n r_{n-1} = r_{n-2} - q_n(r_{n-3} - q_{n-1} r_{n-2}) \\ &= (1 + q_n q_{n-1}) r_{n-2} + (-q_n) r_{n-3}. \end{aligned}$$

We now perform a similar replacement of r_{n-2} by the linear combination of r_{n-3} and r_{n-4}, found on the preceding line. Repeating this process until we use all of the equations in the right column, we have r_n written as a linear combination of a and b.

We demonstrate how this works with a couple of examples.

Example 18.7. We find $\mathrm{GCD}(493, 391)$, and write it as $493x + 391y$ for some $x, y \in \mathbb{Z}$.

We perform the Euclidean algorithm, and solve each of the resulting equations for the remainder.

$$\begin{aligned} 493 &= 1 \cdot 391 + 102 & \qquad 102 &= 493 - 1 \cdot 391 \\ 391 &= 3 \cdot 102 + 85 & \qquad 85 &= 391 - 3 \cdot 102 \\ 102 &= 1 \cdot 85 + 17 & \qquad 17 &= 102 - 1 \cdot 85 \\ 85 &= 5 \cdot 17 + 0 \end{aligned}$$

The last nonzero remainder is 17, so we know that $\mathrm{GCD}(493, 391) = 17$.

Now we see that $17 = 102 - 1 \cdot 85$, from the bottom right equation. Looking at the preceding equation, we see an expression for 85 that we plug into this equation, so

$$\begin{aligned} 17 &= 102 - 1 \cdot (\qquad 85 \qquad) \\ &= 102 - 1 \cdot (391 - 3 \cdot 102) \\ &= 102 - 1 \cdot 391 + 3 \cdot 102 \\ &= 4 \cdot 102 - 1 \cdot 391. \end{aligned}$$

Going one equation higher, we see an expression for 102; namely, $102 = 493 - 1 \cdot 391$. We plug this into our expression for 17,

$$\begin{aligned} 17 &= 4 \cdot (\quad 102 \quad) - 1 \cdot 391 \\ &= 4 \cdot (493 - 1 \cdot 391) - 1 \cdot 391 \\ &= 4 \cdot 493 - 4 \cdot 391 - 1 \cdot 391 \\ &= 4 \cdot 493 - 5 \cdot 391, \end{aligned}$$

and we now have expressed

$$\mathrm{GCD}(493, 391) = 17 = 493 \cdot 4 + 391 \cdot (-5)$$

as a linear combination of 493 and 391. △

Advice 18.8. Probably the most difficult part of this algorithm is the temptation to oversimplify the expression for the GCD. Taken to the extreme, each expression for 17 above can be simplified to equal 17. It is important to keep track of the remainders (perhaps by underlining them) and treat them as if they were variables rather than numbers.

Example 18.9. We will now find the GCD of 221 and 136, and write it as an integral linear combination of 221 and 136.

We perform the Euclidean algorithm, and solve each of the resulting equations for the remainder. In order to remind ourselves to treat the original numbers and the remainders as if they were variables, we will underline them.

$$\begin{aligned} \underline{221} &= 1 \cdot \underline{136} + \underline{85} & \qquad \underline{85} &= \underline{221} - 1 \cdot \underline{136} \\ \underline{136} &= 1 \cdot \underline{85} + \underline{51} & \qquad \underline{51} &= \underline{136} - 1 \cdot \underline{85} \\ \underline{85} &= 1 \cdot \underline{51} + \underline{34} & \qquad \underline{34} &= \underline{85} - 1 \cdot \underline{51} \\ \underline{51} &= 1 \cdot \underline{34} + \underline{17} & \qquad \underline{17} &= \underline{51} - 1 \cdot \underline{34} \\ \underline{34} &= 2 \cdot \underline{17} + \underline{0} \end{aligned}$$

The last nonzero remainder is $\underline{17}$, and we have $\underline{17} = \underline{51} - 1 \cdot \underline{34}$ (from the bottom equation on the right). The previous equation is $\underline{34} = \underline{85} - 1 \cdot \underline{51}$. Substituting for $\underline{34}$, we obtain

$$\begin{aligned} \underline{17} &= \underline{51} - 1 \cdot (\quad \underline{34} \quad) \\ &= \underline{51} - 1 \cdot (\underline{85} - 1 \cdot \underline{51}) \\ &= \underline{51} - 1 \cdot \underline{85} + 1 \cdot \underline{51} \\ &= 2 \cdot \underline{51} - 1 \cdot \underline{85}, \end{aligned}$$

where we have been careful to treat underlined numbers as variables, and not combine them with other numbers.

The equation we use to substitute for $\underline{51}$ is $\underline{51} = \underline{136} - 1 \cdot \underline{85}$.

$$\begin{aligned} \underline{17} &= 2 \cdot (\quad \underline{51} \quad) - 1 \cdot \underline{85} \\ &= 2 \cdot (\underline{136} - 1 \cdot \underline{85}) - 1 \cdot \underline{85} \\ &= 2 \cdot \underline{136} - 2 \cdot \underline{85} - 1 \cdot \underline{85} \\ &= 2 \cdot \underline{136} - 3 \cdot \underline{85} \end{aligned}$$

Finally, we have $\underline{85} = \underline{221} - 1 \cdot \underline{136}$. Substituting in for $\underline{85}$ we obtain the following:

$$\begin{aligned} \underline{17} &= 2 \cdot \underline{136} - 3 \cdot (\quad \underline{85} \quad) \\ &= 2 \cdot \underline{136} - 3 \cdot (\underline{221} - 1 \cdot \underline{136}) \\ &= 2 \cdot \underline{136} - 3 \cdot \underline{221} + 3 \cdot \underline{136} \\ &= 5 \cdot \underline{136} - 3 \cdot \underline{221}. \end{aligned}$$

Hence, $\text{GCD}(221, 136) = 17 = 136 \cdot 5 + 221 \cdot (-3)$. △

18.C Relative primality

The fact that $\text{GCD}(a, b)$ can be written as an integral linear combination of a and b has many important consequences. In particular, we saw in the proof of Theorem 18.5 that $\text{GCD}(a, b)$ is in fact the smallest positive integral linear combination of a and b. This yields the following theorem.

Theorem 18.10. *Let $a, b \in \mathbb{Z}$. Then $\text{GCD}(a, b) = 1$ if and only if $1 = ax + by$ for some $x, y \in \mathbb{Z}$.*

Proof. If $\text{GCD}(a, b) = 1$, Theorem 18.5 tells us that $1 = ax + by$ for some $x, y \in \mathbb{Z}$.

Conversely, if $1 = ax + by$ for some $x, y \in \mathbb{Z}$, then 1 is the smallest positive number that can be written as an integral linear combination of a and b (since there are no positive integers smaller than 1). Hence, $1 = \text{GCD}(a, b)$. □

We give a special name to a pair of numbers that have GCD equal to 1.

Definition 18.11. Let $a, b \in \mathbb{Z}$. If $\text{GCD}(a, b) = 1$, then we say that a and b are *relatively prime.*

Example 18.12. Since $\text{GCD}(15, 7) = 1$, the numbers 15 and 7 are relatively prime. On the other hand, $\text{GCD}(5, 30) = 5$, so 5 and 30 are not relatively prime. △

We now prove two useful properties of relatively prime integers.

Theorem 18.13. *Let $a, b, c \in \mathbb{Z}$. If $a \mid bc$ and $\text{GCD}(a, b) = 1$, then $a \mid c$.*

Proof. Let $a, b, c \in \mathbb{Z}$. Assume that $a \mid bc$ and $\text{GCD}(a, b) = 1$. Since $a \mid bc$, we see that $bc = ak$ for some $k \in \mathbb{Z}$. Also, for some $x, y \in \mathbb{Z}$, we have $1 = ax + by$. Multiplying this last equation by c, we obtain

$$c = c \cdot 1 = c(ax + by) = (ac)x + (bc)y = (ac)x + (ak)y = a(cx + ky).$$

Hence, since $cx + ky \in \mathbb{Z}$, we see that $a \mid c$. □

As a nice application of this theorem we have the following:

Example 18.14. If $2 \mid 3x$, then, since $\text{GCD}(2, 3) = 1$, we see that $2 \mid x$. (Previously we proved the implication "if $3x$ is even then x is even" contrapositively.) △

The next theorem gives us sufficient conditions under which we can expect the product of two numbers to divide into another number.

Theorem 18.15. *Let $a, b, c \in \mathbb{Z}$. If $a \mid c$ and $b \mid c$ and $\text{GCD}(a, b) = 1$, then $ab \mid c$.*

Proof. Let $a, b, c \in \mathbb{Z}$. Assume that $a \mid c$ and $b \mid c$ and $\text{GCD}(a, b) = 1$. Then for some $k, \ell, x, y \in \mathbb{Z}$, we have $c = ak$, $c = b\ell$, and $1 = ax + by$. Multiplying this last equation by c, we get

$$c = cax + cby = (b\ell)ax + (ak)by = ab(x\ell + ky).$$

Since $x\ell + ky \in \mathbb{Z}$, we see that $ab \mid c$. □

Example 18.16. If $2 \mid x$ and $3 \mid x$, we see that $6 \mid x$, since $\text{GCD}(2, 3) = 1$. △

Warning 18.17. Note that neither of the previous two theorems is true if we replace the assumption $\text{GCD}(a, b) = 1$ with $a \nmid b$. For the first theorem, taking $a = 4$, $b = 6$, and $c = 2$, we have that $4 \mid (6 \cdot 2)$, and $4 \nmid 6$, but it is not the case that $4 \mid 2$.

For the second theorem, taking $a = 12$, $b = 18$, and $c = 36$, we see that both a and b divide 36, but $ab \nmid 36$. (Can you find simpler counterexamples?)

18.D Exercises

Exercise 18.1. For each pair of numbers a and b below, calculate $\text{GCD}(a, b)$ and find $x, y \in \mathbb{Z}$ such that $\text{GCD}(a, b) = ax + by$.

(a) Take $a = 15$ and $b = 27$.
(b) Take $a = 29$ and $b = 23$.
(c) Take $a = 91$ and $b = 133$.
(d) Take $a = 221$ and $b = 377$.

Exercise 18.2. Let $a, n \in \mathbb{Z}$. Assume that $\text{GCD}(a, n) = 1$. Prove that there is some $b \in \mathbb{Z}$ such that $ab \equiv 1 \pmod{n}$.

(Hint: Use Theorem 18.10.) This result says that there is an element b which acts like the reciprocal of a, modulo n.

Exercise 18.3. Let $a, b \in \mathbb{Z}$, with $b \neq 0$, and let $d = \text{GCD}(a, b)$.
(a) Prove or disprove the equality $\text{GCD}(a, b/d) = 1$.
(b) Prove or disprove: If c is a positive common divisor of a and b, and $c = ax + by$ for some $x, y \in \mathbb{Z}$, then $c = d$. (Hint: Can you show that $c \leq d$? Can you show that $d \leq c$?)

Exercise 18.4. Let $a, b, n \in \mathbb{Z}$, and put $d = \text{GCD}(a, b)$. Prove that $d|n$ if and only if n is a linear combination of a and b.

Exercise 18.5. Let $a, b, c, d \in \mathbb{Z}$. Assume that $\text{GCD}(a, b) = 1$. Prove that if $c \mid a$ and $d \mid b$, then $\text{GCD}(c, d) = 1$.

Exercise 18.6. The following steps lead one through a proof of the existence and uniqueness of the lowest terms representation of a rational number.
(a) Let $a, b \in \mathbb{Z}$, not both zero, and let $d = \text{GCD}(a, b)$. Prove that

$$\text{GCD}\left(\frac{a}{d}, \frac{b}{d}\right) = 1.$$

(Hint: Use Theorem 18.10.)
(b) Prove that any rational number a/b (with $a, b \in \mathbb{Z}$ and $b \neq 0$) can be represented as a fraction r/s (with $r, s \in \mathbb{Z}$ and $s \neq 0$) satisfying $\text{GCD}(r, s) = 1$. (Hint: Fix $d = \text{GCD}(a, b)$. Define r and s in terms of a, b, and d, and use part (a).)
(c) Prove that in part (b) we can also guarantee that $s > 0$. (Hint: Modify the definitions of r and s if needed.)
(d) Prove that every rational number has a *unique* representation as in part (c). This is the *lowest terms* representation of the rational number. (Hint: Assume that $r/s = r'/s'$ with
(1) $r, r', s, s' \in \mathbb{Z}$,
(2) $s, s' > 0$, and
(3) $\text{GCD}(r, s) = 1 = \text{GCD}(r', s')$.
Prove that $s = s'$, and then that $r = r'$.)

Exercise 18.7. Let a, b be positive integers. A *common multiple* of a and b is an integer n such that $a \mid n$ and $b \mid n$. The *least common multiple* of a and b, written $\text{LCM}(a, b)$, is the smallest positive common multiple of a and b.
(a) Determine the LCM of 12 and 18.
(b) Determine the LCM of 21 and 35.
(c) Prove that $\text{LCM}(a, b) = \frac{ab}{d}$, where $d = \text{GCD}(a, b)$.
(Hint: Show that ab/d is a common multiple of a and b. Then show that it divides (and is thus no larger than) every other positive common multiple of a and b. You may wish to factor $a = a'd$ and $b = b'd$ and use the fact (from Exercise 18.6(a)) that $\text{GCD}(a', b') = 1$.)

19 Prime numbers

Now that we have a good understanding of divisibility in the integers we are prepared to define and study the multiplicative building blocks of the integers. As far as multiplication is concerned, prime numbers are the "atoms" from which other integers are formed.

19.A Definition of prime numbers

Definition 19.1. A *prime number* is an integer $p > 1$ such that the only positive divisors of p are 1 and p. An integer $n > 1$ that is not prime is said to be *composite*.

Example 19.2. We know that all positive divisors of a positive integer n are between 1 and n, so we may check whether a given integer is prime. The integer 2 is prime, since there are no integers between 1 and 2. The integer 3 is prime, since it is not divisible by 2. Similarly 5 is prime, since it is not divisible by 2, 3, or 4. Note that 4 is not prime, since it is divisible by 2.

The first few prime numbers are

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73. \qquad \triangle$$

If a number is composite, then it has a positive factor besides itself and 1. We expand a bit on this fact in the following theorem.

Theorem 19.3. *Let $a \in \mathbb{Z}$ with $a > 1$. If a is composite, then there are positive integers b and c, both strictly between 1 and a, such that $a = bc$.*

Proof. Since a is not prime, it has a positive divisor b with $1 < b < a$. So $a = bc$ for some $c \in \mathbb{Z}$. Clearly c is positive, and c not equal to 1 since $a \neq b$. Hence, $c > 1$.

On the other hand, by Theorem 17.1, since $c|a$ we have $c \leq a$. But $c \neq a$ since $b \neq 1$, so $1 < c < a$. □

We note the following useful fact about prime numbers.

Theorem 19.4. *Let p be a prime number and let $a \in \mathbb{Z}$. Then*

$$\mathrm{GCD}(p, a) = \begin{cases} p & \text{if } p \mid a, \\ 1 & \text{if } p \nmid a. \end{cases}$$

Proof. We know that $\mathrm{GCD}(p, a)$ must be a positive divisor of p, so it must be 1 or p. If $p \mid a$ then p is clearly the largest common divisor; similarly, if $p \nmid a$, then 1 is the largest common divisor. □

When we combine Theorem 19.4 with Theorem 18.13 we obtain the following important description of prime numbers. The implication (1) ⇒ (2) in the following theorem is known as *Euclid's Lemma*, since Euclid proved it in the *Elements*.

Theorem 19.5. *Let $a \in \mathbb{Z}$ with $a > 1$. The following are equivalent:*
(1) *a is a prime number.*
(2) *For any $b, c \in \mathbb{Z}$, if $a \mid bc$, then $a \mid b$ or $a \mid c$.*

Proof. (1) ⇒ (2): Assume that a is a prime number, and that $a \mid bc$ and $a \nmid b$. Then $\mathrm{GCD}(a, b) = 1$, so by Theorem 18.13 we see that $a \mid c$.

(2) ⇒ (1): Working contrapositively, suppose that a is not prime, so that (1) is false. Then a is composite, so $a = bc$ for some integers b, c between 1 and a. Now $a \mid bc$ and $a \nmid b$ and $a \nmid c$ (since b and c are positive and smaller than a). Hence, (2) is false. □

In Exercise 19.2 you will use induction to prove the following extension of Euclid's Lemma.

Theorem 19.6. *Let p be a prime number, let n be a natural number, and let $a_1, \ldots, a_n \in \mathbb{Z}$. If*

$$p \mid a_1 a_2 \cdots a_n$$

then $p \mid a_i$ for some $1 \leq i \leq n$.

19.B Divisibility by primes

We know that if a number is composite, then it has a factorization into smaller numbers. One might wonder if it is possible to guarantee that these smaller numbers *must* also be composite. In other words, is it possible to find a number that is so composite that all of its factors are composite? The following theorem answers that question (in the negative).

Theorem 19.7. *Every integer larger than 1 is divisible by a prime number.*

Proof. Let $a > 1$ be an integer. The set

$$S = \{b \in \mathbb{Z}_{\geq 2} : b \mid a\}$$

is finite and nonempty (since a is an element), so it has a least element p. By Exercise 19.3, p is prime. □

Next, we prove that every positive integer is a product of primes. Note that in this theorem we allow the possibility that a number (namely 1) can be a product of *zero* primes, or (if it is prime) a product of a single prime.

Theorem 19.8. *Let $a \in \mathbb{N}$. Then a is a product of primes.*

Proof. We proceed by strong induction on $n \geq 1$. Let $P(n)$ be the open sentence

$$P(n): \ n \text{ is a product of primes.}$$

We wish to prove that $P(n)$ is true for all $n \geq 1$.

Base Case: We note that $P(1)$ is true, since 1 is a product of zero primes.

Inductive Step: Suppose, for some $k \geq 1$, that $P(1), \ldots, P(k)$ are each true. In other words, assume each integer from 1 to k is a product of primes.

We now break the proof into cases depending on whether $k+1$ is prime or composite.

Case 1. Suppose $k+1$ is prime. Then $P(k+1)$ is true; $k+1$ is a product of the single prime $k+1$.

Case 2. Suppose $k+1$ is composite. Then we may factor $k+1 = bc$, with $1 < b, c < k+1$ (so $2 \leq b, c \leq k$). Hence, by our inductive hypothesis, $P(b)$ and $P(c)$ are both true; both b and c are products of primes. Hence, the product bc is a product of primes, so $P(k+1)$ is true.

Therefore, by induction the theorem is true. □

This theorem can be greatly strengthened; we will now prove that every integer not only has a prime factorization, but that its prime factorization is unique. The importance of this theorem is evident from its name: The Fundamental Theorem of Arithmetic.

Theorem 19.9 (The Fundamental Theorem of Arithmetic)**.** *If $n \in \mathbb{N}$ is greater than 1, then n has a factorization into primes*

$$n = p_1 p_2 \cdots p_r$$

(for some $r \in \mathbb{N}$) with $p_1 \leq p_2 \leq \cdots \leq p_r$, and this factorization is unique.

Proof. We have already seen that n can be written as a product of one or more primes. We order the primes so that they are in nondecreasing order. All that remains to be proved is the uniqueness statement, which we will prove by strong induction.

Let $P(n)$ be the open sentence

$$P(n): \ n \text{ has a unique factorization into prime numbers.}$$

Base Case: Clearly, $P(2)$ is true; 2 is prime, so the only way to factor it into $2 = p_1 \cdots p_r$ is to have $r = 1$ and $p_1 = 2$. A similar argument works for any prime p; hence $P(p)$ is true.

Inductive Step: Assume that $P(2), \ldots, P(k)$ are each true for some $k \geq 2$. In other words, assume each integer from 2 to k has a unique prime factorization. We wish to prove that $P(k+1)$ is true. We divide the proof into two cases.

Case 1: If $k+1$ is prime, then $P(k+1)$ is true as explained above.
Case 2: If $k+1$ is not prime, assume that

$$k+1 = p_1 p_2 \cdots p_m = q_1 q_2 \cdots q_\ell$$

has two prime factorizations, with $p_1, p_2, \ldots, p_m, q_1, q_2, \ldots, q_\ell$ all prime, and such that $p_1 \leq p_2 \leq \cdots \leq p_m$ and $q_1 \leq q_2 \leq \cdots \leq q_\ell$. Note that $p_1 \mid k+1$, so

$$p_1 \mid q_1 q_2 \cdots q_\ell.$$

Hence, by Exercise 19.2, we have that $p_1 \mid q_i$ for some i. Since q_i is prime and $p_1 \neq 1$, we must have $p_1 = q_i$. This yields $q_1 \leq q_i = p_1$.

By a similar argument $p_1 \leq q_1$. Thus, $p_1 = q_1$.

Now, $(k+1)/p_1 = p_2 \cdots p_m = q_2 \cdots q_\ell$. Since $2 \leq (k+1)/p_1 \leq k$, we see that $(k+1)/p_1$ has a unique factorization into primes, by our inductive hypothesis. Hence, $m = \ell$ and each $p_i = q_i$ for i from 2 to m. Since $p_1 = q_1$, we see that the two factorizations that we had for $k+1$ were identical. Hence, $k+1$ has a unique factorization into primes, and $P(k+1)$ is true.

Thus, by induction, every integer greater than 1 has a unique factorization into primes. □

Remark 19.10. We note that typically the factorization of a number into primes will be simplified by combining copies of the same prime together. For instance, if we wish to factor 720, rather than writing $720 = 2 \cdot 2 \cdot 2 \cdot 2 \cdot 3 \cdot 3 \cdot 5$ we might write

$$720 = 2^4 \cdot 3^2 \cdot 5.$$

This is a more compact representation of the factorization. Using this convention, we can restate Theorem 19.9 as saying that every integer $n > 1$ can be uniquely written as

$$n = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k} = \prod_{i=1}^{k} p_i^{a_i},$$

with $k \in \mathbb{N}$, each p_i prime, $p_1 < p_2 \cdots < p_k$, and with each $a_i \in \mathbb{N}$. (Note that the symbol $\prod_{i=1}^{k}$ works just like the symbol $\sum_{i=1}^{k}$, except for multiplication instead of addition.) A factorization of this form has a special name as in the next definition. ▲

Definition 19.11. Let $n > 1$ be an integer. The prime factorization

$$n = \prod_{i=1}^{k} p_i^{a_i}$$

with $k \in \mathbb{N}$, with $p_1 < \cdots < p_k$ each prime, and with each $a_i \in \mathbb{N}$, is called the *canonical factorization* of n.

Example 19.12. The canonical factorization of 5040 is $2^4 \cdot 3^2 \cdot 5^1 \cdot 7^1$. In the notation of Definition 19.11 we have $p_1 = 2$, $a_1 = 4$, $p_2 = 3$, $a_2 = 2$, $p_3 = 5$, $a_3 = 1$, $p_4 = 7$, and $a_4 = 1$. △

Remark 19.13. The statement of the Fundamental Theorem of Arithmetic is not constructive. It provides no algorithm for actually finding a prime factorization. The theorem does guarantee that whatever (correct) method we use to find a factorization will yield the same answer as any other method.

One simple method of finding a prime factorization is trial division. Given a number n, start with $k = 2$ and divide to check if n is divisible by k. If it is, add k to the list of factors of n, replace n by n/k, and repeat the process by finding the factors of n/k. (Stop if $n/k = 1$.) If n is not divisible by k, replace k by $k + 1$ and repeat the process.

We demonstrate this method by finding the prime factorization of 45. First, 45 is not divisible by 2. Hence, we check whether it is divisible by 3. It is, so we add 3 to the list of prime factors, and replace 45 by 15. Now 15 is divisible by 3, so we add 3 again to the list of prime factors, and replace 15 by $15/3 = 5$. Now 5 is not divisible by 3 or by 4, but it is divisible by 5, so we add 5 to the list of prime factors, and replace 5 by $5/5 = 1$. We are now done. The prime factorization of 45 is $3 \cdot 3 \cdot 5$.

The method just described is a very inefficient method of factoring and can in fact (with just a little thought) be improved greatly. However, factoring integers seems to be an inherently difficult problem. The search for efficient factorization techniques is an ongoing research effort, even today. ▲

19.C The infinitude of primes

Thousands of years ago, the ancient Greeks knew that there were infinitely many primes. We give here an adaptation of the proof given by Euclid of this fact. What we will prove is actually that no finite set of primes includes the set of all primes. This clearly implies that the set of all primes must be infinite.

Theorem 19.14. *There are infinitely many prime numbers.*

Proof. Let S be any finite set of prime numbers. Let

$$N = 1 + \prod_{p \in S} p.$$

Then $N \geq 2$ so N is divisible by some prime q by Theorem 19.7.

Using the division algorithm to divide N by any prime $p \in S$ leaves a remainder of 1, so no prime in S divides N. Hence, q must be a prime that is not in S. Therefore, S cannot be the set of all primes.

Since no finite set of primes consists of all the primes, there must be infinitely many primes. □

Remark 19.15. We can test this proof in specific situations by selecting any finite set of primes that we wish to consider, and constructing a prime not in that set. For instance, let $S = \{2, 3, 5, 7, 11\}$. Then $N = 2311$. In this case, N is prime and $N \notin S$.

Now suppose that $S = \{2, 3, 5, 7, 11, 13, 17, 19\}$. Then $N = 9699691 = 347 \cdot 27953$. Both 347 and 27953 are primes not in S. ▲

19.D Exercises

Exercise 19.1. For each of the following integers n, give its canonical prime factorization.

(a) $n = 27$. (b) $n = 3072$. (c) $n = 60$.

Exercise 19.2. Let p be a prime number and let $n \in \mathbb{N}$. Let $P(n)$ be the open sentence

if p divides any product of n integers, then p divides one of those integers.

Prove by induction that $P(n)$ is true for each $n \in \mathbb{N}$. (Hint: The case $n = 2$ is Theorem 19.5. For the inductive step, when showing $P(k+1)$ you should assume its premise, namely that $p|a_1 \cdots a_k a_{k+1}$, for some $a_1, \ldots, a_{k+1} \in \mathbb{Z}$.)

Exercise 19.3. Let $n > 1$ be a natural number. Prove that the smallest divisor d of n that is greater than 1 is prime.

Exercise 19.4. The goal of this exercise is to prove that there are infinitely many primes which are congruent to -1 modulo 3. We will do this in a series of steps.

(a) Prove that, with only one exception, every prime number is congruent to either 1 or -1 modulo 3.

(b) Prove that for any $n \in \mathbb{N}$ and any $a_1, \ldots, a_n \in \mathbb{Z}$, if $a_i \equiv 1 \pmod 3$ for each $1 \leq i \leq n$, then the product $a_1 a_2 \cdots a_n$ is also congruent to 1 modulo 3. (Use induction.)

(c) Suppose that $N \in \mathbb{N}$, and $N \equiv -1 \pmod 3$. Prove that N is divisible by some prime p such that $p \equiv -1 \pmod 3$. (Hint: Working by way of contradiction, assume that no prime factor of N is congruent to -1 modulo 3. What happens if all the prime factors are congruent to 1 modulo 3? What happens if one of the prime factors is the exceptional prime from part (a)?)

(d) Prove that there are infinitely many primes p that are congruent to -1 modulo 3. (Hint: Let $\{p_1, \ldots, p_n\}$ be any finite set of primes that are congruent to -1 modulo 3. Mimic the proof of Theorem 19.14, using $-1 + 3(p_1 p_2 \cdots p_n)$ in place of N.)

Exercise 19.5. Prove that there are infinitely many primes p such that

$$p \equiv -1 \pmod 4.$$

(Hint: Do steps (a) through (d) of the previous exercise with 3 replaced by 4 everywhere.)

Chapter VI

Relations

Assumptions are the termites of relationships. Henry Winkler

Studying *relationships* between objects can yield important information about the objects themselves. For the real numbers, we study inequalities; for sets, we study inclusion; for the integers, we study divisibility and congruences. All of these relationships, whether between real numbers, sets, or integers, will be seen to be special cases of the concept of a *relation* introduced in this chapter.

Because it encompasses so many different examples, the concept of a relation has very broad applicability, but this level of generality limits what we can prove. Any theorems proved about relations must be true for a wide variety of relationships. For this reason, mathematicians have singled out several types of relations for special study. Among these special relations are equivalence relations, which we will study later in this chapter, and functions, which we will study in the next chapter.

By studying special types of relations, the theorems that we can prove will be much more interesting. The relations that we study have been chosen to be widely applicable, while simultaneously providing us with a rich collection of theorems.

20 Properties of relations

20.A What is a relation?

There are many relations that occur in mathematics. For instance, $<$ on the real numbers is a relation. Inclusion, $\subseteq$, is a relation on sets. Divisibility on the natural numbers is a relation. The following definition of a relation encompasses all of these examples and others.

Definition 20.1. Let A and B be sets. A *relation from A to B* is a subset $R \subseteq A \times B$. For arbitrary elements $a \in A$ and $b \in B$ we will write aRb to mean $(a, b) \in R$.

In the case when $A = B$, we say R is a *relation on A*.

This definition is very general; in fact, it is so general that many examples of relations may have no mathematical significance. Nevertheless, some relations are very important. We give a number of examples.

Example 20.2. Let $A = \{1, 2, 3\}$ and $B = \{x, y, z\}$. Set $R = \{(1, y), (3, x)\}$. Then we have $1Ry$ and $3Rx$, but not $1Rz$ (so we write $1 \not R z$). In fact, besides 1 relating to y and 3 relating to x, no element of A relates to any element of B (under the relation R given in this example). △

Example 20.3. Let $A = B = \mathbb{N}$, and let $R = \{(a, b) \in \mathbb{N} \times \mathbb{N} : a - b = 2\}$. Then $3R1$ and $5R3$, but $1 \not R 3$ and $5 \not R 1$. △

Example 20.4. We can define relations on sets of words. Indeed, let W be the set of all words in the English language. Let

$$R = \{(\alpha, \beta) \in W \times W : \alpha \text{ and } \beta \text{ have the same length}\}.$$

Under this relation, two words are related to each other exactly when they have the same length. Thus "tree" is related to "yaks," but "awesome" is not related to "gum." We might name this relation the "have the same length" relation. △

Sometimes we can define relations using symbols other than R. For instance:

Example 20.5. Let $A = \{1, 2, 3, 4, 5\}$, and let $B = \mathscr{P}(A)$. We define a relation from A to B using the following set of ordered pairs:

$$R = \{(a, X) \in A \times B : a \in X\}.$$

Since X is a set, it makes sense to ask whether a is an element of X.

In this case, if we take $a = 2$ and $X = \{1, 2, 3\}$ we see that aRX. However, with $a = 2$ and $X = \{1, 4, 5\}$, we have $a \not R X$.

Given $a \in A$ and $X \in B$, we see that aRX if and only if $a \in X$. Thus, we could have used the symbol "$\in$" instead of "R." △

The following example exhibits a construction of a relation in a way that will be important to us in later sections.

Example 20.6. Let $A = \{1, 2, 3, 4, 5\}$. We break A up into pieces by letting $S = \{\{1, 2\}, \{3, 4\}, \{5\}\}$. (There were many different ways we could have broken A into pieces; this is just one of them.) We now define a relation R on A by

$$R = \{(a, b) \in A \times A : \text{for some } X \in S, \text{ both } a \in X \text{ and } b \in X\}.$$

Let's test whether or not 1 is related to 2. Take $a = 1$ and $b = 2$. Note that a and b are both in $\{1, 2\} \in S$. Hence, there is some X (namely $X = \{1, 2\}$) that is an element of S such that $a \in X$ and $b \in X$, so we see that $1R2$.

If we take $a = 1$ and $b = 4$, there is no element $X \in S$ such that $a \in X$ and $b \in X$, so $(1, 4) \notin R$. If we work through all the possible elements of $A \times A$, we find that

$$R = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 3), (3, 4), (4, 3), (4, 4), (5, 5)\}. \qquad \triangle$$

There are many different ways to define a relation. One option is to simply list all the possible ordered pairs in R. Another option we have seen is to write R using set-builder notation. For instance, we defined the "have the same length" relation this way. Sometimes set-builder notation is too clunky, and so we express the definition of the relation in words. The following example shows how this is commonly done.

Example 20.7. Recall the relation defined in Example 20.3,

$$R = \{(a, b) \in \mathbb{N} \times \mathbb{N} : a - b = 2\}.$$

To define the same relation, we could have instead said the following:

Let R be a relation on $\mathbb{N}$, defined by aRb if $a - b = 2$.

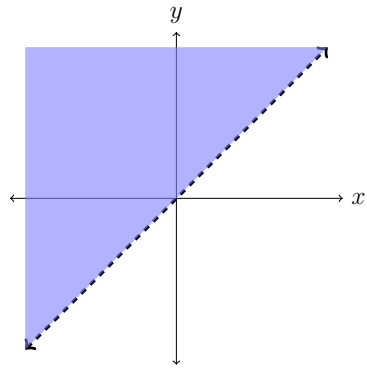
This is shorthand for the more complete sentence: "Let R be the relation on the natural numbers, $\mathbb{N}$, defined as the set of ordered pairs $(a, b) \in \mathbb{N} \times \mathbb{N}$ satisfying $a - b = 2$." When we define a relation R by a condition (such as $a - b = 2$) that condition tells us exactly when we should expect a to relate to b. $\triangle$

Warning 20.8. As with any definition, when we define a relation R with the word "if," the proper interpretation is "if and only if." For instance, in the previous example, the definition really means "aRb if and only if $a - b = 2$." This is a standard convention of mathematical language that can take some getting used to.

Example 20.9. Let $A = \mathbb{R}$, and define a relation on A by xRy if $x - y$ is negative. If we wish to express R in set-builder notation, we could write

$$R = \{(x, y) \in \mathbb{R} \times \mathbb{R} : x - y \text{ is negative}\}.$$

Because $\mathbb{R} \times \mathbb{R}$ is just $\mathbb{R}^2$ we can graph the set R in the coordinate plane, as follows.



Notice that xRy means the same thing as $x < y$. In fact, this is the mathematical relation "less than," and it is usually just denoted by the symbol "$<$" rather than being called R. △

Often, for commonly used relations, we will find that there is a standard symbol. If there is not, we might use a nonstandard symbol, in which case we must be careful to define what the symbol means.

Some other standard symbols for relations on the real numbers are

$$\leq, \quad \geq, \quad >, \quad =, \quad \neq .$$

In Exercise 20.2 you will be asked to graph the set $R \subseteq \mathbb{R} \times \mathbb{R}$ corresponding to each of these relations.

We end with one more example of a relation that is denoted with a standard symbol.

Example 20.10. Let A be the set of all compound sentences formed from P and Q. Define a relation R on A by xRy if x is logically equivalent to y. This relation R is usually written $\equiv$. △

20.B Properties of relations on a set A

We now study different properties that relations on a set A can have. These properties (especially the first three) have proven to be very useful; relations that satisfy these properties tend to be more mathematically interesting than other relations.

Definition 20.11. Let R be a relation on a set A.
(1) We say that R is *reflexive* if: $\forall a \in A,\ aRa$.
(2) We say that R is *symmetric* if: $\forall a, b \in A,\ aRb \Rightarrow bRa$.
(3) We say that R is *transitive* if: $\forall a, b, c \in A,\ ((aRb) \wedge (bRc)) \Rightarrow (aRc)$.
(4) We say that R is *antisymmetric* if: $\forall a, b \in A,\ ((aRb) \wedge (bRa)) \Rightarrow a = b$.

Example 20.12. Consider the "equality" relation on $\mathbb{C}$. Which of the four properties above hold for this relation?

It is reflexive, since given any $a \in \mathbb{C}$ we know $a = a$. It is also symmetric, since given $a, b \in \mathbb{C}$ if $a = b$ then $b = a$. It is transitive; for given $a, b, c \in \mathbb{C}$ if we assume $a = b$ and $b = c$, then $a = c$. Finally, it is antisymmetric. (Can you fill in the proof? Let $a, b \in A$. Assume $a = b$ and $b = a$. Conclude $a = b$.) △

Example 20.13. Consider the relation $<$ on $\mathbb{R}$. We will show that $<$ is not reflexive, is not symmetric, is transitive, and is antisymmetric.

(Not reflexive): Fix $0 \in \mathbb{R}$. We have $0 \not< 0$.

(Not symmetric): Fix $0, 1 \in \mathbb{R}$. We have $0 < 1$ but $1 \not< 0$.

(Transitive): Let $a, b, c \in \mathbb{R}$. Assume $a < b$ and $b < c$. It follows that $a < c$.

(Antisymmetric): Let $a, b \in \mathbb{R}$. Assume $a < b$ and $b < a$. This is impossible, so the implication is vacuously true. △

Example 20.14. Let $A = \{1, 2, 3, 4\}$. Define a relation R on A by

$$R = \{(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3), (4,4)\}.$$

We will prove that R is reflexive, symmetric, and transitive, but not antisymmetric. We can do this by just checking every possibility.

(Reflexive): To see that R is reflexive, we note that each of the four elements of A relates to itself; this is because $(1, 1)$, $(2, 2)$, $(3, 3)$, and $(4, 4)$ are in R.

(Symmetric): To see that R is symmetric, we note that for any pair in R, reversing the order of the elements in the pair yields another pair in R; for example, $(1, 2) \in R$, and reversing the elements, $(2, 1)$ is also in R. (What if $a = 1$ and $b = 4$? In that case the implication is vacuously true, since the premise is false.)

(Transitive): Transitivity is harder to see. Technically, you have 4 options for each of a, b, and c, giving a total of 64 cases. Most of those cases involve a false premise (hence are vacuously true). Let's do a case that is not vacuous. Notice that $(1, 2) \in R$. In addition, there are three elements c such that $(2, c) \in R$, namely $c = 1, 2, 3$. Since each of $(1, 1)$, $(1, 2)$, and $(1, 3)$ are in R, we see that whenever $(1, 2) \in R$ and $(2, c) \in R$, we have $(1, c) \in R$. Repeating this process with each element of R in place of $(1, 2)$, we see that R is transitive.

(Not antisymmetric): Fix $a = 1$ and $b = 2$ in R. We find $1R2$ and $2R1$, but $1 \neq 2$. △

Remark 20.15. Notice that when a relation R is given explicitly by a set, it can be difficult to check transitivity. As we will see in the following examples, transitivity is often easier to check when R is defined by a rule. ▲

Example 20.16. Let R be the relation on $\mathbb{R}$ given by aRb if $b - a \in [0, \infty)$. This relation is really just $\leq$, so we will write it using that symbol. Let's check each of the four properties.

(Reflexive): Let $a \in \mathbb{R}$. We have $a \leq a$.

(Not symmetric): Fix $a = 0$ and $b = 1$ in $\mathbb{R}$. We have $0 \leq 1$ but $1 \not\leq 0$.

(Transitive): Let $a, b, c \in \mathbb{R}$. Assume $a \leq b$ and $b \leq c$. Then $a \leq c$.

(Antisymmetric): Let $a, b \in \mathbb{R}$. Assume $a \leq b$ and $b \leq a$. This forces $a = b$. △

In the examples above we either proved a property holds or it fails. Thus, we have been using the negations of each of the properties. The following table lists all of the negations. (We leave it as an easy exercise to identify which row corresponds to which property.)

Definition	Negation
$\forall x \in A,\ xRx$	$\exists x \in A,\ x \not R x$
$\forall x, y \in A,\ xRy \Rightarrow yRx$	$\exists x, y \in A,\ xRy \wedge y \not R x$
$\forall x, y, z \in A,\ xRy \wedge yRz \Rightarrow xRz$	$\exists x, y, z \in A,\ xRy \wedge yRz \wedge x \not R z$
$\forall x, y \in A,\ (xRy \wedge yRx) \Rightarrow x = y$	$\exists x, y \in A,\ xRy \wedge yRx \wedge x \neq y$

Example 20.17. Let U be a nonempty set, and let $S = \mathscr{P}(U)$. We can define a relation R on S by ARB if $A \subseteq B$ (where $A, B \in S$ are subsets of U). Then R is easily seen to be reflexive (since every set is a subset of itself), transitive (since $A \subseteq B$ and $B \subseteq C$ implies $A \subseteq C$), and antisymmetric (since $A \subseteq B$ and $B \subseteq A$ implies $A = B$), but not symmetric. (Can you prove this last statement?) △

Example 20.18. Let A be any nonempty set, and let $R = \emptyset$ be the empty relation on A. We note that R is not reflexive (since, for $a \in A$, we have $a \not R a$). On the other hand, R is symmetric, transitive, and antisymmetric, since the implications defining these properties are vacuously true for R (since no elements of A are related by R). △

Example 20.19. Let A be any nonempty set, and let $R = A \times A$. Then, for any $a, b \in A$, we have aRb. The relation R is easily seen to be reflexive, symmetric, and transitive. It is antisymmetric if and only if $|A| = 1$. △

Example 20.20. Let $A = \mathbb{N}$, and for $a, b \in \mathbb{N}$, define aRb if, for some $x \in \mathbb{N}$, $ax = b$. Note that saying that aRb is the same as saying that $a \mid b$; this relation is just divisibility, so we will write $a \mid b$ instead of aRb. Before reading further, try to decide which of the four properties hold for this relation.

(Reflexive): Given any $a \in \mathbb{N}$, we know $a \mid a$ (because $a \cdot 1 = a$).

(Transitive): If $a \mid b$ and $b \mid c$ then $a \mid c$, see Proposition 7.13.

(Antisymmetric): In addition, R is antisymmetric, since $a \mid b$ and $b \mid a$ implies that $a = b$. (See Corollary 17.2. Note that the relation in this example is defined on $\mathbb{N}$, not $\mathbb{Z}$. The divisibility relation on $\mathbb{Z}$ is *not* antisymmetric; take $a = 1$ and $b = -1$.)

(Not symmetric): Note that $1 \mid 2$ but $2 \nmid 1$. △

20.C Exercises

Exercise 20.1. Let $A = \{1, 2, 3, 4, 5, 6\}$ and let

$$R = \{(1,1), (1,2), (1,3), (1,4), (2,3), (2,5), (2,6), (3,5), (4,5), (4,6)\}.$$

(a) Give an example of elements $a, b \in A$ such that aRb.
(b) Give an example of elements $a, b \in A$ such that $a \not R b$.
(c) For $a \in A$, let $S_a = \{x \in A : aRx\}$. Thus, S_a is the set of elements to which a relates. Write down the six sets $S_1, \ldots, S_6$.
(d) For $a \in A$, let $T_a = \{x \in A : xRa\}$. Thus, T_a is the set of elements which relate to a. Write down the six sets $T_1, \ldots, T_6$.

Exercise 20.2. For the relations R from $\mathbb{R}$ to $\mathbb{R}$ defined below, write R as a set using set-builder notation, and graph R as a subset of $\mathbb{R}^2$.
(a) Define R by xRy if $x \leq y$.
(b) Define R by xRy if $x \geq y$.
(c) Define R by xRy if $x > y$.
(d) Define R by xRy if $x = y$.
(e) Define R by xRy if $x \neq y$.

Exercise 20.3. Define a relation R on $\mathbb{R}$ by xRy if $xy < 0$.
(a) Describe R as a set using set-builder notation.
(b) Graph R as a subset of $\mathbb{R} \times \mathbb{R}$.
(c) Determine whether R is reflexive, symmetric, transitive, and/or antisymmetric. (Give complete proofs.)

Exercise 20.4. Define a relation R on $\mathbb{R}$ by xRy if $x - y \in \mathbb{Z}$.
(a) Describe R as a set using set-builder notation.
(b) Graph R as a subset of $\mathbb{R} \times \mathbb{R}$.
(c) Determine whether R is reflexive, symmetric, transitive, and/or antisymmetric. (Give complete proofs.)

Exercise 20.5. Define a relation R on $\mathbb{Z}$ by aRb if $a - b$ is even.
(a) Describe R as a set using set-builder notation.
(b) Prove that R is reflexive, symmetric, and transitive.
(c) Prove that R is not antisymmetric.
(d) For which integers b is it the case that $1Rb$?

Exercise 20.6. For each part, give an example of a relation R on the set $A = \{1, 2, 3\}$ with the specified properties. Write R as a set of ordered pairs.
(a) R is reflexive, symmetric, and transitive.
(b) R is reflexive and symmetric, but not transitive.
(c) R is reflexive and transitive, but not symmetric.
(d) R is reflexive, not symmetric, and not transitive.
(e) R is not reflexive, but is symmetric and transitive.
(f) R is not reflexive and not transitive, but is symmetric.
(g) R is not reflexive and not symmetric, but is transitive.
(h) R is not reflexive, not symmetric, and not transitive.

Exercise 20.7. Let $A = \{1, 2, 3, 4, 5\}$ and let $S = \{\{1, 2, 3\}, \{3, 4\}, \{5\}\}$. Define

$$R = \{(a, b) \in A \times A : \text{for some } X \in S, \text{ both } a \in X \text{ and } b \in X\}.$$

(See Example 20.6 for a similar construction.)

(a) Write out the elements of R.
(b) Is R reflexive? Symmetric? Transitive? (Give complete proofs.)

Exercise 20.8. Let $A = \{1, 2, 3, 4, 5\}$ and let $S = \{\{1, 2\}, \{4, 5\}\}$. Define

$$R = \{(a, b) \in A \times A : \text{for some } X \in S, \text{ both } a \in X \text{ and } b \in X\}.$$

(See Example 20.6 for a similar construction.)
(a) Write out the elements of R.
(b) Is R reflexive? Symmetric? Transitive? (Give complete proofs.)

21 Equivalence relations

Knowing that we have a relation R on a set A tells us very little about either R or A. We can find many examples of relations on A, but without further information there is very little that we can say about a relation.

Placing conditions (such as "reflexive", "symmetric", and so forth) on a relation reduces the number of examples that we can find, but increases the amount that we know and can prove about each example.

One class of relations that has proven useful to mathematicians is the class of equivalence relations. Equivalence relations have enough conditions that we can prove useful theorems about them, but they are general enough that there are many examples of them in all areas of mathematics.

21.A Definition and examples

The definition of an equivalence relation is as follows:

Definition 21.1. Let R be a relation on a set A. We say that R is an *equivalence relation* if R is reflexive, symmetric, and transitive.

Equality is the prototypical example of an equivalence relation, but there are many more examples.

Example 21.2. Let $A = \{1, 2, 3, 4\}$. Define a relation R on A by

$$R = \{(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3), (4,4)\},$$

as in Example 20.14. We have seen that R is reflexive, symmetric, and transitive. Hence, R is an equivalence relation. △

Example 21.3. Let $A = \mathbb{Z}$ and for $a, b \in A$, we let aRb if a and b have the same parity. One easily checks that this relation is reflexive (an integer has the same parity as itself), symmetric (if a and b have the same parity, then so do b and a), and transitive (if a and b have the same parity, and b and c have the same parity, then a and c have the same parity). Hence, R is an equivalence relation. △

Example 21.4. Let $A = \mathbb{R}$ and for $a, b \in A$, we let aRb if $|a| = |b|$. Again, one checks easily that this is reflexive ($|a| = |a|$), symmetric ($|a| = |b|$ clearly implies that $|b| = |a|$), and transitive (if $|a| = |b|$ and $|b| = |c|$, then $|a| = |c|$). Hence R is an equivalence relation. △

Example 21.5. Let A be the set of all triangles. For $a, b \in A$, we let aRb if a is similar to b. (Recall from geometry that two triangles are similar if they have the same interior angles.). One sees easily that this is an equivalence relation. △

In many cases, equivalence relations are written using symbols other than R to indicate that two elements are related. A common symbol to use for a generic equivalence relation is $\sim$, which can be read "is equivalent to." Other symbols that might be used to represent equivalence relations include $\equiv$, $\cong$, $\approx$, $\cong$, $\simeq$, $\doteq$. Using these symbols for a relation that is not an equivalence relation can cause confusion. Note that these symbols can have many different meanings, so if you use them it is important to say what they mean. For instance, $\equiv$ could mean congruence modulo n or it could mean logical equivalence (or it could have another meaning) depending on where it occurs.

Example 21.6. Let $A = \mathbb{R} \times \mathbb{R}$. Define a relation $\sim$ on A by $(a, b) \sim (c, d)$ if $a^2 + b^2 = c^2 + d^2$. We check that $\sim$ is an equivalence relation.

Note that for $(a, b) \in \mathbb{R} \times \mathbb{R}$, we have $a^2 + b^2 = a^2 + b^2$. Hence, $(a, b) \sim (a, b)$, so $\sim$ is reflexive.

Suppose that $(a, b) \sim (c, d)$. Then $a^2 + b^2 = c^2 + d^2$, so $c^2 + d^2 = a^2 + b^2$, and $(c, d) \sim (a, b)$. Hence, $\sim$ is symmetric.

Finally, assume that $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$. Then $a^2 + b^2 = c^2 + d^2$ and $c^2 + d^2 = e^2 + f^2$, so $a^2 + b^2 = e^2 + f^2$, and we see that $(a, b) \sim (e, f)$. Hence $\sim$ is transitive.

Therefore, $\sim$ is an equivalence relation. △

Example 21.7. Let P, Q, R be statements and let A be the set of all compound statements formed from P, Q, and R. For $a, b \in A$, we let $a \equiv b$ if a is logically equivalent to b. One may check that $\equiv$ is an equivalence relation. △

Example 21.8. Let $A = \{1, 2, 3, 4, 5\}$ and let $S = \{\{1, 2\}, \{3, 4\}, \{5\}\}$.

Define a relation R on A by

$$R = \{(a, b) \in A \times A : \text{for some } X \in S, \text{ both } a \in X \text{ and } b \in X\}. \tag{21.9}$$

This is the same relation that we constructed in Example 20.6, where we saw that

$$R = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 3), (3, 4), (4, 3), (4, 4), (5, 5)\}. \tag{21.10}$$

We will now prove that R is an equivalence relation. We note two facts about S that will be important: (1) every element of A is a member of some element of S, and (2) the elements of S are disjoint sets; no two of them share an element of A. We will prove that R is an equivalence relation using (21.9), rather than the explicit list (21.10).

(Reflexive): Let $a \in A$. There is some $X \in S$ such that $a \in X$. Hence, $(a, a) \in R$, so aRa. Therefore, R is reflexive.

(Symmetric): Let $a, b \in A$ and assume aRb. Then $(a, b) \in R$, so for some $X \in S$ we have $a, b \in X$. Hence, $b, a \in X$, so $(b, a) \in R$, and we see that bRa. Therefore R is symmetric.

(Transitive): Let $a, b, c \in A$ and assume aRb and bRc. Then there is some $X \in S$ such that $a, b \in X$, and there is some $Y \in S$ such that $b, c \in Y$. Since the elements of S are disjoint, $b \in X$ and $b \in Y$ implies that $X = Y$. Therefore, both a and c are in X, and aRc. Therefore R is transitive. △

21.B Equivalence classes

The properties of equivalence relations are chosen to mimic properties of equality. Often when we have an equivalence relation it is useful to gather together elements that are related under the equivalence relation and to treat them as if they were equal. Forming equivalence classes is the tool that performs this gathering.

Definition 21.11. Let A be a set and let $\sim$ be an equivalence relation on A. For an element $a \in A$, we define the *equivalence class of* a by

$$[a] = \{x \in A : a \sim x\}.$$

The element a is called a *representative* of the class $[a]$.

Other common notations for the equivalence class of a are $\overline{a}$, $\hat{a}$, or $\tilde{a}$. These symbols are used to represent the equivalence classes for many different equivalence relations; thus, if you use one of these notations, you must define it. Similarly, if you see such symbols in mathematical writing, you should look to see how they are defined.

Example 21.12. Let $A = \{1, 2, 3, 4\}$ and define a relation R on A by

$$R = \{(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3), (4,4)\}.$$

We saw in Example 20.14 that R is an equivalence relation. We now determine its equivalence classes.

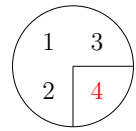
First, $[1]$ consists of all elements $x \in A$ such that $1Rx$. Examining R, we find that $1R1$, $1R2$, $1R3$, and $1\not R4$. Hence, $[1] = \{1, 2, 3\}$.

Now $[2]$ consists of all elements $x \in A$ such that $2Rx$. Examining R, we find that $2R1$, $2R2$, $2R3$, and $2\not R4$. Hence, $[2] = \{1, 2, 3\}$.

Similarly, $[3] = \{1, 2, 3\}$.

Finally, $[4] = \{4\}$, since the only $x \in A$ with $4Rx$ is $x = 4$.

We note that there are two equivalence classes in A, namely $[1] = [2] = [3] = \{1, 2, 3\}$ and $[4] = \{4\}$. These classes divide the set A into two subsets, as in the diagram below.



△

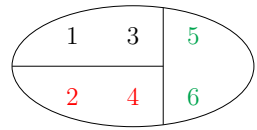
Example 21.13. Let $A = \{1, 2, 3, 4, 5, 6\}$, and define a relation $\sim$ on A by

$$R = \left\{ \begin{array}{l} (1,1), (1,3), (2,2), (2,4), (3,1), (3,3), \\ (4,2), (4,4), (5,5), (5,6), (6,5), (6,6) \end{array} \right\}.$$

Then R is an equivalence relation. We will determine the equivalence classes of R.

To determine [1], we look for all of the ordered pairs $(1, x) \in R$. We see that the only x's which work are 1 and 3. Hence, $[1] = \{1, 3\}$.

Similarly, $[2] = \{2, 4\}$, $[3] = \{1, 3\}$, $[4] = \{2, 4\}$, $[5] = \{5, 6\}$, and $[6] = \{5, 6\}$. We note that there are three equivalence classes, since $[1] = [3]$, $[2] = [4]$ and $[5] = [6]$. Because A is finite, we can easily draw a picture illustrating how the equivalence classes divide A into three pieces.



△

Example 21.14. Let $A = \mathbb{Z}$, and let $\sim$ be the relation defined on A by $a \sim b$ if a and b have the same parity. We saw in Example 21.3 that $\sim$ is an equivalence relation. We will compute the equivalence classes.

The equivalence class [0] consists of all numbers having the same parity as 0. Hence, $[0] = \{\text{even integers}\}$. The equivalence class [1] consists of all numbers having the same parity as 1. Hence, $[1] = \{\text{odd integers}\}$.

Notice that if a is any even integer, $[a] = \{\text{even integers}\} = [0]$, and if a is any odd integer, $[a] = \{\text{odd integers}\} = [1]$. Hence, in this case, there are exactly two equivalence classes, each containing infinitely many elements. Each class also has infinitely many representatives. For instance, $\cdots = [-2] = [0] = [2] = [4] = \cdots$. △

Example 21.15. Let $A = \mathbb{R}$, and for $a, b \in A$, we let $a \mathrel{\doteq} b$ if $|a| = |b|$. We saw in Example 21.4 that $\doteq$ is an equivalence relation. For $a \in A$, we will denote the equivalence class of a by $[a]$.

We see that $[0] = \{x \in \mathbb{R} : 0 \doteq x\} = \{x \in \mathbb{R} : |0| = |x|\}$. There is only one such value of x, namely $x = 0$. Hence, $[0] = \{0\}$.

Now, $[1] = \{x \in \mathbb{R} : 1 \doteq x\} = \{x \in \mathbb{R} : |1| = |x|\}$, or in other words the real numbers with absolute value 1. There are two such numbers: 1 and -1. Hence, $[1] = \{1, -1\}$.

Moving to negative numbers, $[-2] = \{x \in \mathbb{R} : -2 \doteq x\} = \{x \in \mathbb{R} : |-2| = |x|\}$, or in other words, the real numbers with absolute value 2. There are two such numbers: 2 and -2. Hence, $[-2] = \{2, -2\}$.

In general, if $a \in \mathbb{R}$ and $a \neq 0$, we see that $[a] = \{a, -a\}$.

We see that there are infinitely many different equivalence classes in R, each having one or two elements. △

We conclude by stating a very simple theorem that is really just a restatement of the definition of an equivalence class. Nevertheless, the restatement is quite useful to help us recall how to tell whether an element is in an equivalence class.

Theorem 21.16. *Let $\sim$ be an equivalence relation on a set A and let $a, b \in A$. Then $a \sim b$ if and only if $b \in [a]$.*

Proof. Assume that $a \sim b$. Then, by the definition of $[a]$, we have $b \in [a]$.

Assume that $b \in [a]$. Then, by the definition of $[a]$, we have $a \sim b$. □

21.C Exercises

Exercise 21.1. Let $A = \{1, 2, 3\}$ and let R be the relation on A given by

$$R = \{(1, 1), (1, 2), (2, 1), (2, 2), (2, 3), (3, 2), (3, 3)\}.$$

Is R reflexive? Symmetric? Transitive? Antisymmetric? Is R an equivalence relation? (Give proofs.)

Exercise 21.2. Give an example of an equivalence relation R on the set $A = \{1, 2, 3, 4, 5\}$ that has exactly two equivalence classes. Explicitly write out the relation R as a set of ordered pairs.

Exercise 21.3. Let R be an equivalence relation on the set $A = \{1, 2, 3, 4, 5\}$. Assume that $1R3$ and $3R4$. Given these conditions, which ordered pairs must belong to R? (Hint: There are at least 11 such elements.)

Exercise 21.4. Let $A = \{1, 2, 3, 4, 5\}$ and let $S = \{\{1, 2\}, \{3, 4\}, \{5\}\}$. Define

$$R = \{(a, b) \in A \times A : \text{for some } X \in S, \text{ both } a \in X \text{ and } b \in X\}$$

as in Example 21.8. We have seen that R is an equivalence relation. What are the equivalence classes of R?

Exercise 21.5. Let $A = \mathbb{R} - \{0\}$. Define a relation $\sim$ on A by $a \sim b$ if $ab > 0$.

(a) Prove that $\sim$ is an equivalence relation on A.
(b) Determine the equivalence classes of $\sim$.

Exercise 21.6. Let A be the set of humans with English names. Define a relation $\approx$ on A by $\alpha \approx \beta$ if α and β have the same first letter in their first names. (For instance, anyone named "Eugene" is related by $\approx$ to anyone named "Elizabeth".)

(a) Prove that $\approx$ is an equivalence relation on A.
(b) Determine the equivalence classes of $\approx$.

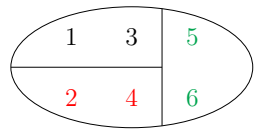
Exercise 21.7. Let A be a set. Let R be a reflexive, symmetric, and antisymmetric relation on A. Prove that R is equality on A. (In other words, prove that for any $x, y \in A$, we have xRy if and only if $x = y$.)

22 Equivalence classes and partitions

In this section we prove that an equivalence relation allows us to break up a set into jigsaw-like pieces that one can fit together to give the entire set. Conversely, we will prove that if you break a set into jigsaw pieces, then one can define a corresponding equivalence relation. We saw an example of this earlier, with the set $A = \{1, 2, 3, 4, 5, 6\}$. The equivalence relation

$$R = \left\{ \begin{array}{l} (1,1),(1,3),(2,2),(2,4),(3,1),(3,3), \\ (4,2),(4,4),(5,5),(5,6),(6,5),(6,6) \end{array} \right\}$$

corresponded to breaking A into the equivalence classes pictured below.



This correspondence holds for arbitrary equivalence relations, as we will prove in this section.

22.A Properties of equivalence classes

The following theorem provides some useful criteria for determining whether two equivalence classes are equal.

Theorem 22.1. *Let $\sim$ be an equivalence relation on a set A, and for any $x \in A$ denote the equivalence class of x by $[x]$. For any $x, y \in A$, the following are equivalent:*

(1) $x \sim y$ *(the elements are related).*
(2) $[x] \cap [y] \neq \emptyset$ *(the classes intersect nontrivially).*
(3) $[x] = [y]$ *(the classes are equal).*

Proof. We prove the equivalence by proving three implications: (1)⇒(3), (3)⇒(2), and (2)⇒(1). This will prove that we can move from any one condition to any other.

(1)⇒(3)**:** Assume $x \sim y$. By symmetry we also have $y \sim x$. We wish to show $[x] = [y]$. We will prove this equality by showing both inclusions.

First we will show $[x] \subseteq [y]$. Let $z \in [x]$. We then know $x \sim z$. Since $y \sim x$ and $x \sim z$, by transitivity we obtain $y \sim z$. Hence $z \in [y]$, and we have shown $[x] \subseteq [y]$.

The other inclusion, $[y] \subseteq [x]$, is proved similarly.

(3)⇒(2)**:** Now assume $[x] = [y]$. By the reflexive property, $x \sim x$. Thus $x \in [x] = [y]$. Hence $[x] \cap [y] \neq \emptyset$ since x is in the intersection.

(2)⇒(1)**:** Finally, assume $[x] \cap [y] \neq \emptyset$. Fix $z \in [x] \cap [y]$. We then have $x \sim z$ and $y \sim z$. By symmetry and transitivity, $x \sim y$. □

This theorem, along with Theorem 21.16 from the previous section, allows us to prove three central properties for equivalence classes. Let A be a set and let $\sim$ be an equivalence relation on A. The following all hold:

(1) **(Nonempty pieces)** No equivalence class is empty.

Proof. Let $[x]$ be an equivalence class. Since $x \sim x$ by the reflexive property, we then have $x \in [x]$. □

(2) **(Covering)** Every element of A is an element of some equivalence class.

Proof. An arbitrary element $a \in A$ belongs to the equivalence class $[a]$ (again, by the reflexive property). □

(3) **(Disjoint pieces)** Any two classes are either disjoint or equal.

Proof. Since (2) implies (3) in Theorem 22.1, two classes that are not disjoint are equal. □

We can think of these three conditions as describing how to break a set into jigsaw pieces. First, none of the pieces should be empty. (Empty jigsaw pieces would just fall out of the box, and they wouldn't be missed anyway!) Second, when you put all of the pieces together, you should get the entire puzzle (in this case, the entire set A). Third, none of the pieces should overlap (else they don't fit together into a jigsaw puzzle).

We have proved that these three facts about equivalence classes are necessary. In the next subsection, we will show that they are sufficient to describe a (unique) equivalence relation with the same equivalence classes.

22.B Partitions and equivalence classes

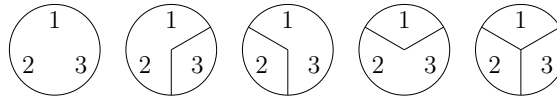
The three facts mentioned at the end of the previous subsection help us to define the concept of a partition, as follows:

Definition 22.2. Let A be a set. A collection P of subsets of A is called a *partition* of A if the following three conditions hold:
(Nonempty pieces) No set in P is empty.
(Covering) Every element of A is a member of some element of P.
(Disjoint pieces) Any two distinct elements of P are disjoint.
The elements of P are called the *parts* of the partition.

Example 22.3. Let $A = \{1, 2, 3\}$. Then the following are partitions of A:

$$P_1 = \{\{1, 2, 3\}\},$$
$$P_2 = \{\{1, 2\}, \{3\}\},$$
$$P_3 = \{\{1, 3\}, \{2\}\},$$
$$P_4 = \{\{2, 3\}, \{1\}\},$$
$$P_5 = \{\{1\}, \{2\}, \{3\}\}.$$

If we wish to represent these five partitions graphically, we can do so as below in which the set $A = \{1, 2, 3\}$ is represented as a circle. The parts of the partition are separated by line segments.



Careful examination shows that there are no other partitions of A.

Each of the following collections is *not* a partition of A. Try to figure out what is wrong in each case before reading the answers below.

(1) $\{\emptyset, \{1, 2\}, \{3\}\}$
(2) $\{\{1, 3\}\}$
(3) $\{\{1, 2\}, \{2, 3\}\}$
(4) $\{\{1, 2\}, \{3, 4\}\}$
(5) $\{1, 2, 3\}$

Here are the reasons each of the sets above is not a partition of A:

(1) This collection fails the "nonempty pieces" condition, since one of the pieces is empty.
(2) This collection fails the "covering" condition, since we are missing 2.
(3) This collection fails the "disjoint pieces" condition, since the two distinct pieces overlap.
(4) This collection does not consist of subsets of A. (It would be a partition for the new set $B = \{1, 2, 3, 4\}$.)
(5) This collection is not a set of sets. (It would be a partition with an extra pair of set braces around it, and would then equal P_1 above.)

△

Example 22.4. Let's partition $\mathbb{N}$ into four parts. One such partition would be the following:

$$P = \{\{1\}, \{2\}, \{3\}, \{x \in \mathbb{N} : x \geq 4\}\}.$$

Three of the pieces are very small, and the other piece is infinite.

Another possibility would be the following, somewhat more natural, example. Given $n \in \mathbb{N}$ let $S_n = \{x \in \mathbb{N} : x \equiv n \pmod 4\}$. Our new partition is

$$P' = \{S_1, S_2, S_3, S_4\}.$$

This is a partition because every natural number is congruent to exactly one of the four numbers $1, 2, 3, 4$. △

Example 22.5. There are fifteen different partitions of the set $A = \{1, 2, 3, 4\}$. We do not list them all, but mention that there is one partition with one part, seven partitions with two parts, six partitions with three parts, and one partition with four parts. As an example, $\{\{1, 2\}, \{3, 4\}\}$ is a partition of A with two parts, and $\{\{1\}, \{2\}, \{3, 4\}\}$ is a partition of A with three parts. △

The following theorem asserts that from an equivalence relation we can build a partition.

Theorem 22.6. *Let $\sim$ be an equivalence relation on a set A. For any $a \in A$, denote its equivalence class by $[a]$. Then the set*

$$P = \{[a] : a \in A\}$$

is a partition of A.

Proof. We already proved above that the set of equivalence classes satisfies all three of the defining conditions for a partition. □

The next two examples will demonstrate how we pass from an equivalence relation to a partition.

Example 22.7. Let $A = \{1, 2, 3\}$ and let R be the relation defined by

$$R = \{(1, 1), (2, 2), (2, 3), (3, 2), (3, 3)\}.$$

We can easily confirm that R is an equivalence relation. The equivalence classes are $[1] = \{1\}$ and $[2] = \{2, 3\} = [3]$. Hence, the set of equivalence classes is $\{\{1\}, \{2, 3\}\}$, which is the partition P_4 from Example 22.3. △

Example 22.8. Consider the "same parity" relation on $\mathbb{Z}$. The two equivalence classes are the sets of even integers and odd integers. This partitions $\mathbb{Z}$ into two disjoint sets. △

The next theorem allows us to pass from a partition to an equivalence relation, thereby reversing the direction of the previous theorem.

Theorem 22.9. *Let P be a partition of a set A. There is an equivalence relation $\sim$ on A such that the equivalence classes of $\sim$ are precisely the parts of P.*

Proof. Let P be a partition of A. Define the relation $\sim$ on A by $a \sim b$ if a and b are both in a common part of P. We will show that $\sim$ is an equivalence relation.

First note that if A is empty (so that P is also empty), then symmetry and transitivity hold vacuously, while reflexivity holds since there is nothing to check. So, hereafter we assume A is nonempty.

(Reflexive): Let $a \in A$. By the covering property, a is an element of some part of P. Hence, $a \sim a$, so $\sim$ is reflexive.

(Symmetric): Let $a, b \in A$ and assume $a \sim b$. Thus, there is some part $X \in P$, such that $a \in X$ and $b \in X$. Hence, $b, a \in X$, so $b \sim a$. Thus, $\sim$ is symmetric.

(Transitive): Let $a, b, c \in A$ and assume $a \sim b$ and $b \sim c$. So $a, b \in X$ and $b, c \in Y$, where X and Y are parts of P. Since b is an element of both X and Y, we

must have $X = Y$ (by the disjoint pieces property), so $a, c \in X$, and hence $a \sim c$. Thus, $\sim$ is transitive, and this finishes the proof that $\sim$ is an equivalence relation.

Finally, we will prove that the equivalence classes of $\sim$ are exactly the parts of P. For any $a \in A$ there is exactly one part $X \in P$ with $a \in X$ (by the covering and disjoint pieces conditions). By the definition of $\sim$, the elements of $[a]$ are exactly the elements of X, so $[a] = X$. Hence, the equivalence classes of $\sim$ are parts of P. Conversely, we show that the parts of P are equivalence classes. Given a part $X \in P$, it is nonempty. So fix $a \in X$. By the same argument above we obtain $[a] = X$, so it is an equivalence class. □

Together, Theorems 22.6 and 21.3 tell us that partitions of a set A and equivalence relations on A correspond to each other; every partition gives an equivalence relation and every equivalence relation gives a partition. The following example demonstrates this fact concretely.

Example 22.10. We partition $\mathbb{Z}$ into three pieces, in the following way:

$$P = \{\{0\}, \{x \in \mathbb{Z} : x < 0\}, \{x \in \mathbb{Z} : x > 0\}\}.$$

In other words, we partition $\mathbb{Z}$ into its zero, negative, and positive pieces.

What is the corresponding equivalence relation? It is the "two elements are related if they are both positive, both zero, or both negative" relation. More formally,

$$R = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x, y \in \mathbb{Z}_{>0} \text{ or } x = y = 0 \text{ or } x, y \in \mathbb{Z}_{<0}\}. \qquad \triangle$$

22.C Transversals of equivalence relations

On occasion it is important to be able to refer to representatives of equivalence classes. When we pick one representative for each class, the collection of those representatives has a special name.

Definition 22.11. Let $\sim$ be an equivalence relation on a set A. A *transversal* of $\sim$ is a set $S \subseteq A$, such that S consists of exactly one representative of every equivalence class of $\sim$.

Example 22.12. Let $A = \{1, 2, 3, 4, 5\}$ and let

$$R = \{(1,1), (1,2), (2,1), (2,2), (3,3), (4,4), (4,5), (5,4), (5,5)\}.$$

Then R is an equivalence relation on A. The equivalence classes of R are

$$[1] = [2] = \{1, 2\}, \quad [3] = \{3\}, \quad [4] = [5] = \{4, 5\}.$$

Then the set $\{1, 3, 5\}$ is a transversal, since it consists of a single element from each equivalence class. Other transversals are $\{1, 3, 4\}$, $\{2, 3, 4\}$, and $\{2, 3, 5\}$. $\triangle$

Example 22.13. Define a relation $\sim$ on $\mathbb{R}$ by $a \sim b$ if $a - b \in \mathbb{Z}$. Thus, two real numbers are related by $\sim$ if they are an integer apart from one another. So, for instance, π is related to $\pi - 3$. We will show that $\sim$ is an equivalence relation, and that the set $[0, 1)$ is a transversal of $\sim$.

(Reflexive): Let $a \in \mathbb{R}$. We have $a - a = 0 \in \mathbb{Z}$, and so $a \sim a$.

(Symmetric): Let $a, b \in \mathbb{R}$, and assume $a \sim b$. Thus, $a - b \in \mathbb{Z}$, so $b - a = -(a - b) \in \mathbb{Z}$, and $b \sim a$.

(Transitive): Let $a, b, c \in \mathbb{R}$. Assuming $a \sim b$ and $b \sim c$, then we know that $a - b, b - c \in \mathbb{Z}$. Adding, we find that $a - c = (a - b) + (b - c) \in \mathbb{Z}$, so $a \sim c$.

Hence, $\sim$ is an equivalence relation.

We now proceed to show that $[0, 1)$ is a transversal of $\sim$. Define the *floor* of x, written $\lfloor x \rfloor$, to be the unique integer such that $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$. Thus, $\lfloor x \rfloor$ is obtained by "rounding down," even when x is negative. (So $\lfloor 5.6893 \rfloor = 5$, $\lfloor -3.49583 \rfloor = -4$, and $\lfloor \pi \rfloor = 3$.) The quantity $x - \lfloor x \rfloor$ is called the *fractional part* of x and is in $[0, 1)$.

First, every element of $\mathbb{R}$ is related to its fractional part, since x minus its fractional part $x - \lfloor x \rfloor$ is the integer $\lfloor x \rfloor$. This shows that some element from each equivalence class is a member of $[0, 1)$. Finally, if $a, b \in [0, 1)$ with $a \neq b$, then $a \not\sim b$, since $0 < |a - b| < 1$, so $a - b \notin \mathbb{Z}$. Thus, no more than one element from any equivalence class belongs to $[0, 1)$. △

Advice 22.14. There are typically two steps to proving that a given set T is a transversal of an equivalence relation $\sim$ on a set A.

First, you need to prove that every element of A is related to at least one element of T. Often this can be done by giving an explicit construction; for any element a of A, construct an element $t \in T$ such that $a \sim t$.

Second, you need to prove that every element of A is related to at most one element of T. This can be done directly, by assuming, for $t_1, t_2 \in T$, that $a \sim t_1$ and $a \sim t_2$, and proving that t_1 must equal t_2; it can also be done simply by showing that any two elements of T that are related to each other are actually equal. (The reader should convince themselves that either of these two statements suffices.)

Example 22.15. Let $\sim$ be the relation on $\mathbb{R} \times \mathbb{R}$ defined by $(a, b) \sim (c, d)$ if $a + d = b + c$. We will show that $\sim$ is an equivalence relation and give a transversal of $\sim$.

(Reflexive): Let $(a, b) \in \mathbb{R} \times \mathbb{R}$. Then $(a, b) \sim (a, b)$ since $a + b = a + b$.

(Symmetric): Let $(a, b), (c, d) \in \mathbb{R} \times \mathbb{R}$. Assume that $(a, b) \sim (c, d)$. Then $a + d = b + c$. If follows that $c + b = a + d$, and we see that $(c, d) \sim (a, b)$.

(Transitive): Let $(a, b), (c, d), (e, f) \in \mathbb{R} \times \mathbb{R}$. Assume that $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$. Then $a + d = b + c$ and $c + f = d + e$. Adding these equations, we see that $a + d + c + f = b + c + d + e$. Subtracting $c + d$ from both sides of this equality, we see that $a + f = b + e$, so $(a, b) \sim (e, f)$.

Hence, $\sim$ is reflexive, symmetric, and transitive, so it is an equivalence relation.

Let $T = \{(0, d) : d \in \mathbb{R}\}$. We will show that T is a transversal for $\sim$.

Let $(a, b) \in \mathbb{R}\times\mathbb{R}$. Then $(0, b-a) \in T$, and $(a, b) \sim (0, b-a)$, since $a+(b-a) = b+0$. Hence, every element of $\mathbb{R} \times \mathbb{R}$ is related to at least one element of T.

Now, suppose that $(0, d_1) \sim (0, d_2)$. Then $0 + d_2 = d_1 + 0$, hence $d_1 = d_2$. So $(0, d_1) = (0, d_2)$. This shows that no element of $\mathbb{R} \times \mathbb{R}$ is related to more than one element of T. Therefore T is a transversal. △

22.D Exercises

Exercise 22.1. Let $A = \{1, 2, 3, 4\}$. List one partition of A with one part, seven partitions of A with two parts, six partitions of A with three parts, and one partition of A with four parts. This gives a total of fifteen partitions of A. (There are no more, but you do not need to prove this.)

Exercise 22.2. Let A be a set with $|A| = 10$, and let $\sim$ be an equivalence relation on A. Denote the equivalence classes of $\sim$ by $[x]$, for each $x \in A$. Suppose that we have elements $a, b, c \in A$ with $|[a]| = 3$, $|[b]| = 5$, and $|[c]| = 1$.

(a) Are any of a, b, and c related by $\sim$?
(b) How many equivalence classes for $\sim$ are there in A?

Exercise 22.3. Define a relation $\sim$ on $\mathbb{R}^2$ by $(a, b) \sim (c, d)$ if $a^2 + b^2 = c^2 + d^2$. We saw in Example 21.6 that $\sim$ is an equivalence relation.

(a) Describe the equivalence class $[(3, 4)]$, both as a set and geometrically.
(b) For an arbitrary element $(a, b) \in \mathbb{R}^2$, describe $[(a, b)]$.
(c) Prove that the set $[0, \infty) \times \{0\}$ is a transversal of $\sim$.

Exercise 22.4. Let W be the set of all words in the English language. Define a relation on W by $\alpha \approx \beta$ if α and β have the same first letter.

(a) Prove that $\approx$ is an equivalence relation.
(b) Let $[\alpha]$ be the equivalence class of $\alpha \in W$. For $\alpha =$ "cat", list six elements of $[\alpha]$.
(c) How many equivalence classes are there in W for $\approx$?
(d) Describe a transversal of $\approx$. (You do not need to write it down in full.)

Exercise 22.5. Let A be a set with n elements. Define a relation $\sim$ on $\mathscr{P}(A)$ by $X \sim Y$ if $|X| = |Y|$, for any $X, Y \in \mathscr{P}(A)$.

(a) Prove that $\sim$ is an equivalence relation.
(b) Describe the equivalence classes for $\sim$.
(c) How many equivalence classes are there for $\sim$?
(d) Describe a transversal of $\sim$.
(e) How many elements of $\mathscr{P}(A)$ are in each equivalence class?

Exercise 22.6. Let $A = \{1, 2, \ldots, 10\}$. For each $i \in A$, define

$$S_i = \{X \in \mathscr{P}(A) \,:\, i \text{ is the least element of } X\}$$

Let $P = \{\{\emptyset\}, S_1, \ldots, S_{10}\}$.

(a) Prove that P is a partition of $\mathscr{P}(A)$. (Hint: Are the parts of P empty? Disjoint? Is every element of $\mathscr{P}(A)$ in some part?)

(b) Let $\sim$ be the equivalence relation on $\mathscr{P}(A)$ corresponding to P. How many equivalence classes does $\sim$ have?
(c) Write down a transversal of $\sim$.
(d) Determine the equivalence class $[\{8, 9, 10\}]$ by listing its elements.
(e) How large is the equivalence class $[\{2, 3, 4\}]$? (Do not write down all of its elements.)

Exercise 22.7. Give another proof of Theorem 22.1, by proving that (1) implies (2), that (2) implies (3), and that (3) implies (1). Note that your proof should not use any theorems after or including Theorem 22.1; it should only use basic properties of equivalence relations and equivalence classes.

Exercise 22.8. In this exercise we will show that an equivalence relation is uniquely determined by its equivalence classes, as follows: Let A be any set. Let $\sim$ and $\approx$ be two equivalence relations on A. Show that if their equivalence classes are the same, then the relations are the same. (In other words, conclude that for all $a, b \in A$, we have $a \sim b$ if and only if $a \approx b$.)

(Hint: For ease of notation, write the equivalence classes for $\sim$ using $[a]$, and the equivalence classes for $\approx$ using $\overline{a}$.)

23 Integers modulo n

In this section we will study integer congruence and prove it is an equivalence relation.

23.A Review of integer congruence

Recall the definition of congruence modulo n (Definition 8.11).

Definition 23.1. Let $a, b, n \in \mathbb{Z}$. We say that $a \equiv b \pmod{n}$ if $n \mid (a - b)$.

We read $a \equiv b \pmod{n}$ as "a is congruent to b modulo n" or "a is congruent to b mod n." This defines a relation on $\mathbb{Z}$. (Two elements are related if they are congruent modulo n.) We refer to the relation as "congruence modulo n."

Theorem 23.2. *Congruence modulo n is an equivalence relation on $\mathbb{Z}$.*

Proof. **(Reflexive):** Let $a \in \mathbb{Z}$. We have $n \mid 0$, so $n \mid (a - a)$. Hence, $a \equiv a \pmod{n}$, and so congruence modulo n is reflexive.

(Symmetric): Let $a, b \in \mathbb{Z}$, and assume $a \equiv b \pmod{n}$. Then $n \mid (a - b)$, so $(a - b) = nk$ for some $k \in \mathbb{Z}$. This implies that $b - a = n(-k)$, so $n \mid (b - a)$, and $b \equiv a \pmod{n}$. Hence, congruence modulo n is symmetric.

(Transitive): Let $a, b, c \in \mathbb{Z}$, and assume both $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$. Then $a - b = nk$ and $b - c = n\ell$ for some $k, \ell \in \mathbb{Z}$. Then

$$a - c = (a - b) + (b - c) = nk + n\ell = n(k + \ell),$$

so $a \equiv c \pmod{n}$. Hence, congruence modulo n is transitive. □

23.B Congruence classes modulo n

Having verified that congruence modulo n is an equivalence relation, we can utilize our knowledge of equivalence relations from the previous sections. In particular, we can talk about equivalence classes under the relation of congruence modulo n.

For instance, congruence modulo 3 is an equivalence relation. The set of integers that are congruent to 0 modulo 3 is

$$\{\ldots, -6, -3, 0, 3, 6, 9, \ldots\}.$$

This is the equivalence class of 0, under the relation "congruence modulo 3," which is also the equivalence class of 3, 6, -3, and so forth. We will see in a moment that there are only two other equivalence classes, which are

$$\{\ldots, -5, -2, 1, 4, 7, \ldots\}$$

and

$$\{\ldots, -4, -1, 2, 5, 8, \ldots\}.$$

We will now set up the notation to prove this in much greater generality.

Definition 23.3. An equivalence class in $\mathbb{Z}$ under the relation congruence modulo n is called a *congruence class modulo* n. The congruence class of $a \in \mathbb{Z}$ will be denoted $\overline{a}$.

Given $a, b \in \mathbb{Z}$, we see by Theorem 22.1 that $\overline{a} = \overline{b}$ if and only if $a \equiv b \pmod{n}$.

We know by Theorem 22.6 that the equivalence classes for an equivalence relation on $\mathbb{Z}$ form a partition of $\mathbb{Z}$. We will now proceed to determine the congruence classes mod n and the number of distinct classes. To do this we will make use of the division algorithm. In the following theorem we restrict to the case $n \in \mathbb{N}$, since allowing n to be 0 would make the theorem false (why?) and allowing n to be negative would require absolute values in the statement.

Theorem 23.4. *Let $n \in \mathbb{N}$. There are n congruence classes modulo n, namely the classes*

$$\overline{0}, \overline{1}, \overline{2}, \ldots, \overline{n-1}.$$

For any $a \in \mathbb{Z}$, we have

$$\overline{a} = \{a + kn : k \in \mathbb{Z}\}.$$

Proof. We first prove that the only congruence classes are the ones listed in the theorem. Let $a \in \mathbb{Z}$. By the division algorithm $a = qn + r$ for some $q, r \in \mathbb{Z}$ and $0 \leq r < n$. Thus $a - r = qn$, so $a \equiv r \pmod{n}$. Hence, $\overline{a} = \overline{r}$ is one of the n congruence classes listed in the theorem.

It remains to show that no two congruence classes listed in the theorem are equal. Suppose that $i, j \in \mathbb{Z}$, with $0 \leq i < n$ and $0 \leq j < n$ such that $i \neq j$. We will show that $\overline{i} \neq \overline{j}$, by showing that $i \not\equiv j \pmod{n}$. Assume, without loss of generality, that $0 \leq j < i < n$. Then $0 < i - j < n$, and since there are no multiples of n between 0 and n, we see that $n \nmid i - j$. Hence $i \not\equiv j \pmod{n}$, so $\overline{i} \neq \overline{j}$.

Finally, we note that for any $a \in \mathbb{Z}$,

$$\begin{aligned}
\overline{a} &= \{x \in \mathbb{Z} : x \equiv a \pmod{n}\} \\
&= \{x \in \mathbb{Z} : x - a = kn \text{ for some } k \in \mathbb{Z}\} \\
&= \{x \in \mathbb{Z} : x = a + kn \text{ for some } k \in \mathbb{Z}\} \\
&= \{a + kn : k \in \mathbb{Z}\}
\end{aligned}$$

as claimed. □

The set of congruence classes modulo n is so important that we give it a special symbol.

Definition 23.5. Let $n \in \mathbb{N}$. We define $\mathbb{Z}_n$ to be the set

$$\{\overline{0}, \overline{1}, \overline{2}, \ldots, \overline{n-1}\}$$

of congruence classes modulo n.

We note the following facts about $\mathbb{Z}_n$:
(a) $|\mathbb{Z}_n| = n$.
(b) Every element of $\mathbb{Z}$ is a member of exactly one element of $\mathbb{Z}_n$.
(c) Every element of $\mathbb{Z}_n$ is an infinite subset of $\mathbb{Z}$.

Example 23.6. The set $\mathbb{Z}_4$ has four elements. They are written $\overline{0}, \overline{1}, \overline{2}, \overline{3}$. Each of these elements, even though it looks like a number, is really an infinite set. For instance,

$$\overline{1} = \{\ldots, -7, -3, 1, 5, 9, \ldots\}. \qquad \triangle$$

Remark 23.7. We have said that the n elements $\overline{0}, \ldots, \overline{n-1}$ are the only elements in $\mathbb{Z}_n$. There are many other integers, in fact infinitely many, which are not elements of the set $\{0, 1, \ldots, n-1\}$. What happens to their congruence classes?

The proof of Theorem 23.4 shows that for *every* integer $a \in \mathbb{Z}$, the congruence class $\overline{a}$ is the same as one of the classes $\overline{r}$ with $0 \leq r < n$. We can think of $\overline{a}$ and $\overline{r}$ as different names for the same congruence class. For instance, in $\mathbb{Z}_6$ the congruence class $\overline{1}$ has many different representatives:

$$\cdots = \overline{-11} = \overline{-5} = \overline{1} = \overline{7} = \overline{13} = \overline{19} = \cdots . \qquad \blacktriangle$$

23.C Operations on $\mathbb{Z}_n$

Even though the elements of $\mathbb{Z}_n$ are sets (not numbers), we still want to treat these elements as if they *were* numbers. In fact, you have probably already done this without realizing it. For instance, when you add clock times together you are doing arithmetic in $\mathbb{Z}_{12}$. When you use the fact that "odd plus even equals odd" you are doing arithmetic in $\mathbb{Z}_2$.

The following theorem will help us make a sensible definition of addition and multiplication on the set $\mathbb{Z}_n$.

Theorem 23.8. *Let $n \in \mathbb{N}$. For $a, b, c, d \in \mathbb{Z}$, if $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, then*
(1) *$a + c \equiv b + d \pmod{n}$ and*
(2) *$ac \equiv bd \pmod{n}$.*

Proof. Suppose that $a \equiv b \pmod{n}$ and that $c \equiv d \pmod{n}$. Then $a - b = nk$ and $c - d = n\ell$ for some $k, \ell \in \mathbb{Z}$.

In order to prove that (1) holds, we examine $(a + c) - (b + d)$ and find

$$(a + c) - (b + d) = (a - b) + (c - d) = nk + n\ell = n(k + \ell),$$

so $a + c \equiv b + d \pmod{n}$. Hence, (1) is true.

To prove that (2) holds, we note that $a = b + nk$ and $c = d + n\ell$. Multiplying, we find that

$$ac = (b + nk)(d + n\ell) = bd + bn\ell + nkd + n^2k\ell = bd + n(b\ell + kd + nk\ell).$$

Thus $ac - bd = n(b\ell + kd + nk\ell)$, so $ac \equiv bd \pmod{n}$. Hence, (2) is true. $\square$

The next example demonstrates how using this theorem can help simplify computations modulo n.

Example 23.9. We have that $28 \equiv 2 \pmod{26}$ and $29 \equiv 3 \pmod{26}$. Hence, according to the theorem, $28 \cdot 29 \equiv 2 \cdot 3 \pmod{26}$. Indeed, computation shows that

$$28 \cdot 29 - 2 \cdot 3 = 812 - 6 = 806 = 26 \cdot 31.$$

Multiplying 2 and 3 is much easier than multiplying 28 and 29. Thus, finding $28 \cdot 29$ modulo 26 is very easy. You just multiply $2 \cdot 3 = 6$. △

The upshot of this theorem is that when adding and multiplying, if we are only interested in what the result is modulo n, then we only need to worry about what the inputs are modulo n. Motivated by this, we make a definition of what it means to add and multiply congruence classes mod n. (The following definition may be strange at first. If so, work through the examples which follow.)

Definition 23.10. Let $n \in \mathbb{N}$. Given $X, Y \in \mathbb{Z}_n$, then $X = \overline{a}$ and $Y = \overline{b}$ for some $a, b \in \mathbb{Z}$. We define

$$X + Y = \overline{a + b},$$

and

$$X \cdot Y = \overline{a \cdot b},$$

where the $+$ and $\cdot$ on the right denote addition and multiplication of integers.

Warning 23.11. It is important to notice that on the left-hand side of the equations in this definition, the $+$ and $\cdot$ symbols are *not* the usual addition and multiplication of integers. They are telling us how to add and multiply elements of $\mathbb{Z}_n$, which are sets rather than just numbers. The definition shows that these operations are closely related to addition and multiplication on the integers, but they are not the same!

We will illustrate this definition by working modulo 7 and modulo 11. In these examples, when we work in $\mathbb{Z}_n$ we will always write our results as $\overline{r}$ with $0 \leq r < n$, even though we may use numbers larger than n during the computations.

Example 23.12. Let $n = 7$. We perform the following computations in $\mathbb{Z}_7$:

$$\begin{aligned}
\overline{5} + \overline{6} &= \overline{5 + 6} = \overline{11} = \overline{4}, \\
\overline{1} + \overline{6} &= \overline{1 + 6} = \overline{7} = \overline{0}, \\
\overline{5} \cdot \overline{6} &= \overline{5 \cdot 6} = \overline{30} = \overline{2}, \\
\overline{3} \cdot \overline{5} &= \overline{3 \cdot 5} = \overline{15} = \overline{1}.
\end{aligned}$$

If we want to add two equivalence classes, we choose representatives, add the representatives, and then simplify (in this case modulo 7). △

Example 23.13. We perform the following calculations in $\mathbb{Z}_{11}$.

$$\begin{aligned}\overline{5}+\overline{6}&=\overline{5+6}=\overline{11}=\overline{0},\\ \overline{1}+\overline{6}&=\overline{1+6}=\overline{7},\\ \overline{5}\cdot\overline{6}&=\overline{5\cdot 6}=\overline{30}=\overline{8},\\ \overline{3}\cdot\overline{5}&=\overline{3\cdot 5}=\overline{15}=\overline{4}.\end{aligned}$$

If we want to add two equivalence classes, we choose representatives, add the representatives, and then simplify (in this case modulo 11). △

Warning 23.14. It is important to note that it is not obvious that the definitions of multiplication and addition in $\mathbb{Z}_n$ make sense. We have defined $X+Y$ for $X,Y\in\mathbb{Z}_n$ by choosing integers a,b with $X=\overline{a}$ and $Y=\overline{b}$, and defining $X+Y=\overline{a+b}$. However, other choices of integers would be possible. Suppose that we had chosen (possibly different) representative c and d with $X=\overline{c}$ and $Y=\overline{d}$. Then our definition would claim that $X+Y=\overline{c+d}$. If the result of addition depends on an arbitrary choice, then we do not have a good definition.

Fortunately, in this case, Theorem 23.8 comes to our rescue. If we have $X=\overline{a}=\overline{c}$ and $Y=\overline{b}=\overline{d}$, then we know that $a\equiv c\pmod n$ and $b\equiv d\pmod n$. Hence, $a+b\equiv c+d\pmod n$, by Theorem 23.8, so $\overline{a+b}=\overline{c+d}$. Hence, our choices made no difference in the definition of $X+Y$.

When we have a definition that appears to depend on arbitrary choices, but for which the arbitrary choices can be shown to make no difference in the definition, we say that the object being defined is *well-defined*. In other words, we have shown that addition on $\mathbb{Z}_n$ is well-defined.

Example 23.15. We write the complete addition and multiplication tables for $\mathbb{Z}_6$.

$+$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$	$\overline{0}$
$\overline{2}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$	$\overline{0}$	$\overline{1}$
$\overline{3}$	$\overline{3}$	$\overline{4}$	$\overline{5}$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{4}$	$\overline{4}$	$\overline{5}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$
$\overline{5}$	$\overline{5}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{4}$	$\overline{0}$	$\overline{2}$	$\overline{4}$
$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{0}$	$\overline{3}$	$\overline{0}$	$\overline{3}$
$\overline{4}$	$\overline{0}$	$\overline{4}$	$\overline{2}$	$\overline{0}$	$\overline{4}$	$\overline{2}$
$\overline{5}$	$\overline{0}$	$\overline{5}$	$\overline{4}$	$\overline{3}$	$\overline{2}$	$\overline{1}$

In these tables we wrote every entry using $\overline{0},\overline{1}\ldots,\overline{5}$. We can do this because $S=\{0,1,\ldots,5\}$ is a transversal for the equivalence classes. We could have used any other transversal such as $\{1,2,3,4,5,6\}$, $\{-2,-1,0,1,2,3\}$, or even $\{-10,-5,0,5,10,15\}$ (but it is best to keep things simple).

We notice some interesting facts about multiplication and addition in $\mathbb{Z}_n$ that are different from addition and multiplication in $\mathbb{Z}$.

If a is an integer and $a+a=0$, then a must equal 0. However, in $\mathbb{Z}_6$ we have $\overline{3}+\overline{3}=\overline{0}$ even though $\overline{3}\neq\overline{0}$.

If $a,b\in\mathbb{Z}$ with $a\cdot b=0$, then either $a=0$ or $b=0$. However, in $\mathbb{Z}_6$ we have $\overline{2}\cdot\overline{3}=\overline{0}$ even though neither $\overline{2}$ nor $\overline{3}$ is equal to $\overline{0}$. △

This example shows that we cannot take facts about addition or multiplication in $\mathbb{Z}_n$ for granted. We must prove those facts about addition and multiplication which we wish to use in $\mathbb{Z}_n$, rather than just assuming they hold. In the exercises you will prove some basic facts about operations in $\mathbb{Z}_n$ that do match our intuition from $\mathbb{Z}$. We demonstrate with an example.

Example 23.16. Let $n \in \mathbb{N}$. We will prove that for each $X \in \mathbb{Z}_n$, we have $X+\overline{0} = X$.

Let $X \in \mathbb{Z}_n$. Then for some $a \in \mathbb{Z}$, we have $X = \overline{a}$. (Note that we could take $0 \leq a < n$, but we do not need this extra information.) We find

$$X + \overline{0} = \overline{a} + \overline{0} = \overline{a+0} = \overline{a} = X. \qquad \triangle$$

23.D Exercises

Exercise 23.1. Let $n \in \mathbb{N}$ and $a \in \mathbb{Z}$. Prove that $0 \in \overline{a}$ if and only if $n \mid a$.

Exercise 23.2. Compute the following. Write the results as $\overline{r}$, with $r \in \mathbb{Z}$ nonnegative and as small as possible.

(a) $\overline{6} + \overline{7}$ in $\mathbb{Z}_9$.
(b) $\overline{6} \cdot \overline{7}$ in $\mathbb{Z}_9$.
(c) $\overline{59} \cdot \overline{119}$ in $\mathbb{Z}_{30}$.
(d) $\overline{6} \cdot \overline{5} + \overline{85}$ in $\mathbb{Z}_7$.
(e) $\overline{2}^{10}$ in $\mathbb{Z}_5$. (By $\overline{a}^n$ we mean $\underbrace{\overline{a} \cdot \overline{a} \cdots \overline{a}}_{n \text{ times}}$).

Exercise 23.3. Create addition and multiplication tables for $\mathbb{Z}_5$. Be sure to write each entry of the tables as one of $\overline{0}$, $\overline{1}$, $\overline{2}$, $\overline{3}$, or $\overline{4}$.

Exercise 23.4. Let $n \in \mathbb{N}$. Prove the following facts about addition and multiplication in $\mathbb{Z}_n$.

(a) For all $X, Y \in \mathbb{Z}_n$, $X + Y = Y + X$.
(b) For all $X, Y \in \mathbb{Z}_n$, $X \cdot Y = Y \cdot X$.
(c) For all $X \in \mathbb{Z}_n$, $X \cdot \overline{0} = \overline{0}$.
(d) For all $X \in \mathbb{Z}_n$, $X \cdot \overline{1} = X$.
(e) For all $X \in \mathbb{Z}_n$, $X \cdot \overline{2} = X + X$.
(f) For all $X \in \mathbb{Z}_n$, there is some $Y \in \mathbb{Z}_n$ such that $X + Y = \overline{0}$.
(g) For all $X, Y, Z \in \mathbb{Z}_n$, $(X + Y) \cdot Z = (X \cdot Z) + (Y \cdot Z)$.

Exercise 23.5. Demonstrate that for each $X \neq \overline{0}$ in $\mathbb{Z}_5$, there is some $Y \in \mathbb{Z}_5$ such that $X \cdot Y = \overline{1}$. (Hint: Use the multiplication table that you created in Exercise 23.3.)

Exercise 23.6. Is it true that for each $X \neq \overline{0}$ in $\mathbb{Z}_6$, there is some $Y \in \mathbb{Z}_6$ such that $X \cdot Y = \overline{1}$?

Exercise 23.7. In this exercise we generalize what was done in the previous two exercises.

(a) If $n \in \mathbb{N}$ is composite, prove that there are elements $\overline{a}, \overline{b} \in \mathbb{Z}_n$ with $\overline{a} \cdot \overline{b} = \overline{0}$ even though $\overline{a}, \overline{b} \neq \overline{0}$.

(b) If $n \in \mathbb{N}$ is prime, prove that given $\overline{a}, \overline{b} \in \mathbb{Z}_n$, if $\overline{a} \cdot \overline{b} = \overline{0}$, then $\overline{a} = \overline{0}$ or $\overline{b} = \overline{0}$. (Hint: Euclid's Lemma, Theorem 19.5.)

(c) If $n \in \mathbb{N}$ is prime, prove that for any nonzero $\overline{a} \in \mathbb{Z}_n$, there exists $\overline{b} \in \mathbb{Z}_n$ with $\overline{a} \cdot \overline{b} = \overline{1}$. (Hint: First prove that if $\overline{a} \neq \overline{0}$ then $\text{GCD}(a, n) = 1$. Then write 1 as a linear combination of a and n.)

Chapter VII

Functions

Greatness is not a function of circumstance. Greatness, it turns out, is largely a matter of conscious choice, and discipline. James C. Collins

Functions play an important role in mathematical applications to physics, engineering, biology, economics, etc. If you have taken algebra and calculus, then you evaluated functions, graphed functions, and determined derivatives and integrals of functions. However, the precise definition of what a function is may have been somewhat vague.

In this chapter we will make precise exactly what modern mathematicians mean by a function. The definition will allow us to study functions on sets much more general than real numbers. We will see examples of functions defined on the real numbers, the integers, the natural numbers, congruence classes, power sets, and other sets.

We will also discuss various properties of functions. The properties that we study have been chosen for their usefulness; they show up throughout higher mathematics.

As you progress in mathematics, you will find that an important tool in studying any class of mathematical objects is the study of functions between the objects. The definitions and techniques that we learn in this chapter will prepare you for this study.

24 Defining functions

The main purpose of this section is to formally define what we mean by a function. We will also give examples.

24.A What is a function?

Recall that we have defined a relation from a set A to a set B to be a subset of $A \times B$. We now define a function from A to B as a relation that satisfies certain conditions.

Definition 24.1. Let A and B be sets. A *function* from A to B is a relation f from A to B (i.e., a subset of $A \times B$), such that each $a \in A$ is a first coordinate of exactly one element of f. In other words

$$\forall a \in A, \exists! \, b \in B, \ (a, b) \in f.$$

We call A the *domain* of f, and we call B the *codomain* of f.

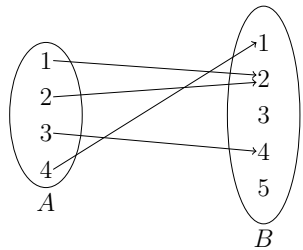
To specify that f is a function from A to B, we may write "Let $f \colon A \to B$ be a function."

Example 24.2. Let $A = \{1, 2, 3, 4\}$ and $B = \{1, 2, 3, 4, 5\}$. Define a relation f by

$$f = \{(1, 2), (2, 2), (3, 4), (4, 1)\}.$$

We see that f is a function, because each of the four elements of A is a first coordinate of exactly one ordered pair in f.

It is sometimes useful to visualize a function via a diagram. For instance, the function f described above could be visualized in the following diagram.



Here, we draw an arrow from $a \in A$ to $b \in B$ if $(a, b) \in f$. So there is an arrow from $1 \in A$ to $2 \in B$, since $(1, 2) \in f$.

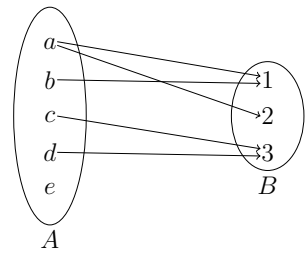
Note that since f is a function, we have exactly one arrow emanating from each element of the domain. There is no such restriction on the codomain; some elements of the codomain may have multiple arrows pointing at them and others may have no arrows pointing at them. △

Example 24.3. Let $A = \{a, b, c, d, e\}$ and $B = \{1, 2, 3\}$. Define a relation f by

$$f = \{(a, 1), (b, 1), (c, 3), (d, 3), (a, 2)\}.$$

This relation f fails to be a function for two reasons: First, the element $a \in A$ is the first coordinate of two different ordered pairs, namely $(a, 1)$ and $(a, 2)$. Second, the element $e \in A$ is not the first coordinate of any ordered pair in f.

We can see these failures if we draw a diagram of the relation f.



The element a has two arrows emanating from it; the element e has none. △

Since a function f from A to B is a relation from A to B, for $a \in A$ and $b \in B$ such that $(a, b) \in f$ we could write $a\, f\, b$. However, for functions we typically use a different notation.

Definition 24.4. Let $f\colon A \to B$ be a function. We write

$$f(a) = b$$

to mean that $(a, b) \in f$. In other words, $f(a)$ is the second coordinate of the unique ordered pair having a as its first coordinate.

When $f(a) = b$, we say that b is the *image* of a under the function f.

Example 24.5. Let $A = \{1, 2, 3, 4\}$ and $B = \{5, 6, 7, 8\}$. Let $f\colon A \to B$ be the function

$$f = \{(1, 5), (2, 6), (3, 7), (4, 8)\}.$$

Then we may write $f(1) = 5$, $f(2) = 6$, $f(3) = 7$, and $f(4) = 8$. We say that the image of 4 under f is 8. If it is understood that f is the only function under consideration, we could just say that the image of 4 is 8 (since there is no possibility of confusion about which function we mean). △

In many cases that we encounter, we will be able to define a function $f\colon A \to B$ by some kind of mathematical formula. If we wish to do this, we would want to give a rule that would tell us the value of $f(a)$ for any given element $a \in A$.

Example 24.6. Let $f\colon A \to B$ be the function defined in Example 24.5. Rather than defining f by listing all of its ordered pairs, we might define f as follows:

$$\text{For each } a \in A, \text{ we have } f(a) = a + 4.$$

This tells us that $f(1) = 1+4 = 5$, so $(1,5) \in f$. Also $f(2) = 2+4 = 6$, so $(2,6) \in f$, and $f(3) = 3+4 = 7$, so $(3,7) \in f$. Finally, $f(4) = 4+4 = 8$, so $(4,8) \in f$.

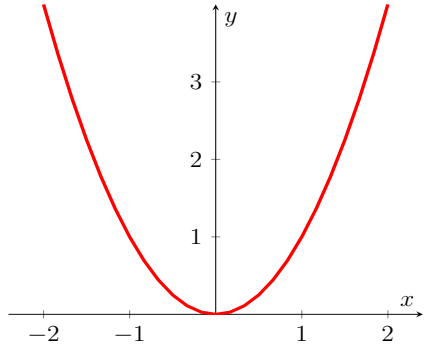
Another way to describe this function by a formula would be to say that

$$f = \{(a, a+4) : a \in A\}.$$ △

Example 24.7. Define a function $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$ for each $x \in \mathbb{R}$. We may find the images of various real numbers under this function; $f(2) = 4$, $f(\sqrt{3}) = 3$, and so on. Note that for any real number x, the output x^2 is also a real number, so the function does in fact go from $\mathbb{R}$ to $\mathbb{R}$. Note that as a set

$$f = \{(x, x^2) : x \in \mathbb{R}\},$$

and we graph this set below; the graph is called the *graph of the function*.



△

Warning 24.8. When defining a function $f\colon A \to B$ by a formula, as above, it is *very* important to verify that for each element of A, the output of the formula is actually an element of B. If even one of the values is not an element of B, then f is not a function from A to B.

Example 24.9. Suppose we try to define a function $f\colon \mathbb{R} \to \mathbb{Z}$ by the formula $f(x) = x^2$ for all $x \in \mathbb{R}$. We have exactly the same rule as in the previous example, but now, for many real numbers x, the image of x under f is not in $\mathbb{Z}$. For instance, $f(0.5) = 0.25 \notin \mathbb{Z}$. Hence, f is not a function from $\mathbb{R}$ to $\mathbb{Z}$. △

Sometimes we will wish to determine if two functions are equal. For this, we will use the following definition.

Definition 24.10. Two functions f and g are *equal* if they have the same domain, the same codomain, and they are equal as sets of ordered pairs.

Example 24.11. If we define $f\colon \mathbb{Z} \to \mathbb{Z}$ by $f(x) = 2x$, and we define $g\colon \mathbb{Z} \to \mathbb{R}$ by $g(x) = 2x$, then f and g consist of the same set of ordered pairs (namely, the set $\{(x, 2x) : x \in \mathbb{Z}\}$), but they have different codomains, so they are not the same function. △

Example 24.12. If we define $f\colon \mathbb{Z} \to \mathbb{Z}$ by $f(x) = 2x + 2$, and we define $g\colon \mathbb{Z} \to \mathbb{Z}$ by $g(x) = 2(x + 1)$, then is it the case that these functions are equal? Yes. First, they have the same domain, $\mathbb{Z}$. Second, they have the same codomain, $\mathbb{Z}$. Third, they have the same ordered pairs, since $f(x) = 2x + 2 = 2(x + 1) = g(x)$ for each $x \in \mathbb{Z}$. △

24.B Piecewise defined functions

The rule or formula defining a function need not be a simple one-line mathematical formula. Sometimes we will wish to define a function $f\colon A \to B$ by one formula on one part of A and by another formula on another part of A. For instance, the absolute value function is defined on $\mathbb{R}$ as

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

This is a *piecewise defined function.*

We will make this precise in the following theorem, and give other examples afterward.

Theorem 24.13. *Let A and B be sets, and suppose that $\{P, Q\}$ is a partition of A with two parts. If we are given a function $g\colon P \to B$ and a function $h\colon Q \to B$, then $f = g \cup h$ defines a function $f\colon A \to B$.*

Remark 24.14. The rule for f is

$$f(a) = \begin{cases} g(a) & \text{if } a \in P, \\ h(a) & \text{if } a \in Q. \end{cases}$$

See Exercise 24.7 for a more general version of this theorem. ▲

Proof. We may consider g as a subset of $P \times B$, which is a subset of $A \times B$, and h as a subset of $Q \times B$, which is a subset of $A \times B$. Hence, both g and h (considered as collections of ordered pairs) are subsets of $A \times B$. In fact, $g = \{(x, g(x)) : x \in P\}$ and $h = \{(x, h(x)) : x \in Q\}$. Let

$$f = g \cup h.$$

Then f is a subset of $A \times B$, so f is a relation from A to B. To see that f is a function, it suffices to check that each element of A is the first coordinate of exactly one ordered pair in f.

To that end, let $a \in A$. Since A is partitioned into two parts, there are two cases; either $a \in P$ or $a \in Q$. Without loss of generality, we may assume $a \in P$. We know that a is the first coordinate of exactly one ordered pair in g, namely $(a, g(a))$ (using the fact that $g\colon P \to B$ is a function). On the other hand, a is not the first coordinate of any ordered pairs in h (since every ordered pair in h has first coordinate in Q, and $a \notin Q$). Hence, $(a, g(a))$ is the one and only ordered pair in f that has first coordinate a. Since a was an arbitrary element of A, we see that f is a function. □

The following example involves finite sets. You might try drawing the arrows yourself to check that the diagram in the example is correct.

Example 24.15. Let $A = P \cup Q$, where $P = \{1, 2, 3\}$ and $Q = \{4, 5, 6\}$, and let $B = \{1, 2, 3\}$. Then $\{P, Q\}$ is a partition of A. If we define $g\colon P \to B$ by $g(x) = 4 - x$, and we define $h\colon Q \to B$ by $h(x) = x - 3$, we see that

$$g = \{(1, 3), (2, 2), (3, 1)\}$$

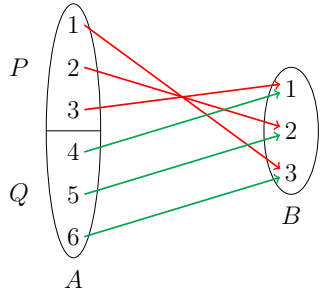
is a function from P to B and

$$h = \{(4, 1), (5, 2), (6, 3)\}$$

is a function from Q to B. By the theorem we may define a function $f\colon A \to B$ by the rule

$$f(a) = \begin{cases} 4 - a & \text{if } a \in \{1, 2, 3\}, \\ a - 3 & \text{if } a \in \{4, 5, 6\}. \end{cases}$$

Graphically, we have the following diagram, where the top three arrows indicate the function g, and the bottom three arrows are h.



As a collection of ordered pairs, we see that f is given by

$$f = g \cup h = \{(1, 3), (2, 2), (3, 1), (4, 1), (5, 2), (6, 3)\}. \qquad \triangle$$

Example 24.16. Question: What is wrong with the following piecewise defined function? Try to define $f\colon \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} x^2 + 2 & \text{if } x \geq 0, \\ -1 & \text{if } x \leq 0. \end{cases}$$

Answer: It has two different values at 0, and hence is not a function. The problem is that the two domains "$x \geq 0$" and "$x \leq 0$" are not a partition of $\mathbb{R}$ (they overlap). $\triangle$

Warning 24.17. Typically, we don't need to explicitly state that conditions given to define a piecewise defined function will partition the domain (usually because this fact is obvious). We must, however, be careful that the conditions really do define a partition.

In the following example, the conditions in the piecewise defined function implicitly define a partition on the domain. We have no need to explicitly name the parts of the partition.

Example 24.18. We can define a function $f\colon \mathbb{Z} \to \mathbb{Z}$ by

$$f(x) = \begin{cases} 3x + 1 & \text{if } x \text{ is odd,} \\ \dfrac{x}{2} & \text{if } x \text{ is even.} \end{cases}$$

Note that when x is even then $x/2 \in \mathbb{Z}$, so this really does give a function with codomain $\mathbb{Z}$. We have, for instance, $f(1) = 4$, $f(2) = 1$, and so forth. △

It is easy to generalize Theorem 24.13 to partitions of A with more than two parts. You will do this in Exercise 24.7. For now, we give an example.

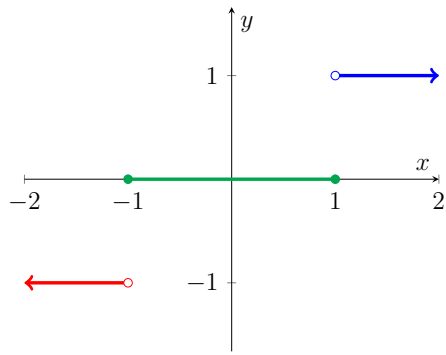
Example 24.19. We can define a function $f\colon \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} -1 & \text{if } x < -1 \\ 0 & \text{if } -1 \leq x \leq 1 \\ 1 & \text{if } 1 < x \end{cases}$$

Note that the conditions on x partition $\mathbb{R}$. The partition is

$$\{\{x \in \mathbb{R} : x < -1\}, \{x \in \mathbb{R} : -1 \leq x \leq 1\}, \{x \in \mathbb{R} : 1 < x\}\}.$$

We graph this function below, with the portions given by the separate conditions in different colors.



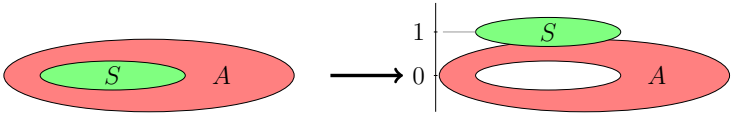
△

To end this subsection, we define a special kind of (piecewise defined) function, which is useful in many parts of mathematics.

Definition 24.20. Let A be a set and let S be a subset of A. The *characteristic function* of S, denoted $\chi_S \colon A \to \{0,1\}$, is the function defined by the rule

$$\chi_S(a) = \begin{cases} 1 & \text{if } a \in S \\ 0 & \text{if } a \notin S \end{cases}$$

We can view this situation, as depicted below.

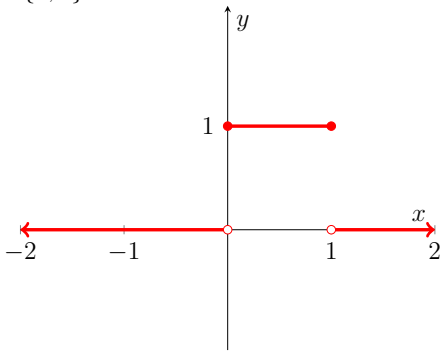


The elements in S are mapped to 1, and everything else is mapped to 0. The function χ_S indicates (by the 1 or 0 output, thought of as a "yes/no") whether an element belongs to S or not.

Advice 24.21. It can be useful to think of the characteristic function as a metal detector. It beeps (gives an output of 1) when you are near elements of S (the metal), but is silent (gives an output of 0) outside S.

Notice that the notation χ_S does not indicate the domain of the characteristic function. The domain is typically either understood from context or explicitly stated, for instance by saying "$\chi_S \colon A \to \{0,1\}$".

Example 24.22. Let $A = \mathbb{R}$ and let $S = [0,1]$ be the closed interval from 0 to 1. The graph of $\chi_S \colon \mathbb{R} \to \{0,1\}$ is



△

24.C Well-defined functions

Sometimes when defining a function by a rule, the rule depends upon a choice we make. The following is an example.

Example 24.23. Define a function $f \colon \mathbb{Z}_5 \to \mathbb{Z}_5$ by

$$f(\overline{a}) = \overline{a^2}.$$

Where are we making a choice? We choose a representative $a \in \mathbb{Z}$ for the congruence class $\overline{a}$. Thus, our rule really is: choose a representative for the class, square the representative, and then take the class corresponding to this square.

For f to be a function, we must make sure that the output does *not* depend on the choice of representative that we make. For instance, the congruence class $\overline{3}$ is the same as the congruence class $\overline{8}$. Do we get the same output using the representative 3 as we do with 8? Yes, $\overline{3^2} = \overline{9} = \overline{4}$, and similarly $\overline{8^2} = \overline{64} = \overline{4}$.

We will now show that this works in general. Let $a, b \in \mathbb{Z}$, and assume that $\overline{a} = \overline{b}$. We then know that $a \equiv b \pmod 5$. Multiplying this equation by itself (using Theorem 23.8(2)), we find that $a^2 \equiv b^2 \pmod 5$, so $\overline{a^2} = \overline{b^2}$. Hence, we see that

$$f(\overline{a}) = \overline{a^2} = \overline{b^2} = f(\overline{b}).$$

As a set of ordered pairs, we may write f as

$$f = \{(\overline{0}, \overline{0}), (\overline{1}, \overline{1}), (\overline{2}, \overline{4}), (\overline{3}, \overline{4}), (\overline{4}, \overline{1})\}.$$

Note that in order for the definition of f to make sense, it is important that the ordered pair $(\overline{8}, \overline{64})$ is equal to the pair $(\overline{3}, \overline{4})$, since their first coordinates are equal. △

We now give two examples where we try to define a function by a rule, but the rule does depend on a choice we make, and thus does not define a function.

Example 24.24. Let's *try* to define the "numerator function" $f\colon \mathbb{Q} \to \mathbb{Z}$ by the rule $f(n/d) = n$. In other words, this rule tells us that if we plug in a fraction like $1/2$, we should get as an output the integer 1.

Unfortunately this is not a function, because we can write the same fraction with two different numerators. For instance $1/2 = 2/4$, however we have $f(1/2) = 1$ and $f(2/4) = 2$, thus $f(1/2) \neq f(2/4)$.

(The modified rule "take the numerator of a fraction n/d, when the fraction is in lowest terms with $d > 0$" does yield a function, because there are not two different ways of writing a fraction in lowest terms.) △

Example 24.25. Suppose we try to define a function $f\colon \mathbb{Z}_3 \to \mathbb{Z}$ by

$$f(\overline{a}) = a.$$

We find the following ordered pairs in f:

$$f = \{(\overline{0}, 0), (\overline{1}, 1), (\overline{2}, 2), (\overline{3}, 3), (\overline{4}, 4), \ldots\}.$$

Note, however, that $\overline{0} = \overline{3}$ (since we are in $\mathbb{Z}_3$), so f possesses two *distinct* ordered pairs, $(\overline{0}, 0)$ and $(\overline{0}, 3)$ that both have $\overline{0}$ as a first coordinate. We say that this f is *not a well-defined function.* △

Definition 24.26. Suppose that we try to define a function $f\colon A \to B$ by a rule. If the rule produces multiple images for a single element of A, then we say that f is *not a well-defined function.* If the rule for f produces a single image for each element of A, then we say that f *is a well-defined function.*

Advice 24.27. When given a rule, it is usually unnecessary to check that it produces a well-defined function. Most mathematical operations produce a fixed result for a given input.

The key indicator which determines whether you should check that a function is well-defined is that the elements of the domain can be written in different ways when plugged into the rule, such as when working with equivalence classes.

Example 24.28. Here is another example of a rule that does not yield a well-defined function. Look for how elements in the domain can be represented in multiple ways.

Suppose a professor wishes to assign grades to his class in a nontraditional way. He takes $X = \{\text{students in the class}\}$ and $Y = \{A, B, C, D, F\}$, and decides to assign grades by a rule $g\colon X \to Y$ given by

$$g(\text{student}) = \begin{cases} \text{The first letter of the student's name} & \text{if the result is in } \{A, B, C, D\}, \\ F & \text{otherwise.} \end{cases}$$

How would students react to this? Amelia Andrews would be overjoyed. Robert Smith would likely say "Call me Bob." What grade would John Adams receive? Since the professor has not made clear whether the grade is determined by the first or the last name, John could make a case that he should receive either an A or an F (he would probably argue for the former).

In this case, the proposed function that the professor wishes to use is not well-defined. It depends on a choice of which name to use for the student. △

Warning 24.29. Note that if a proposed rule for a function f does not produce a well-defined function, we would be incorrect to say that "the function f is not well-defined," since f is, in fact, not a function. We should instead say "the rule defining f does not produce a well-defined function," "the proposed function is not well-defined," "f is not a well-defined function," or something similar (as long as we do not call f a function).

Example 24.30. In this example we will be working both modulo 6 and modulo 3, so we need different notations for the different equivalence classes. As usual, we will denote congruence classes modulo 6 by $\overline{a}$, where $a \in \mathbb{Z}$. For this example, we will denote congruence classes modulo 3 by $[a]$, where $a \in \mathbb{Z}$. If we define

$$f\colon \mathbb{Z}_3 \to \mathbb{Z}_6$$

by $f([a]) = \overline{a}$ and

$$g\colon \mathbb{Z}_6 \to \mathbb{Z}_3$$

by $g(\overline{a}) = [a]$, one of f and g is a function and the other is not a function. Which is which?

We first prove that g is a well-defined function. Let $a, b \in \mathbb{Z}$, and assume that $\overline{a} = \overline{b}$, so that $a \equiv b \pmod 6$. Hence, $a - b = 6k = 3(2k)$ for some $k \in \mathbb{Z}$. This implies that $a \equiv b \pmod 3$, so $[a] = [b]$. Therefore, $g(\overline{a}) = [a] = [b] = g(\overline{b})$, and we see that g is well-defined. Its definition does not depend on the representative we take for the congruence class.

Next we prove that f is not a well-defined function. Since $1 \equiv 4 \pmod 3$, we have that $[1] = [4]$. However, using the rule above gives $f([1]) = \overline{1} \neq \overline{4} = f([4])$. Hence, the output of f *does* depend on the representative that we take for $[1]$. Therefore, f is *not* well-defined. △

24.D Exercises

Exercise 24.1. Which of the following relations are functions from the set $A = \{1, 2, 3, 4\}$ to the set $B = \{1, 2, 3, 4, 5\}$?

(a) $f_1 = \{(1, 3), (2, 3), (3, 3), (4, 3)\}$.
(b) $f_2 = \{(1, 2), (2, 3), (3, 5), (4, 6)\}$.
(c) $f_3 = \{(1, 2), (2, 3), (2, 4), (4, 5)\}$.
(d) $f_4 = \{(1, 2), (1, 3), (2, 3), (3, 4), (4, 1)\}$.
(e) $f_5 = \{(1, 2), (2, 3), (4, 5)\}$.
(f) $f_6 = \{(1, 2), (1, 2), (2, 3), (3, 4), (4, 1)\}$. (Hint: Do repetitions matter?)

Exercise 24.2. In some textbooks it is claimed that $f\colon \mathbb{R} \to \mathbb{R}$ given by the rule $f(x) = 1/x$ is a function. Technically this is incorrect because the rule is not defined at $x = 0$, and so the largest available domain should be $\mathbb{R} - \{0\}$. Find the largest domain for the following functions (supposedly defined from $\mathbb{R}$ to $\mathbb{R}$):

(a) $f(x) = \sin(x)$.
(b) $g(x) = \tan(x)$.
(c) $h(x) = \ln(x)$.
(d) $p(x) = \sqrt{1 - x}$.
(e) $q(x) = \sqrt[3]{x}/(2 + x)$.

Exercise 24.3. Let A be a finite set and let B be any set. Let $f\colon A \to B$ be a function. Considering f as a set of ordered pairs, prove that $|f| = |A|$.

Exercise 24.4. For $a \in \mathbb{Z}$, denote the congruence class of a modulo 8 by $\overline{a}$, and the congruence class of a modulo 4 by $[a]$. Determine which of the following definitions give well-defined functions. For those that are well-defined, give a proof. For those that are not well-defined, give an example to demonstrate this fact.

(a) Define $f\colon \mathbb{Z}_8 \to \mathbb{Z}_4$ by $f(\overline{a}) = [a]$.
(b) Define $g\colon \mathbb{Z}_4 \to \mathbb{Z}_8$ by $g([a]) = \overline{a}$.
(c) Define $h\colon \mathbb{Z}_4 \to \mathbb{Z}_8$ by $h([a]) = \overline{2a}$.
(d) Define $j\colon \mathbb{Z}_4 \to \mathbb{Z}_8$ by $j([a]) = \overline{3a}$.

Exercise 24.5. Define $f\colon \mathbb{Z}_5 \to \mathbb{Z}_5$ by $f(\overline{a}) = \overline{a}^5$. Define $g\colon \mathbb{Z}_5 \to \mathbb{Z}_5$ by $g(\overline{a}) = \overline{a}$. You may assume that f and g are both well-defined (which they are). Are f and g equal? (Hint: Both f and g are collections of exactly five ordered pairs.)

Exercise 24.6. Let A be a set, and let S_1, S_2 be subsets of A. Given an arbitrary $a \in A$, prove the following facts for characteristic functions defined on the domain A:

(a) If $T = S_1 \cup S_2$ and $S_1 \cap S_2 = \emptyset$, then $\chi_T(a) = \chi_{S_1}(a) + \chi_{S_2}(a)$.

(b) If $T = S_1 \cap S_2$, then $\chi_T(a) = \chi_{S_1}(a) \cdot \chi_{S_2}(a)$.

(c) If $T = S_1 \cup S_2$, then $\chi_T(a) = \chi_{S_1}(a) + \chi_{S_2}(a) - \chi_{S_1}(a) \cdot \chi_{S_2}(a)$.

(Hint: Cases.)

Exercise 24.7. Let A and B be sets. Let I be an indexing set, and let $P = \{P_i : i \in I\}$ be an arbitrary partition of A. For each $i \in I$, let $f_i \colon P_i \to B$ be a function. Prove that the relation

$$f = \bigcup_{i \in I} f_i$$

is a function from A to B and that the rule for f (as a piecewise defined function) is

$$f(a) = f_i(a) \text{ when } a \in P_i \text{ for some } i \in I.$$

(Hint: Modify the proof of Theorem 24.13.)

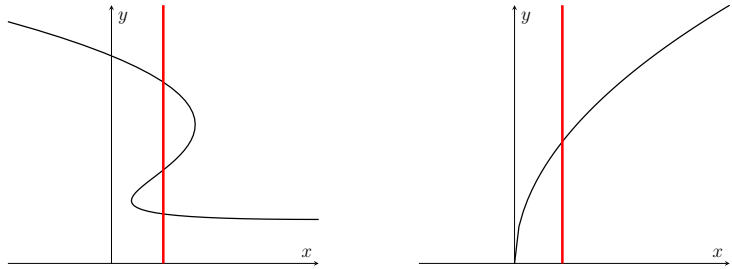
25 Injective and surjective functions

Although a function from A to B consists of ordered pairs with the condition that each element $a \in A$ is the first coordinate of exactly one ordered pair, there is no such constraint on elements of B. As we have seen, it is possible that an element of B can be the second coordinate of several ordered pairs in a function; it can also be the second coordinate of no ordered pairs.

We will now discuss two conditions on functions that together guarantee that every element of B is the second coordinate of exactly one ordered pair. Even though the strongest results occur when both conditions hold, each condition separately is important enough to study on its own.

25.A Injective functions

Given some curve in the plane (such as in the two pictures below), in order for it to be the graph of a function it must pass the *vertical line test*: every vertical line hits the curve in at most one point. Thus, the curve on the left does not represent a function, while the curve on the right passes the test.



Thus, the vertical line test checks that each x-input yields at most one y-output. If we turn this around, and ask which functions pass the *horizontal line test* (i.e., no output comes from two different inputs) we get the following definition:

Definition 25.1. Let $f\colon A \to B$ be a function. We say f is *injective* if distinct elements of A are mapped by f to distinct elements of B. In symbols,

$$\forall a_1, a_2 \in A,\ a_1 \neq a_2 \Rightarrow f(a_1) \neq f(a_2).$$

In order to avoid working with inequalities, one often instead uses the contrapositive

$$\forall a_1, a_2 \in A,\ f(a_1) = f(a_2) \Rightarrow a_1 = a_2. \tag{25.2}$$

Advice 25.3. Students should memorize (25.2), as that is the formulation of injectivity commonly used in proofs. Injectivity will be usually accomplished by doing a direct proof using (25.2), although on occasion other methods can be utilized.

Injective functions are also called "one-to-one" (or 1-1) functions, meaning each element of B is the image of at most one element from A. In other words, for each $b \in B$ either there is a unique $a \in A$ such that $f(a) = b$, or there is no element in A that maps to B. This is in contrast to functions that might take two, three, or more elements of A to one fixed element of B. In other words, a function $f\colon A \to B$ is injective exactly when:

(25.4) Each $b \in B$ is the second coordinate of at most one ordered pair in f.

The name "one-to-one" does not refer to the fact that the function takes each individual element of A to one element of B (as opposed to multiple elements of B); this is a property that all functions have (corresponding to the vertical line test).

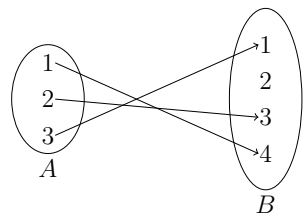
The next example illustrates what injectivity looks like for functions between finite sets.

Example 25.5. Let $A = \{1, 2, 3\}$ and let $B = \{1, 2, 3, 4\}$. Define $f\colon A \to B$ by

$$f = \{(1, 4), (2, 3), (3, 1)\}.$$

Clearly, distinct elements of A map to distinct elements in B. In other words, 1, 2, and 3 each go to different places. Hence, f is injective by checking (25.4) in every case.

Thinking of f pictorially, we get

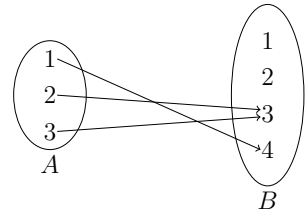


In terms of the diagram, we can think of *injectivity* as meaning that there is never more than one arrow pointing at an element of B. △

Example 25.6. Let $A = \{1, 2, 3\}$ and let $B = \{1, 2, 3, 4\}$. Define $f\colon A \to B$ by

$$f = \{(1, 4), (2, 3), (3, 3)\}.$$

Thinking of f pictorially, we get



We see from the diagram that $f(2) = f(3)$ even though $2 \neq 3$. Thus, this function is not injective. △

In most of the following examples we take Advice 25.3 to heart and prove injectivity by showing that (25.2) holds.

Example 25.7. Let $f\colon \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = 2x + 1$. Notice that f passes the horizontal line test, so it should be injective. We will prove that claim.

Let $x_1, x_2 \in \mathbb{R}$. Assume that $f(x_1) = f(x_2)$. Then $2x_1 + 1 = 2x_2 + 1$. Subtracting 1 and dividing by 2, we obtain $x_1 = x_2$. Hence, f is injective. △

Example 25.8. Let $f\colon \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$. This function fails the horizontal line test, so it should not be injective. Indeed, $f(1) = f(-1)$, but $1 \neq -1$. Hence, f is not injective. △

Example 25.9. Let $f\colon [0, \infty) \to \mathbb{R}$ be defined by $f(x) = x^2$. We will show that f is injective. (Notice that it passes the horizontal line test, since it is only the right half of the parabola.)

Let $x_1, x_2 \in [0, \infty)$. Assume $f(x_1) = f(x_2)$, so $x_1^2 = x_2^2$. Hence $x_1^2 - x_2^2 = 0$, and factoring we have

$$(x_1 - x_2)(x_1 + x_2) = 0.$$

One of the two factors must be zero, and thus there are two cases to consider.

Case 1: $x_1 + x_2 = 0$. In this case $x_1 = -x_2$. But since x_1 and x_2 are nonnegative, the only way this can happen is if $x_1 = x_2 = 0$.

Case 2: $x_1 - x_2 = 0$. Therefore, $x_1 = x_2$.

In both cases $x_1 = x_2$, and thus we have shown that f is injective. △

The previous two examples show that injectivity depends as much on the domain as on the rule used to define the function.

Example 25.10. Let $f\colon \mathbb{R} - \{2\} \to \mathbb{R} - \{1\}$ be defined by

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

First, notice that f is indeed a function with the given domain and codomain. For $a \in \mathbb{R} - \{2\}$, we see that $f(a)$ is a real number. If $f(a)$ were equal to 1, we would have

$$\frac{a - 1}{a - 2} = 1,$$

so that $a - 1 = a - 2$, and hence $-1 = -2$. This is not possible, so $f(a) \in \mathbb{R} - \{1\}$.

Now, we prove injectivity. Let $x_1, x_2 \in \mathbb{R} - \{2\}$ and suppose that $f(x_1) = f(x_2)$. We have

$$\frac{x_1 - 1}{x_1 - 2} = \frac{x_2 - 1}{x_2 - 2}.$$

Cross-multiplying, we find

$$(x_1 - 1)(x_2 - 2) = (x_2 - 1)(x_1 - 2).$$

Expanding both sides, we obtain

$$x_1x_2 - 2x_1 - x_2 + 2 = x_1x_2 - x_1 - 2x_2 + 2$$

Cancelling equal terms on both sides, and adding $2x_1 + 2x_2$ to both sides yields

$$x_2 = x_1.$$

Hence, f is injective. △

Example 25.11. Define a function $f\colon \mathbb{Z} \to \mathbb{Z}$ by

$$f(n) = \begin{cases} n + 1 & \text{if } n \geq 0, \\ n & \text{if } n < 0. \end{cases}$$

We will show that f is injective.

Working contrapositively, let $a, b \in \mathbb{Z}$ and assume $a \neq b$. We note that $a+1 \neq b+1$. There are now three possible cases.

Case 1: Assume $a \geq 0$ and $b \geq 0$. Then $f(a) = a + 1 \neq b + 1 = f(b)$.

Case 2: Assume $a < 0$ and $b < 0$. Then $f(a) = a \neq b = f(b)$.

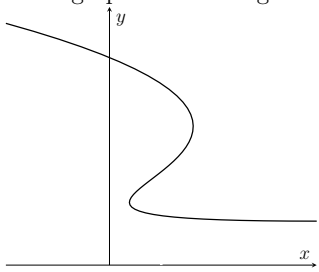
Case 3: Assume (without loss of generality) that $a < 0$ and $b \geq 0$. Then $f(a) = a < 0$ and $f(b) = b + 1 > 0$, so $f(a) \neq f(b)$.

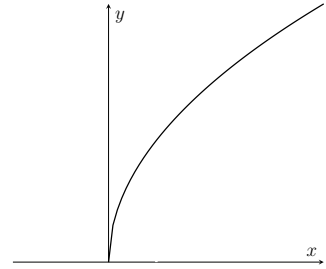
In all cases we have shown $f(a) \neq f(b)$. Hence, f is injective. △

Remark 25.12. Proving that a piecewise defined function is injective will almost always involve a proof by cases. ▲

25.B Surjective functions

Recall the two graphs of curves given in the previous subsection:





The curve on the left is not a function because it fails the vertical line test. If we are asking for a function from $\mathbb{R}$ to $\mathbb{R}$, then the curve on the right *also* fails to be a function because it is not defined everywhere. In other words, some vertical lines do not hit the curve; there are "holes" for all negative values of x.

Thus, any function $f\colon A \to B$ satisfies two basic properties:

- (Vertical line test): For each input, there is at most one output.
- (No holes in the domain): For each input, there is at least one output.

Putting these two properties together we get exactly the definition of a function (for each element of the domain, there is exactly one output in the codomain).

In the previous subsection we studied what happens if we reverse the roles of the domain and codomain, turning the vertical line test into the horizontal line test. Here, we study the same reversal, but with the "no holes in the domain" rule changed to the "no holes in the codomain."

Definition 25.13. A function $f\colon A \to B$ is *surjective* if every element of B occurs (at least once) as the image under f of some $a \in A$. In symbols,

$$\forall b \in B, \exists a \in A,\ f(a) = b. \tag{25.14}$$

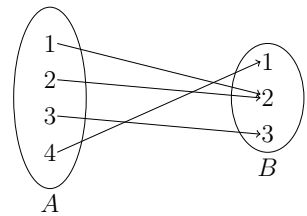
Surjective functions are sometimes called "onto" (meaning, they map onto *all* elements of the codomain). In terms of ordered pairs, surjectivity of a function $f\colon A \to B$ means:

(25.15) Each $b \in B$ is a second coordinate of at least one ordered pair in f.

Example 25.16. Let $A = \{1, 2, 3, 4\}$ and let $B = \{1, 2, 3\}$. Define a function $f\colon A \to B$ by

$$f = \{(1, 2), (2, 2), (3, 3), (4, 1)\}.$$

We note that each element $1, 2, 3 \in B$ appears at least once as the second coordinate of an ordered pair in f (in fact, 2 appears twice). Visualizing f as a diagram, we have



In terms of the diagram, we can think of *surjectivity* as meaning that there is always at least one arrow pointing at any element of B. △

In order to prove that $f\colon A \to B$ is surjective we normally verify (25.14). We start by letting $b \in B$ be an arbitrary element of B. We then need to somehow use b to prove the existence of an element $a \in A$ such that $f(a) = b$. We will demonstrate how this is to be done with many examples.

Example 25.17. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function defined by $f(x) = 2x + 1$. If we graph this function, each horizontal line crosses the graph, so there are no holes in the codomain. Thus, we should believe that f is surjective. We will now prove it.

Let $b \in \mathbb{R}$, and fix

$$a = \frac{b-1}{2}.$$

Note that $a \in \mathbb{R}$ (the domain). We compute

$$f(a) = f\left(\frac{b-1}{2}\right) = 2\left(\frac{b-1}{2}\right) + 1 = (b-1) + 1 = b.$$

Hence, f is surjective. △

Remark 25.18. This proof may be unsatisfying because, like many existence proofs, it clearly shows that a exists with $f(a) = b$, but it does so without explaining how a might be found. This is due to the fact that the correct a was found using scratchwork, but that work was not included in the proof. We will do that scratchwork now.

Given $b \in \mathbb{R}$, we wish to find a such that $f(a) = b$. Hence, we wish to solve

$$2a + 1 = b.$$

Subtracting 1, we see that $2a = b - 1$. Dividing by 2 yields

$$a = \frac{b-1}{2}.$$

This gives the desired element.

Notice that the preceding paragraph derives the desired a. However, the work done there is *not* the proof of surjectivity. To do the proof we need to check two things. First, that this a we found in our scratchwork belongs to the domain. Second, that $f(a) = b$. You should do the work of calculating the desired a on scratch paper and afterwards write out the proof. ▲

Example 25.19. Define $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. Does this function pass the "no holes in the codomain" test? No, we see that all negative values are missed. We will now formally prove it is not surjective. Fixing $b = -1$, then for any $a \in \mathbb{R}$ we have $f(a) = a^2 \neq -1 = b$. Hence, f is not surjective. △

Example 25.20. Define $f\colon \mathbb{R} \to [0, \infty)$ by $f(x) = x^2$. Let $b \in [0, \infty)$. Since b is nonnegative it has a square root, so fix $a = \sqrt{b}$. Clearly a is real, hence it belongs to the domain. Now,

$$f(a) = f(\sqrt{b}) = (\sqrt{b})^2 = b.$$

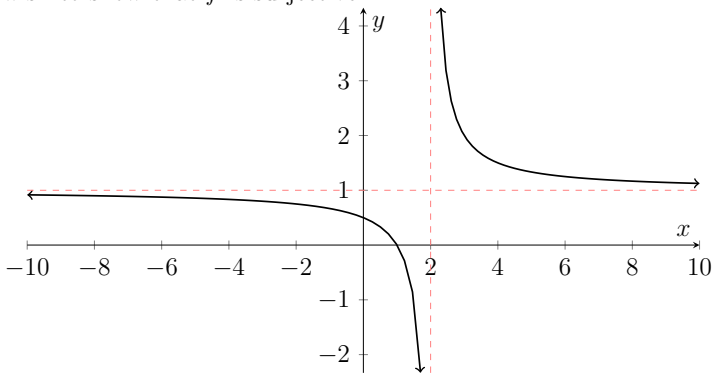
Since b was an arbitrary element of the codomain, f is surjective. △

The previous two examples show that surjectivity can depend as much on the codomain as on the rule used to define the function.

Example 25.21. Let $f\colon \mathbb{R}-\{2\} \to \mathbb{R}-\{1\}$ be defined by

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

We wish to show that f is surjective.



(Note: This function only has a "hole" in the codomain if our codomain is $\mathbb{R}$. We remove that hole by limiting the codomain to $\mathbb{R}-\{1\}$.)

Let $b \in \mathbb{R}-\{1\}$. Fix

$$a = \frac{2b-1}{b-1}.$$

We need to show that a belongs to the domain $\mathbb{R}-\{2\}$. Clearly $a \in \mathbb{R}$ (since $b \neq 1$). Assuming, by way of contradiction, that $a = 2$, we would have $2b-1 = 2(b-1)$. Simplifying, we get $-1 = -2$, a contradiction. Hence, $a \in \mathbb{R}-\{2\}$.

Now

$$f(a) = \frac{\dfrac{2b-1}{b-1} - 1}{\dfrac{2b-1}{b-1} - 2} = \frac{2b-1-(b-1)}{2b-1-2(b-1)} = \frac{b}{1} = b.$$

Hence, f is surjective. △

Remark 25.22. Again, this proof of surjectivity is quite unsatisfying. It gives no idea how a was found. Note that we wish to find a such that

$$f(a) = \frac{a-1}{a-2} = b.$$

Clearing denominators by multiplying by $a-2$, we have $a-1 = b(a-2)$, or in other words $a-1 = ab-2b$. Hence, we want $2b-1 = ab-a = a(b-1)$. Dividing by $b-1$, we see that we want

$$a = \frac{2b-1}{b-1}.$$

This calculation is an important step in finding the correct a, but is not a necessary part of the written proof. ▲

Example 25.23. Let $S = \{1, 2, 3, 4, 5\}$, and let $A = \mathscr{P}(S) - \{\emptyset\}$. Define a function $f\colon A \to S$ by

$$f(X) = \text{the least element of } X,$$

for every $X \in A$. You may check directly that f is a function from A to S (for instance $f(\{3, 4, 5\}) = 3$ and $f(\{2, 4\}) = 2$). We will now prove that f is surjective.

Let $b \in S$. Fix $a = \{b\}$. Since $\{b\} \in \mathscr{P}(S)$ and $\{b\} \neq \emptyset$ we have that $a \in A$. Clearly, the least element of $\{b\}$ is b. Hence, $f(a) = f(\{b\}) = b$. This proves that f is surjective. △

Example 25.24. Let $f\colon \mathbb{Z} \to \mathbb{Z}$ be defined by

$$f(n) = \begin{cases} n & \text{if } n \geq 0, \\ n+1 & \text{if } n < 0. \end{cases}$$

We will show that f is surjective.

Let $b \in \mathbb{Z}$. We consider two cases:

Case 1. If $b \geq 0$, then fixing $a = b$ we have $a \in \mathbb{Z}$ and $f(a) = f(b) = b$.

Case 2. If $b < 0$, then fixing $a = b - 1 < 0$ we have $a \in \mathbb{Z}$ and $f(a) = f(b-1) = b$.

Hence, f is surjective. (Note that $f(-1) = f(0) = 0$, so f is not injective.) △

Remark 25.25. As with injectivity, proving the surjectivity of a piecewise defined function will generally involve proof by cases. ▲

25.C The range of a function

A useful concept in studying surjectivity is the set of outputs of a function.

Definition 25.26. Let A and B be sets, and let $f\colon A \to B$ be a function. Define

$$\operatorname{im}(f) = \{f(a) \,:\, a \in A\}.$$

We call $\operatorname{im}(f)$ the *range* or the *image* of f. Notice that $\operatorname{im}(f) \subseteq B$.

Example 25.27. Let $f\colon \{1, 2, 3, 4\} \to \{a, b, c, d\}$ be given as a set of ordered pairs by

$$f = \{(1, a), (2, a), (3, b), (4, c)\}.$$

Then

$$\operatorname{im}(f) = \{f(1), f(2), f(3), f(4)\} = \{a, a, b, c\} = \{a, b, c\}.$$ △

Example 25.28. Consider the function $f\colon \mathbb{R} \to \mathbb{R}$ given by the rule $f(x) = x^2$. What is the image of this function? In other words, what is the set of outputs? Since every output is nonnegative, and every nonnegative number a is equal to $f(\sqrt{a})$, we see that we have $\operatorname{im}(f) = \mathbb{R}_{\geq 0}$. △

Example 25.29. Let $A = \{1, 2, 3\}$, and let $f\colon A \to \mathscr{P}(A)$ be defined by $f(a) = \{a, 2\}$. The range of f is the set

$$\{f(a) : a \in A\} = \{f(1), f(2), f(3)\} = \{\{1, 2\}, \{2\}, \{2, 3\}\}. \qquad \triangle$$

A useful way to think about $\operatorname{im}(f)$ is as the set of all elements of B that are second coordinates of some element of f.

Theorem 25.30. *Let A and B be sets, and $f\colon A \to B$ a function. Then f is surjective if and only if the range of f is equal to B.*

Proof. The function f is surjective if and only if every $b \in B$ is equal to $f(a)$ for some $a \in A$. This happens if and only if $B = \{f(a) : a \in A\} = \operatorname{im}(f)$. □

25.D Bijective functions

We have seen two important properties of functions: injectivity and surjectivity. When we combine these two properties we obtain a very special type of function.

Definition 25.31. A *bijective* function is a function that is both injective and surjective.

Remark 25.32. A bijective function is often called a *bijection*. (Similarly, injective functions are called *injections*, and surjective functions are called *surjections*.) It is also called a "one-to-one correspondence." Although the latter term risks being confused with the very similar phrase "one-to-one function," which does not imply surjectivity, it is nevertheless frequently used. ▲

Example 25.33. Let $A = \{1, 2, 3\}$ and let $B = \{4, 5, 6\}$. Define a function $f\colon A \to B$ by

$$f = \{(1, 5), (2, 6), (3, 4)\}.$$

Since each element of A occurs exactly once as a first coordinate of an ordered pair in f, we know f is a function. Since each element of B occurs exactly once as a second coordinate, f is both injective and surjective by (25.4) and (25.15). Hence, f is bijective. △

Example 25.34. Define $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = 2x + 1$.

We have seen (Example 25.7) that f is injective and (Example 25.17) that f is surjective. Hence, f is a bijection. △

Example 25.35. Let $f\colon \mathbb{R} - \{2\} \to \mathbb{R} - \{1\}$ be defined by

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

We have seen (Example 25.10) that f is injective and (Example 25.21) that f is surjective. Hence, f is a bijection. △

In the next chapter we will see that bijections help us show that two sets have the same size.

25.E Exercises

Exercise 25.1. Let $A = \{1, 2, 3\}$ and $B = \{x, y\}$. List all functions from $A \to B$, and for each function state (without proof) whether it is injective, surjective, both (bijective), or none of the above.

Now do the same for all functions from $B \to A$.

Exercise 25.2. For each of the following, determine (with proof) whether the function is injective and/or surjective.

(a) Define $f\colon \mathbb{Z} \to \mathbb{Z}$ by $f(n) = 2n + 1$.
(b) Define $g\colon \mathbb{R} \to \mathbb{R}$ by $g(x) = x^2 + 2x + 2$.
(c) Define $h\colon \mathbb{Z} \to \mathbb{Z}$ by $h(n) = n + 3$.

Exercise 25.3. Define $f\colon \mathbb{Z}_5 \to \mathbb{Z}_5$ by $f(\overline{a}) = \overline{2a + 3}$.

(a) Prove that f is well-defined.
(b) Is f injective? Surjective? Give proofs. (Hint: You cannot divide by $\overline{2}$, but you can multiply by $\overline{3}$. Alternatively, write out the ordered pairs and check all cases.)

Exercise 25.4. Find a function $f\colon \mathbb{R} \to \mathbb{R}$ that is

(a) neither injective nor surjective.
(b) injective but not surjective.
(c) surjective but not injective.
(d) both injective and surjective.

In all cases give proofs. (Hint: For some of these, piecewise defined functions may be useful.)

Exercise 25.5. Define $f\colon \mathbb{R} - \{2\} \to \mathbb{R} - \{1\}$ by

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

(a) Prove that f is a function from $\mathbb{R} - \{2\}$ to $\mathbb{R} - \{1\}$. (The only question is whether $f(x) \in \mathbb{R} - \{1\}$ whenever $x \in \mathbb{R} - \{2\}$.)
(b) Prove that f is injective.
(c) Prove that f is surjective.

Exercise 25.6. Define $f\colon \mathbb{Z}^2 \to \mathbb{Z}$ by $f(m, n) = 3m - 2n$. Is f injective? Surjective? (Give proofs.)

Exercise 25.7. Describe (without proof) the image of each of the following functions from $\mathbb{R} \to \mathbb{R}$.

(a) $\sin(x)$.
(b) e^x.
(c) x^3.
(d) $\sqrt{|x|}$.

Exercise 25.8. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function. Suppose that we graph f in the xy-plane, with the domain being the horizontal axis, and the codomain being the vertical axis. Prove the following:

(a) (The vertical line test): Since f is a function, every vertical line intersects the graph of f at most once.
(b) (No holes in the domain): Since f is a function, every vertical line intersects the graph of f at least once.
(c) (The horizontal line test): The function f is injective if and only if every horizontal line intersects the graph of f at most once.
(d) (No holes in the codomain): The function f is surjective if and only if every horizontal line intersects the graph of f.

26 Composition of functions

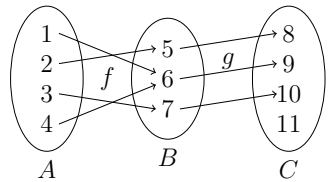
Fix $A = \{1, 2, 3, 4\}$, $B = \{5, 6, 7\}$, and $C = \{8, 9, 10, 11\}$, and fix functions $f\colon A \to B$ and $g\colon B \to C$ defined by

$$f = \{(1, 6), (2, 5), (3, 7), (4, 6)\},$$

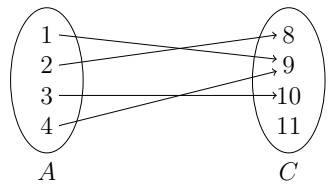
and

$$g = \{(5, 8), (6, 9), (7, 10)\}.$$

A picture of this situation appears below:



One can ask what might happen if we were to begin at an element of A, and first follow the arrow for the function f (to an element of B), and then follow the arrow for the function g (to an element of C). We would then get a drawing like the following:



Each arrow in the new diagram is obtained by traversing two arrows in the original diagram. We notice that the new diagram describes a function, since each element of A has a single arrow emanating from it. We call this function the composition. In this section we will formally define compositions and study how function properties (such as injectivity or surjectivity) behave on composites of functions.

26.A Defining function composition

Here is the formal definition of what it means to compose two functions.

Definition 26.1. Let $f\colon A \to B$ and $g\colon B \to C$ be functions. We define the *composition* of g with f, written $g \circ f$, to be the new function $g \circ f\colon A \to C$ given by the relation

$$g \circ f = \{(a, g(f(a))) \in A \times C \,:\, a \in A\}.$$

Remark 26.2. We note that this definition makes sense: since $a \in A$, we have that $f(a)$ is in B, so $g(f(a))$ exists and is in C. ▲

Advice 26.3. When the codomain of f is not equal to the domain of g, there is no composite function $g \circ f$ (i.e., it is not defined).

Example 26.4. Let $A = \{1, 2, 3, 4\}$, $B = \{5, 6, 7\}$, and $C = \{8, 9, 10, 11\}$, and suppose we have functions $f\colon A \to B$ and $g\colon B \to C$ defined by

$$f = \{(1, 6), (2, 5), (3, 7), (4, 6)\},$$

and

$$g = \{(5, 8), (6, 9), (7, 10)\}.$$

Since $f(1) = 6$ and $g(6) = 9$, we have $(g \circ f)(1) = g(f(1)) = g(6) = 9$.
Since $f(2) = 5$ and $g(5) = 8$, we have $(g \circ f)(2) = g(f(2)) = g(5) = 8$.
Similarly, $(g \circ f)(3) = 10$ and $(g \circ f)(4) = 9$. Hence, we have that

$$g \circ f = \{(1, 9), (2, 8), (3, 10), (4, 9)\}. \qquad \triangle$$

Example 26.5. Let $f\colon \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x + 1$ and let $g\colon \mathbb{R} \to \mathbb{R}$ be given by $g(x) = x^2$. Then $g \circ f\colon \mathbb{R} \to \mathbb{R}$ is given by

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

In this case, since the codomain of g is the same as the domain of f, we may also construct the function $f \circ g\colon \mathbb{R} \to \mathbb{R}$. This function is given by

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

Notice that $f \circ g$ is *not* the same function as $g \circ f$. The order in which the functions are composed does matter. Because of this, we say that *function composition is not commutative*. This is in contrast to operations like addition of real numbers, where $a + b = b + a$. $\triangle$

Although function composition is not commutative, it does satisfy a different property that is very useful. This property is motivated by the question: What happens if we compose three functions?

Theorem 26.6. *Let A, B, C, and D be sets, and let $f\colon A \to B$, $g\colon B \to C$, and $h\colon C \to D$ be three given functions. The two new functions $h \circ (g \circ f)\colon A \to D$ and $(h \circ g) \circ f\colon A \to D$ are equal.*

Proof. Both $h \circ (g \circ f)$ and $(h \circ g) \circ f$ have domain A and codomain D. Hence, to prove that they are equal, we need only check that for each $a \in A$ both functions give the same image.

Let $a \in A$. Then

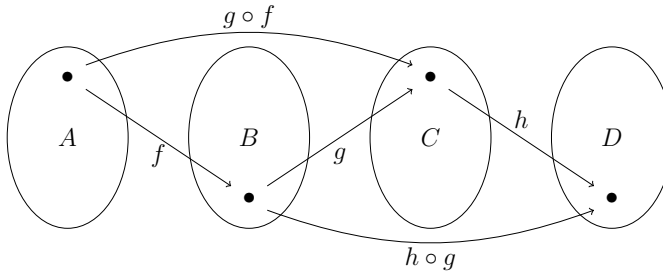
$$\big(h \circ (g \circ f)\big)(a) = h\big((g \circ f)(a)\big) = h(g(f(a))),$$

and

$$\big((h \circ g) \circ f\big)(a) = (h \circ g)\big(f(a)\big) = h(g(f(a))).$$

Hence, we see that the two functions give the same image for a. Therefore, they are equal. □

Remark 26.7. A graphical representation of this theorem is given below.



The statement of the theorem is that following the arrow marked $g \circ f$ and then the arrow marked h, in other words $h \circ (g \circ f)$, yields the same result as following f and then following $h \circ g$, in other words $(h \circ g) \circ f$. ▲

Remark 26.8. The fact that $(h \circ g) \circ f = h \circ (g \circ f)$ is called *associativity* of function composition. Associativity is a property that shows up in many places in math. For $a, b, c \in \mathbb{R}$, we have two different forms of associativity, namely $(a + b) + c = a + (b + c)$ for addition and $(ab)c = a(bc)$ for multiplication. If you have seen matrix multiplication, then you know that if A, B, and C are matrices that can be multiplied, then $(AB)C = A(BC)$. Different manifestations of associativity are often, in some way, related to the associativity of function composition. ▲

There is one special function that behaves very well with respect to composition.

Definition 26.9. Let A be a set. Define the function $\mathrm{id}_A \colon A \to A$ by $\mathrm{id}_A(a) = a$ for each $a \in A$. This is the *identity function* on A.

Example 26.10. Let $A = \{1, 2, 3\}$. Then, as a collection of ordered pairs,

$$\mathrm{id}_A = \{(1, 1), (2, 2), (3, 3)\}.$$

Let $B = \{r, s, t, u, v\}$. Then, as a collection of ordered pairs,

$$\mathrm{id}_B = \{(r, r), (s, s), (t, t), (u, u), (v, v)\}.$$ △

Theorem 26.11. *Let A and B be sets, and let $f \colon A \to B$ be a function. We have*

(1) $\mathrm{id}_B \circ f = f$ *and*
(2) $f \circ \mathrm{id}_A = f$.

Proof. We prove (1) and leave (2) as an exercise.

Note that $\mathrm{id}_B \circ f\colon A \to B$ and $f\colon A \to B$ have the same domain and codomain. Let $a \in A$. Notice that $f(a) \in B$. Then

$$(\mathrm{id}_B \circ f)(a) = \mathrm{id}_B(f(a)) = f(a).$$

Hence, we see that $\mathrm{id}_B \circ f = f$. □

26.B Composition of injective and surjective functions

When we compose functions, properties of the component functions can sometimes be deduced from the composition, and properties of the composition can sometimes be deduced from the component functions. We investigate both of these possibilities.

Theorem 26.12. *Let A, B, and C be sets, and let $f\colon A \to B$ and $g\colon B \to C$ be functions.*

(1) *If $g \circ f$ is injective, then f is injective.*
(2) *If $g \circ f$ is surjective, then g is surjective.*
(3) *If f and g are injective, then $g \circ f$ is injective.*
(4) *If f and g are surjective, then $g \circ f$ is surjective.*

Proof. (1) Assume that $g \circ f\colon A \to C$ is injective. To prove that f is injective, let $a_1, a_2 \in A$ and assume $f(a_1) = f(a_2)$; we will show $a_1 = a_2$. Applying g to both sides of the assumed equality yields $g(f(a_1)) = g(f(a_2))$, and so

$$(g \circ f)(a_1) = (g \circ f)(a_2).$$

Since $g \circ f$ is injective, we see that $a_1 = a_2$. Hence, f is injective.

(2) See Exercise 26.1.

(3) See Exercise 26.1.

(4) Assume that f and g are surjective. We wish to show that $g \circ f$ is surjective. Let $c \in C$. Since $g\colon B \to C$ is surjective, there is some $b \in B$ such that $g(b) = c$. Since $f\colon A \to B$ is surjective, there is some $a \in A$ such that $f(a) = b$. Then $(g \circ f)(a) = g(f(a)) = g(b) = c$. □

26.C Inverse relations

We now define a way of constructing new relations from old, and apply it to functions.

Definition 26.13. Let R be a relation from A to B. The *inverse relation* is the set

$$R^{-1} = \{(b, a) \,:\, (a, b) \in R\}$$

where we reverse each ordered pair in R. It is a relation from B to A.

Example 26.14. Let $A = \{1, 2, 3, 4\}$ and $B = \{1, 2, 3, 4, 5\}$. Define a relation R by

$$R = \{(1, 2), (2, 2), (3, 4), (4, 1)\}.$$

Then

$$R^{-1} = \{(2, 1), (2, 2), (4, 3), (1, 4)\}.$$

Is R a function? (Yes.)
Is R^{-1} a function? (No.) △

Suppose that $f\colon A \to B$ is a function. Then f is a relation, so it has an inverse relation. What can keep this inverse relation from being a function? There are two possibilities.

(1) If some element of B is not a second coordinate of any ordered pair in f, then it will not be a first coordinate of any ordered pair in f^{-1}. Hence, the inverse relation will not be a function. In other words, if f is not surjective, then f^{-1} will not be a function because it is not defined somewhere.
(2) If some element of B is a second coordinate of more than one ordered pair in f, then it will be a first coordinate of multiple ordered pairs in the inverse relation. Hence, the inverse relation will not be a function. In other words, if f is not injective, then f^{-1} will not be a function because it "maps" some element to two different "outputs."

It would appear that in order for the inverse relation of f to be a function, each element of B should be the second coordinate of exactly one ordered pair in f. In other words, f should be a bijection. We state this as a theorem and give a proof.

Theorem 26.15. *Let $f\colon A \to B$ be a function. Let g be the inverse relation to f. Then g is a function from B to A if and only if f is a bijection.*

Proof. Let $f\colon A \to B$ be an arbitrary function, and let $g = f^{-1}$ be the inverse relation.

($\Leftarrow$): Assume that f is a bijection. Then every element of B is the second coordinate of exactly one ordered pair in f. Hence, every element of B is the first coordinate of exactly one ordered pair in g. Therefore, g is a function.

($\Rightarrow$): Conversely, assume that $g\colon B \to A$ is a function. Hence, every element of B is the first coordinate of exactly one ordered pair in g. Thus, every element of B is the second coordinate of exactly one ordered pair in f. Therefore, f is a bijection. □

Example 26.16. Let $f\colon \mathbb{R} \to \mathbb{R}$ be given by $f(x) = 2x + 1$. We've shown previously that f is a bijection. As a collection of ordered pairs, we have

$$f = \{(x, y) \,:\, x \in \mathbb{R}, y = 2x + 1\}.$$

Hence, the inverse relation is the set

$$f^{-1} = \{(y, x) \,:\, x \in \mathbb{R}, y = 2x + 1\} = \{(y, x) \,:\, y \in \mathbb{R}, x = (y - 1)/2\},$$

since if $y = 2x + 1$, then we may solve for x and we find $x = (y - 1)/2$. So we see that f^{-1} is a function given by the rule $f^{-1}(y) = (y - 1)/2$. △

Advice 26.17. Suppose a function $f(x) = y$ is given by a simple rule. If we can solve for x in terms of y, this gives us the rule $f^{-1}(y) = x$ for the inverse relation.

In the example above, we had $f(x) = 2x + 1$. We set $y = 2x + 1$, subtract 1 and divide by 2 to obtain

$$x = \frac{y-1}{2},$$

showing that the inverse relation to f is $f^{-1}(y) = (y-1)/2$.

Example 26.18. Let $f\colon \mathbb{R} \to \mathbb{R}$ be given by the rule $f(x) = x^2$. Setting $y = x^2$, we might think that we can solve for x to get $x = \sqrt{y}$, and then guess $f^{-1}(y) = \sqrt{y}$. This is not correct, because when we solved for x we should have set $x = \pm\sqrt{y}$. Thus, the inverse relation is not a function. (This can also be seen from the fact that f is neither injective, nor surjective.) △

Example 26.19. We have seen that $f\colon \mathbb{R} - \{2\} \to \mathbb{R} - \{1\}$ given by

$$f(x) = \frac{x-1}{x-2}$$

is a bijection. Setting $y = f(x)$, we will solve for x. Multiplying by $x - 2$, we find

$$xy - 2y = x - 1,$$

so that

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

and, dividing by $y - 1$, we obtain

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

Hence, the inverse relation to f is $f^{-1}\colon \mathbb{R} - \{1\} \to \mathbb{R} - \{2\}$ given by

$$f^{-1}(y) = \frac{2y-1}{y-1}.$$ △

26.D Composition of inverse functions

In general, if f is a bijective function, we denote the inverse relation (which is also a function) by f^{-1}, and we call it the inverse (function) of f. In cases where f is a function that is not bijective, we typically do not use the notation f^{-1} for the inverse relation (in fact, we seldom discuss the inverse relation). However, we may use the notation f^{-1} in other contexts (for instance, see the next section). We have the following theorem concerning a bijective function and its inverse.

Theorem 26.20. *Let $f\colon A \to B$ be a bijective function. The following each hold:*
(1) $(f^{-1})^{-1} = f$.
(2) $f^{-1}\colon B \to A$ *is a bijective function.*
(3) $f^{-1} \circ f\colon A \to A$ *is equal to* id_A.
(4) $f \circ f^{-1}\colon B \to B$ *is equal to* id_B.

Proof. Throughout the proof we let $f\colon A \to B$ be a bijection.
(1) This part holds for *any* relation f. If we reverse the ordered pairs in f, and then reverse again, we are back to the original relation.
(2) By Theorem 26.15, we know f^{-1} is a function. Now, by part (1) above, its inverse relation is f, which is a function. Thus, by Theorem 26.15 again, we know f^{-1} is a bijection.
(3) We note that $f^{-1} \circ f$ and id_A have the same domain and codomain. Now for any $a \in A$, we have $(a, f(a)) \in f$, so we see that $(f(a), a) \in f^{-1}$. This tells us that $f^{-1}(f(a)) = a$, proving that $(f^{-1} \circ f)(a) = a = \mathrm{id}_A(a)$. Hence, $f^{-1} \circ f = \mathrm{id}_A$.
(4) This is similar to part (3), by switching the roles of f and f^{-1}. □

26.E Exercises

Exercise 26.1. Let $f\colon A \to B$ and $g\colon B \to C$ be functions.
(a) Prove that if f and g are injective, then $g \circ f$ is injective.
(b) Prove that if $g \circ f$ is surjective, then g is surjective.

Exercise 26.2. Let $f\colon A \to B$ be a function. Prove that $f \circ \mathrm{id}_A = f$.

Exercise 26.3. Prove or disprove: If $f\colon A \to B$ and $g\colon B \to C$ are functions, and g is surjective, then $g \circ f$ is surjective.

Exercise 26.4. Prove or disprove: If $f\colon A \to B$ and $g\colon B \to C$ are functions, and $g \circ f$ is injective, then g is injective.

Exercise 26.5. Let $f\colon A \to B$ and $g\colon B \to C$ be functions. Prove that if f and g are both bijective, then $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.
(Hint: How do you check if functions are equal?)

Exercise 26.6. Prove that the function $f\colon \mathbb{R} - \{5\} \to \mathbb{R} - \{3\}$ given by

$$f(x) = \frac{3x+1}{x-5}$$

is bijective. Find $f^{-1}(y)$ for $y \in \mathbb{R} - \{3\}$.

Exercise 26.7. Let $A = \{1, 2, 3\}$ and let $f\colon A \to A$ be given as

$$f = \{(1, 2), (2, 3), (3, 1)\}.$$

(a) Determine f^{-1}.
(b) Determine $f \circ f$.
(c) Determine $f \circ f \circ f$.
(d) Define

$$f^n = \underbrace{f \circ \cdots \circ f}_{n \text{ times}}.$$

Then $f^1 = f$, $f^2 = f \circ f$, and $f^3 = f \circ f \circ f$, and so forth. Determine f^n, as a collection of ordered pairs, for each natural number n.

27 Additional facts about functions

27.A Functions between finite sets

For finite sets A and B, we can draw conclusions about the sizes of the sets based on properties of functions between them.

Given $f\colon A \to B$, if either A or B is a finite set, we can sometimes obtain information about whether f is injective or surjective by studying the size of $\operatorname{im}(f)$.

Theorem 27.1. *Let A and B be sets, and let $f\colon A \to B$ be a function. If A is finite, then $|\operatorname{im}(f)| \leq |A|$. Moreover (still assuming A is finite) f is injective if and only if $|\operatorname{im}(f)| = |A|$.*

Proof. As a collection of ordered pairs, f has $|A|$ elements. Hence, there are at most $|A|$ second coordinates. Thus, there cannot be more than $|A|$ elements in the image of f.

We now prove the last sentence. For the forward direction, assume that f is injective. Then all second coordinates of pairs in f are distinct, so the number of such second coordinates, which is $|\operatorname{im}(f)|$, is equal to $|f| = |A|$.

For the converse we work contrapositively. Assume that f is not injective. Then f has fewer than $|A|$ distinct second coordinates (since at least two of the $|A|$ second coordinates must be equal). Hence, $|\operatorname{im}(f)| < |A|$. □

Theorem 27.2. *Let A and B be sets, and let $f\colon A \to B$ be a function. If B is finite, then $|\operatorname{im}(f)| \leq |B|$. Moreover (still assuming B is finite) f is surjective if and only if $|\operatorname{im}(f)| = |B|$.*

Proof. Left as Exercise 27.1. □

Theorem 27.3. *Let A and B both be finite sets, and let $f\colon A \to B$ be a function.*
(1) *If f is injective, then $|A| \leq |B|$.*
(2) *If f is surjective, then $|A| \geq |B|$.*
(3) *If f is bijective, then $|A| = |B|$.*

Proof. We make use of Theorems 27.1 and 27.2.
(1) If f is injective, then $|A| = |\operatorname{im}(f)| \leq |B|$ (since $\operatorname{im}(f) \subseteq B$).
(2) If f is surjective, then $|A| \geq |\operatorname{im}(f)| = |B|$.
(3) This follows from (1) and (2). □

Example 27.4. Let $A = \{1, 2, 3, 4\}$ and let $B = \{1, 2, 3, 4, 5\}$. Then $|A| < |B|$. Hence, by the contrapositive of part (2) above, we know that there can be no surjective function from A to B. This also implies that there can be no bijective function from A to B. △

A special property of finite sets is that, in some cases, injectivity and surjectivity may imply each other.

Theorem 27.5. *Let A and B be finite sets and assume $|A| = |B|$. A function $f\colon A \to B$ is injective if and only if it is surjective.*

Proof. Suppose A and B are finite, $|A| = |B|$, and that $f\colon A \to B$ is a function.
($\Rightarrow$): Assume f is not surjective. Then $|\mathrm{im}(f)| < |B| = |A|$, so f is not injective.
($\Leftarrow$): Assume f is not injective. Then $|\mathrm{im}(f)| < |A| = |B|$, so f is not surjective. □

Warning 27.6. Theorem 27.5 does not apply if A and B are infinite sets. Firstly, we don't yet know what it means for two infinite sets to have the same size. However, even if $A = B$, there are still problems. For instance, define $f\colon \mathbb{N} \to \mathbb{N}$ by

$$f(n) = n + 1$$

and define $g\colon \mathbb{N} \to \mathbb{N}$ by

$$g(n) = \begin{cases} 1 & \text{if } n = 1, \\ n - 1 & \text{if } n > 1. \end{cases}$$

One may quickly check that f is injective but not surjective, and g is surjective but not injective.

Remark 27.7. A common use of Theorem 27.5 occurs when A is a finite set and $f\colon A \to A$ is a function. In this case, the sizes of the domain and codomain are clearly equal, so f is injective if and only if f is surjective. ▲

Example 27.8. Let $A = \{1, 2, 3\}$. If $f\colon A \to A$ is a function, is it automatically bijective?

No, it is not. But if it is injective, then it is automatically surjective (and vice versa), since the domain and codomain are both finite of the same size. △

27.B Partitions and pasting functions

Often, it will happen that we wish to define a function from a set A to a set B. However, it might be the case that A is too large or too complicated to make it convenient for us to define such a function. However, if we can partition A, then we may define a piecewise function on A. Proving injectivity and surjectivity of piecewise functions can be somewhat more complicated than for functions defined by a single simple rule, but by carefully using the definitions, and proceeding with a proof by cases, it is usually not too difficult.

We will now examine a special case of piecewise defined functions for which injectivity and surjectivity are easy to prove. These functions will be useful to us later.

Definition 27.9. Let A and B be sets, and suppose $\{P_1, \ldots, P_n\}$ is a partition of A with n parts and $\{Q_1, \ldots, Q_n\}$ is a partition of B with n parts. Assume that for each i with $1 \leq i \leq n$, we are given a function $f_i \colon P_i \to Q_i$. We may define a function $f \colon A \to B$ by

$$f = \bigcup_{i=1}^{n} f_i$$

with the rule $f(a) = f_i(a)$ if $a \in P_i$. We call f the function obtained by *pasting together* the f_i.

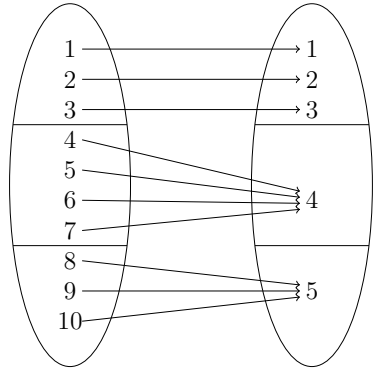
Example 27.10. Let $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and let $P_1 = \{1, 2, 3\}$, $P_2 = \{4, 5, 6, 7\}$, and $P_3 = \{8, 9, 10\}$. Then $\{P_1, P_2, P_3\}$ is a partition of A.

Let $B = \{1, 2, 3, 4, 5\}$, $Q_1 = \{1, 2, 3\}$, $Q_2 = \{4\}$, and $Q_3 = \{5\}$. Then $\{Q_1, Q_2, Q_3\}$ is a partition of B.

Define $f_1 \colon P_1 \to Q_1$ by $f_1(x) = x$. Define $f_2 \colon P_2 \to Q_2$ by $f_2(x) = 4$, and define $f_3 \colon P_3 \to Q_3$ by $f_3(x) = 5$. Pasting together the f_i we obtain a function f given by the rule

$$\begin{aligned} f(x) &= \begin{cases} f_1(x) & \text{if } x \in P_1, \\ f_2(x) & \text{if } x \in P_2, \\ f_3(x) & \text{if } x \in P_3. \end{cases} \\ &= \begin{cases} x & \text{if } x \in P_1, \\ 4 & \text{if } x \in P_2, \\ 5 & \text{if } x \in P_3. \end{cases} \end{aligned}$$

We can view this situation as follows:



△

The following theorem tells us when a pasted together function is injective, surjective, or both.

Theorem 27.11 (The Pasting Together Theorem)**.** *Using the notation as given in* Definition 27.9*, each of the following holds:*

(1) *If each f_i is injective, then f is injective.*
(2) *If each f_i is surjective, then f is surjective.*
(3) *If each f_i is bijective, then f is bijective.*

Proof. (1) Suppose that each f_i is injective. We will show that f is injective, so let $a_1, a_2 \in A$ and assume $f(a_1) = f(a_2)$. We know that $a_1 \in P_i$ and $a_2 \in P_j$ for some i and j (since the P's partition A). Hence $f(a_1) = f_i(a_1) \in Q_i$, and $f(a_2) = f_j(a_2) \in Q_j$. Since $f(a_1) = f(a_2)$ we have $Q_i \cap Q_j \neq \emptyset$. But the Q's partition B, hence $Q_i = Q_j$, or in other words $i = j$. Thus

$$f_i(a_1) = f(a_1) = f(a_2) = f_i(a_2).$$

Since f_i is injective, $a_1 = a_2$.

(2) Suppose that each f_i is surjective. Let $b \in B$, so $b \in Q_i$ for some i. Since f_i is surjective we can fix some $a \in P_i$ with $f_i(a) = b$. Hence, $f(a) = f_i(a) = b$. Therefore, f is surjective.

(3) This part follows from (1) and (2). □

Example 27.12. We partition $\mathbb{N}$ into $\{P_1, P_2\}$, where P_1 is the set of even natural numbers, and P_2 is the set of odd natural numbers. We partition $\mathbb{Z}$ into $\{Q_1, Q_2\}$, where Q_1 is the set of positive integers, and Q_2 is the set of nonpositive integers. Define $f_1 \colon P_1 \to Q_1$ by $f_1(n) = n/2$, and $f_2 \colon P_2 \to Q_2$ by $f_2(n) = -(n-1)/2$.

Pasting together, define $f \colon \mathbb{N} \to \mathbb{Z}$ by

$$f(n) = \begin{cases} n/2 & \text{if } n \text{ is even,} \\ -(n-1)/2 & \text{if } n \text{ is odd.} \end{cases}$$

One checks easily that f_1 and f_2 are bijections (see Exercise 27.2). Hence, by the Pasting Together Theorem, f is a bijection. △

27.C Restrictions of functions

Suppose $f \colon A \to B$ is a function and S is a subset of A. It will often be the case that we wish to study the function f limited to the set S. This situation occurs often enough that we have a special name and notation for it.

Definition 27.13. Let A and B be sets and let $f \colon A \to B$ be a function. If $S \subseteq A$, we can define a new function

$$f|_S \colon S \to B$$

by the rule $f|_S(x) = f(x)$ for each $x \in S$. We call $f|_S$ the *restriction of f to S*, and read it as "f restricted to S."

Example 27.14. Let $f\colon \mathbb{R} \to \mathbb{R}$ be the function defined by $f(x) = x^2$. We note that f is not injective, since $f(-1) = f(1)$. If we let S be the set of *nonnegative* real numbers, then the function $f|_S\colon S \to \mathbb{R}$ is in fact injective. Note that it is defined by the same rule as f, except that we restrict the allowed inputs. △

Remark 27.15. If we think about the restriction of f to S in terms of ordered pairs, we come up with the following description. If $f\colon A \to B$ is the function

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

then

$$f|_S = \{(a, f(a)) \,:\, a \in S\}.$$

Hence, $f|_S$ is just the set of ordered pairs in f whose first coordinate is in S. ▲

It is also possible to adjust a function by changing its codomain. There is no standard notation (or even name) for this process. We will define a notation, but note that it is not standard.

Definition 27.16. Let $f\colon A \to B$ be a function. We define a new function $\hat{f}\colon A \to \operatorname{im}(f)$ by the rule $\hat{f}(a) = f(a)$ for all $a \in A$. We call $\hat{f}$ the *surjective reduction of* f.

Example 27.17. Let $f\colon \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$. We know f is not surjective, because -1 is not in the range of f. Note that the image of f is the set $\operatorname{im}(f) = [0, \infty)$.

If we let $\hat{f}\colon \mathbb{R} \to [0, \infty)$ be the surjective reduction of f, then we see that $\hat{f}$ is a surjective function with the same defining rule as f, but with the new codomain $\operatorname{im}(f)$. △

You should try to prove the following theorem before reading the proof.

Theorem 27.18. *Given a function $f\colon A \to B$, the surjective reduction*

$$\hat{f}\colon A \to \operatorname{im}(f)$$

is a surjective function.

Proof. Since $\hat{f}$ consists of the same ordered pairs as f, it is a function. (Each element of the domain A is the first coordinate of exactly one ordered pair, and the second coordinate is an element of $\operatorname{im}(f)$.) For each $b \in \operatorname{im}(f)$, there is some $a \in A$ such that $f(a) = b$ (by the definition of the image). Then $\hat{f}(a) = b$, so $\hat{f}$ is surjective. □

Remark 27.19. In essence, Theorem 27.18 says "a function f is surjective onto its image." ▲

27.D Images and preimages

Given a function $f\colon A \to B$, we previously defined the image $f(a) \in B$ of an element $a \in A$. We now give a new meaning of the word "image" that applies to subsets of A, rather than individual elements.

Definition 27.20. Let A and B be sets and let $f\colon A \to B$ be a function. If S is a subset of A, we define the *image of S under f* (or just "the image of S") to be

$$f(S) = \{f(x) \,:\, x \in S\}.$$

We note that $f(S)$ is a subset of B.

Example 27.21. Let $A = \{1, 2, 3, 4\}$, $B = \{5, 6, 7, 8\}$, and

$$f = \{(1, 6), (2, 5), (3, 8), (4, 7)\}.$$

Then $f(\{1, 2\}) = \{5, 6\}$ and $f(\{1, 3\}) = \{6, 8\}$. △

Example 27.22. Define $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = 2x + 1$. Let $[0, 1]$ be the closed interval of real numbers from 0 to 1. Then the image of $[0, 1]$ under f is the set

$$f([0, 1]) = [1, 3].$$

To see this, note that if $0 \le x \le 1$, then $0 \le 2x \le 2$, so $1 \le 2x + 1 \le 3$. Hence, if $x \in [0, 1]$, then $f(x) \in [1, 3]$.

Conversely, if $y \in [1, 3]$, then $(y - 1)/2 \in [0, 1]$, and $f((y - 1)/2) = y$. Hence, every element of $[1, 3]$ is the image of some element in $[0, 1]$. △

Remark 27.23. Note that for any function $f\colon A \to B$, we have $f(A) = \mathrm{im}(f)$. ▲

A final concept that is often mentioned in relation to functions is the preimage of a set.

Definition 27.24. Let $f\colon A \to B$ be a function, and let T be a subset of B. We define the *preimage* of T to be the set

$$f^{-1}(T) = \{a \in A \,:\, f(a) \in T\}.$$

Note that $f^{-1}(T)$ is a subset of A.

Warning 27.25. Although the same symbol is used for the preimage and the inverse function of f (if f is a bijection), we note that the two concepts are quite different. The preimage exists even if f is not a bijection, while we have seen that the inverse function only exists if f is a bijection.

Example 27.26. Let $A = \{1, 2, 3, 4, 5\}$ and $B = \{6, 7, 8, 9\}$. Let

$$f = \{(1, 7), (2, 6), (3, 7), (4, 6), (5, 9)\}.$$

Then we can calculate the preimages

$$f^{-1}(\{6, 7\}) = \{1, 2, 3, 4\}, \qquad f^{-1}(\{8\}) = \emptyset,$$
$$f^{-1}(\{9\}) = \{5\}, \qquad f^{-1}(\{7, 8, 9\}) = \{1, 3, 5\},$$

as well as preimages of many other subsets of B. (How many subsets of A are preimages of subsets of B? There are less than 16.) △

Example 27.27. Let $f\colon \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$. Then for $x > 0$ we have $f^{-1}(\{x\}) = \{\sqrt{x}, -\sqrt{x}\}$, for $x < 0$ we have $f^{-1}(\{x\}) = \emptyset$, and $f^{-1}(\{0\}) = \{0\}$. △

27.E Exercises

Exercise 27.1. Prove Theorem 27.2.

Exercise 27.2. Prove that the functions f_1 and f_2 defined in Example 27.12 are both bijections.

Exercise 27.3. Give an example of a bijective function $f\colon \mathbb{Z} \to \{0, 1\} \times \mathbb{N}$ and include a proof that it is bijective.

(Hint: Partition $\mathbb{Z}$ into positive and nonpositive integers. Partition $\{0, 1\} \times \mathbb{N}$ into

$$\{\{(0, n) : n \in \mathbb{N}\}, \{(1, n) : n \in \mathbb{N}\}\}.$$

Define two bijections, and then use the Pasting Together Theorem.)

Exercise 27.4. Let $A = \{n \in \mathbb{Z} : -3 \leq n \leq 3\}$, and let $f\colon A \to \mathbb{Z}$ be defined by $f(x) = x^2 + 2x + 2$.

(a) Write f as a set of ordered pairs.
(b) Find $\operatorname{im}(f)$.
(c) Find a subset C of A so that $f|_C$ is injective and $\operatorname{im}(f|_C) = \operatorname{im}(f)$.

Exercise 27.5. Let $f\colon A \to B$ be an injective function, and let S be an arbitrary subset of A.

(a) Prove that $f|_S\colon S \to B$ is injective.
(b) Prove that $\hat{f}$ is a bijection.

Exercise 27.6. Let $f\colon A \to B$ be a function.
(a) Prove that f is surjective if and only if $f^{-1}(\{b\}) \neq \emptyset$ for each $b \in B$.
(b) Prove that f is injective if and only if $|f^{-1}(\{b\})| \leq 1$ for each $b \in B$.

Exercise 27.7. Let $f\colon A \to B$ be a function, and let $X, Y \subseteq A$ and $C, D \subseteq B$. Prove or disprove each of the following equalities.
(a) $f(X \cup Y) = f(X) \cup f(Y)$.
(b) $f(X \cap Y) = f(X) \cap f(Y)$.
(c) $f^{-1}(C \cup D) = f^{-1}(C) \cup f^{-1}(D)$.
(d) $f^{-1}(C \cap D) = f^{-1}(C) \cap f^{-1}(D)$.
(Hint: At least one of these is true and at least one of these is false.)

Chapter VIII

Cardinality

...it's very much like your trying to reach Infinity. You know that it's there, but you just don't know where—but just because you can never reach it doesn't mean that it's not worth looking for. Norton Juster, The Phantom Tollbooth

In the very first chapter of this book, we defined the cardinality of a finite set to equal the number of its elements. Thus, for instance, the sets $\{a, b, c\}$ and $\{1, 2, 3\}$ have the same cardinality, which is 3. For infinite sets we cannot define the cardinality to be the *number* of elements, because such sets do not have any (finite) number of elements.

However, there is a reason we do not just define the cardinality of an infinite set to be the symbol ∞; there is a better way to measure the *size* of sets! This came as a shock to mathematicians in the late 1800's, who expected all infinite sets to have the same size. This theory was developed by Cantor, who showed that the set of real numbers $\mathbb{R}$ has *bigger* cardinality than $\mathbb{N}$. In this chapter we develop Cantor's theory of cardinality, which has become an important part of modern mathematics.

28 Definitions regarding cardinality

In this section, we will define what it means for two sets to have the same cardinality, discuss what it means for sets to have the same size as the natural numbers, and prove some basic facts about cardinality.

28.A How do we measure the size of sets?

The first question we must answer is: How do we measure the size of a set? For finite sets we simply count out the elements. For instance, if $A = \{68, 5, 39291, 90\}$, then we just count off each element, and we see that $|A| = 4$. By counting this way we form a bijection

$$f\colon \{1, 2, 3, 4\} \to A$$

given by

$$f(1) = 68, \qquad f(2) = 5, \qquad f(3) = 39291, \qquad \text{and} \qquad f(4) = 90.$$

Intuitively, we can think of sets having the same "size" if there is a bijection between them. This motivates the following definition.

Definition 28.1. Let A and B be sets. We say that A and B have the *same cardinality* if there exists a bijection $f\colon A \to B$. If this holds, we write $|A| = |B|$.

If there is no bijection from A to B, we say that they have *different cardinalities* and write $|A| \neq |B|$.

Remark 28.2. We will prove, shortly, that this relation is an equivalence relation, which will justify our use of an equality sign for the relation. ▲

The following are some examples and nonexamples of sets with the same cardinality.

Example 28.3. (1) Consider the three sets $A = \{a, b, c, d, e, f\}$, $B = \{1, 2, 3, 4, 5, 6\}$, and $C = \{0, 1, 2, 3, 4, 5\}$. It is easy to construct a bijection from A to B (since both sets have exactly six elements). So $|A| = |B|$. There are also bijections from A to C, and from B to C (since C also has six elements), so $|A| = |C|$, and $|B| = |C|$.

(2) Let $S = \{1, 2, 3\}$ and $T = \mathbb{N}$. There is no bijection from S to T (since T has more than 3 elements). Thus $|S| \neq |T|$.

(3) It can be tricky when working with infinite sets to tell whether they have the same cardinality. For instance, does $\mathbb{N}$ have the same cardinality as $2\mathbb{N}$? The answer is yes! There is a bijection $f\colon \mathbb{N} \to 2\mathbb{N}$, given by $f(n) = 2n$. In other words,

$$f(1) = 2, \qquad f(2) = 4, \qquad f(3) = 6, \qquad f(4) = 8, \ldots$$

is a bijection from $\mathbb{N}$ to $2\mathbb{N}$. Thus, we do have

$$|\mathbb{N}| = |2\mathbb{N}|.$$ △

The following example is so important that we'll call it a theorem.

Theorem 28.4. *The cardinalities of* $\mathbb{N}$ *and* $\mathbb{Z}$ *are the same.*

Proof. We need to construct a bijection $f\colon \mathbb{N} \to \mathbb{Z}$; or, in other words, we need to "count" the elements of $\mathbb{Z}$ using the natural numbers. The idea is to first count 0, and then successively count the positive and negative integers, as in the following table

n	1	2	3	4	5	6	7	8	9	10
$f(n)$	0	1	-1	2	-2	3	-3	4	-4	5

which yields the needed bijection between $\mathbb{N}$ and $\mathbb{Z}$. □

Remark 28.5. The previous proof was very informal. First, the definition of the function f is sloppy. To be more precise we should define f as a piecewise function from $\mathbb{N} \to \mathbb{Z}$ by the rule

$$f(n) = \begin{cases} n/2 & \text{if } n \text{ is even,} \\ -(n-1)/2 & \text{if } n \text{ is odd.} \end{cases}$$

This is exactly the same function as in Example 27.12.

Second, we didn't prove that the function f is injective and surjective. This was previously assigned as Exercise 27.2.

Amazingly, it turns out that this function can be expressed by a (somewhat complicated) single formula

$$f(n) = \frac{1 + (-1)^n(2n-1)}{4}.$$

(It is not expected that a student would be able to come up with this formula without a lot of help!) ▲

Advice 28.6. To prove that two sets have the same cardinality you are required to find a bijection between the two sets. Generally there are many different bijections. Try to look for a simple one.

We end this subsection with one more example.

Example 28.7. We will prove that the open interval $A = (0, 1)$ and the open interval $B = (1, 4)$ have the same cardinality. We thus want to construct a bijection between these two sets. The most obvious option would be to stretch by a factor of 3 and then shift right by 1. So we define $g\colon (0, 1) \to (1, 4)$ by the rule

$$g(x) = 1 + 3x.$$

It is straightforward to check that g is a function from A to B, and that g is both injective and surjective. △

28.B Basic results and a picture

In the previous subsection, we defined what it means for two sets to have the same cardinality. To really justify that definition, we should show that the relation of "having the same cardinality" is an equivalence relation on sets.

Theorem 28.8. *The relation of "having the same cardinality" as given in* Definition 28.1 *is an equivalence relation on the collection of sets.*

Proof. We first prove this relation is reflexive. Let X be any set. The identity function $\mathrm{id}_X \colon X \to X$ is a bijection. Thus X is related to X.

Next, we prove this relation is symmetric. Let X and Y be any sets, and assume X relates to Y. In other words, assume there is a bijection $f \colon X \to Y$. Then f has an inverse function $f^{-1} \colon Y \to X$ which is also a bijection. Hence Y relates to X.

Finally, we prove transitivity. Let X, Y, and Z be any sets, and assume there are bijections $f \colon X \to Y$ and $g \colon Y \to Z$. The composite function $g \circ f \colon X \to Z$ is a bijection, as needed. □

The equivalence classes of this equivalence relation are *precisely* the collections of sets with the same cardinality.

The observant reader will have noticed that we defined when two sets A and B have the same cardinality, $|A| = |B|$, but that we *have not* defined what the cardinality of an individual set is. Mathematicians solve this problem by choosing a (special) transversal of this equivalence relation; the representatives in the transversal are the *cardinal numbers*. Thus, the cardinality of a set, denoted $|A|$, is a special element of the equivalence class of A under the relation "having the same cardinality." There are specific symbols used to represent the cardinality of a set. For finite sets, that symbol is just the *actual* size of the set. Thus, we still have $|\{2, 79, -4\}| = 3$.

For infinite sets things are much more complicated. (Did you expect otherwise?) The smallest infinite cardinal $|\mathbb{N}|$ is written as $\aleph_0$ (read as "aleph-nought"). The next infinite cardinal is $\aleph_1$, and so forth. The diagram below gives some perspective to this chain. (We put question marks in places where we do not yet have any examples.)

28.C Definition of countable sets

The easiest infinite set to understand is the set of natural numbers. Its cardinality is given a special symbol $|\mathbb{N}| = \aleph_0$. We think of the natural numbers as "counting numbers" which motivates the following two definition.

Definition 28.9. Given a set A, we say that A is *countably infinite* if $|A| = |\mathbb{N}|$.

Definition 28.10. Given a set A, we say that A is *countable* if it is either finite or countably infinite.

The following are some examples and nonexamples involving these definitions.

Example 28.11. (1) The empty set is countable, since it is finite. It is not countably infinite (since it isn't infinite).

(2) The set $\{1, 2, 93828283928\}$ is countable and finite, but not infinite, and hence not countably infinite.

(3) Theorem 28.4 tells us that the integers are a countably infinite set. Similarly, Example 28.3 tells us that $2\mathbb{N}$ is a countably infinite set.

(4) The set of integer square, $\{n^2 : n \in \mathbb{Z}\}$, is infinite; we will see shortly that it is countably infinite.

(5) Are there any sets which are infinite but not countably infinite? These would be sets which occur strictly above $\aleph_0$ in the diagram below. We will prove in Section 30 that, yes, there are such sets! △

Cardinalities		Examples
$\vdots$	Infinite	$\vdots$
$\aleph_2$		?
$\aleph_1$		?
$\aleph_0$		$\mathbb{N}, 2\mathbb{N}, \mathbb{Z}, \ldots$
$\vdots$	Finite	$\vdots$
2		$\{1,2\}, \{1,3\}, \ldots$
1		$\{1\}, \{2\}, \ldots$
0		$\emptyset$

Advice 28.12. To show that a set A is countably infinite, you just need to arrange its elements in a nonrepeating, infinite list

$$A = \{a_1, a_2, a_3, \ldots\}.$$

This is precisely what we did when we proved that $\mathbb{Z}$ is countably infinite, we put its elements in the list $0, 1, -1, 2, -2, 3, -3, \ldots$.

Warning 28.13. If you are proving that a set is countably infinite by putting its elements into a list, then do not skip elements and do not repeat elements. Otherwise, you didn't really create a bijection.

Question: Which of the following lists proves that $\mathbb{Z}$ is countably infinite?

(a) $\{0, 1, 2, 3, \ldots\}$.
(b) $\{0, 1, 0, -1, 0, 2, 0, -2, \ldots\}$.
(c) $\{1, 0, 2, -1, 3, -2, 4, -3, \ldots\}$.

Answer: Only (c) works. The list in (a) skips the negative integers. (However, it does prove that the nonnegative integers are countably infinite.) The list in (b) repeats 0. Of the choices, only (c) lists every integer exactly once, hence gives the bijection with $\mathbb{N}$.

28.D Subsets of countable sets

Many operations involving countable sets yield new countable sets. One of the most useful results is the following:

Theorem 28.14. *Any infinite subset of a countably infinite set is countably infinite.*

Proof. Let A be a countably infinite set. We can write the elements of A in an infinite list $a_1, a_2, a_3, \ldots$. Let B be an infinite subset of A.

Let n_1 be the smallest natural number with $a_{n_1} \in B$, which exists since $B \neq \emptyset$ as B is infinite. Put $b_1 = a_{n_1}$.

Next let n_2 be the smallest natural number with $n_2 > n_1$ and $a_{n_2} \in B$, which exists since $B - \{b_1\} \neq \emptyset$ as B is infinite. Put $b_2 = a_{n_2}$.

Repeating this process (by induction) we create an infinite list $b_1, b_2, \ldots$. Clearly there are no repetitions in this list. This new list covers every element of B because we can also prove (by induction) that $n_i \geq i$ for each $i \in \mathbb{N}$; hence, we have worked all the way through the list of elements of A. □

Corollary 28.15. *Any subset of a countable set is countable.*

Proof. We leave this as Exercise 28.7. □

Example 28.16. Not every subset of $\mathbb{N}$ is countably infinite. For instance $\{3, 7, 19\}$ is a subset but not countably infinite.

However, every *infinite* subset of $\mathbb{N}$ is countably infinite by Theorem 28.14. For instance, since there are infinitely many primes (by Theorem 19.14), then we know that the set of all primes

$$\{2, 3, 5, 7, 11, 13, 17, \ldots\}$$

is countably infinite.

Is $S = \{x^3 : x \in \mathbb{Z}\} = \{\ldots, -27, -8, -1, 0, 1, 8, 27, 64, \ldots\}$ countably infinite? Yes! First, it is an infinite set since as you cube larger and larger integers, you get infinitely many different cubes (because the cubing function is strictly increasing). As S is an infinite subset of the countably infinite set $\mathbb{Z}$, we know S is countably infinite by Theorem 28.14. △

Example 28.17. Is $|2\mathbb{Z}| = \aleph_0$? Yes, since $2\mathbb{Z}$ is an infinite subset of the countably infinite set $\mathbb{Z}$.

Alternatively, you could create an explicit bijection $f\colon \mathbb{N} \to 2\mathbb{Z}$, although this is more difficult. We see that we can list the elements of $2\mathbb{Z}$ as $0, 2, -2, 4, -4, 6, -6, \ldots$. From this pattern, one possible bijection is given by the rule

$$f(n) = \frac{1 + (-1)^n(2n-1)}{2}.$$ △

28.E Exercises

Exercise 28.1. Indicate whether the following statements are true or false, with proof/reason or counterexample:

(a) All finite sets have the same cardinality.
(b) If $f\colon A \to B$ is a function between two sets, then $|f| = |A|$ (thinking of f as a set of ordered pairs).
(c) Every subset of $\mathbb{N}$ is countably infinite.
(d) Every subset of an infinite set has cardinality $\aleph_0$.
(e) If $f\colon A \to B$ is a surjective function then $|f| = |B|$ (thinking of f as a set of ordered pairs).

Exercise 28.2. Prove that the set of those natural numbers with exactly one digit equal to 7 is countably infinite. For instance, the number 103792 has exactly one of its digits equal to 7, while 8772 has two digits equal to seven.

Exercise 28.3. Consider the set

$$S = \{x \in \mathbb{Z} : x = a^2 + b^2 \text{ for some } a, b \in \mathbb{Z}\}.$$

Prove that $|S| = |\mathbb{N}|$.

Exercise 28.4. Define $h\colon (0, \infty) \to (0, 1)$ by the rule

$$h(x) = \frac{x}{x+1}.$$

Verify that h is a function by showing that for any $x \in (0, \infty)$ we have $\frac{x}{x+1} \in (0, 1)$. Then prove that h is a bijection. What does this say about the cardinality of these open intervals?

Exercise 28.5. Adjust the formula in Exercise 28.4 to define a bijective function $j\colon (-\infty, 0) \to (-1, 0)$. Verify that your new rule j defines a function to the given codomain and that it is a bijection.

Exercise 28.6. Prove that $|\mathbb{R}| = |(0, 1)|$. (Hint: Use the previous two exercises to define a piecewise bijection from $\mathbb{R}$ to $(-1, 1)$. Then find a bijection from $(-1, 1)$ to $(0, 1)$ and compose the two bijections.)

Exercise 28.7. Prove Corollary 28.15.

Exercise 28.8. Prove that a set A is countable if and only if there exists an injective function $f\colon A \to \mathbb{N}$. (This exercise may be useful for future exercises.)

29 More examples of countable sets

Usually a combination of two countable sets is countable. For instance, in this section we will prove that unions of countable sets are countable, as well as Cartesian products. Finally, we will show that $\mathbb{Q}$ is countably infinite.

29.A Unions

Theorem 29.1. *If A and B are countable sets, then $A \cup B$ is countable.*

Proof. Let $B' = B - A$. Notice that $A \cup B = A \cup B'$. Also since B is countable so is B', by Corollary 28.15. Thus, to prove the theorem, it suffices to work with B' in place of B. The benefit of working with B' instead of B is that $A \cap B' = A \cap (B - A) = \emptyset$; i.e., A and B' are disjoint.

We will handle the case when A and B' are both countably infinite, leaving the other cases as Exercise 29.1.

So assume that A and B' are countably infinite. Thus, we can write the elements of A in an infinite list $a_1, a_2, a_3, \ldots$. Similarly list the elements of B' as $b_1, b_2, b_3, \ldots$. We need to list the elements of $A \cup B'$ in an infinite list, which is easy to do by interlacing the two lists, as

$$a_1, b_1, a_2, b_2, a_3, b_3, \ldots$$

There is no repetition in this list, since $A \cap B' = \emptyset$ and the two original lists have no repetitions. Also this new list contains each of the elements of $A \cup B' = A \cup B$. □

Advice 29.2. The type of argument used in the first paragraph of the proof above is referred to as *reducing* to a simpler situation. For instance, in the proof above we could say there that we reduced to the case where the two sets are disjoint.

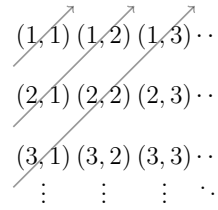
After a reduction, mathematicians will simply assert that they now need only consider the simpler situation. For example, after the first paragraph of the proof above, we could simply say "We thus may assume $A \cap B = \emptyset$." This is because we would recognize that, after replacing B by B', this situation actually occurs.

29.B Products

Taking a union is not the only operation we can do with two sets. Another operation is intersection. When we intersect two sets, the cardinality can get much smaller. There is a third operation: the Cartesian product. Cantor came up with a very clever method for showing that the product of two countably infinite sets is still countably infinite. Thus we have:

Theorem 29.3. *If A and B are countably infinite sets, then $A \times B$ is as well.*

Proof. Without loss of generality, we just need to show that $\mathbb{N} \times \mathbb{N}$ is countably infinite. Consider the following diagram:



Travel along each arrow, starting at the smallest arrow, and passing to the next smallest arrow. This allows us to list the elements of $\mathbb{N} \times \mathbb{N}$ as

$$(1,1), (2,1), (1,2), (3,1), (2,2), (1,3), \ldots,$$

according to when we pass through each ordered pair. We will hit each ordered pair exactly once. □

Remark 29.4. The bijection $f\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ described in Theorem 29.3 was only described implicitly because the *idea* of the proof is much more important than the details. However, we can explicitly describe the function. It is given by the rule

$$f(m,n) = \frac{(m+n-1)(m+n-2)}{2} + n.$$

The first term $\frac{(m+n-1)(m+n-2)}{2}$ is the size of the triangle covered by the previous arrows, and the last term n counts how far along the current arrow we have travelled. (The details of proving that f is a bijection are left to the motivated reader!)

There are other ways of showing $|\mathbb{N} \times \mathbb{N}| = |\mathbb{N}|$. For instance, we could show that the function $g\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ given by the rule

$$g(m,n) = 2^{m-1}(2n-1)$$

is also a bijection. (Proving that g is a bijection requires the lemma that every natural number can be written as a unique power of 2 times a unique odd integer; for existence, see Exercise 15.4.)

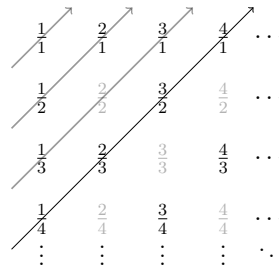
There are *many* other options. For instance, we could have used arrows pointing down and to the left, instead of up and to the right. Alternatively, we could have "snaked" back and forth along each finite diagonal. ▲

29.C Rational numbers

We are almost ready to describe the cardinality of $\mathbb{Q}$. First, let's deal with the positive rational numbers $\mathbb{Q}_{>0}$.

Theorem 29.5. *The set $\mathbb{Q}_{>0}$ is countably infinite.*

Proof. Put the elements of $\mathbb{Q}_{>0}$ into a diagram as below. (We put fractions which are not in lowest terms as light gray.)



Now, we just list the elements as before, skipping over the elements in light gray, since they will be counted when they are in lowest terms. This counting procedure never repeats elements (since we skip those fractions not in lowest terms), and continues forever since $\mathbb{Q}_{>0}$ is infinite (since $\mathbb{N}$ is a subset; in other words, the top row of the diagram is infinite). □

Remark 29.6. Here is an alternative proof of Theorem 29.5. Define a function $f\colon \mathbb{Q}_{>0} \to \mathbb{N} \times \mathbb{N}$ by sending a rational number a/b (written in lowest terms) to the ordered pair (a, b). The function f is injective, but not surjective. The set $\mathrm{im}(f)$ is infinite, and a subset of the countably infinite set $\mathbb{N} \times \mathbb{N}$. Thus $\mathrm{im}(f)$ is countably infinite. But f is a bijection from $\mathbb{Q}_{>0}$ to $\mathrm{im}(f)$, hence $\mathbb{Q}_{>0}$ is countably infinite. ▲

Corollary 29.7. *The set $\mathbb{Q}$ is countably infinite.*

Proof sketch. We have $\mathbb{Q} = \mathbb{Q}_{>0} \cup \{0\} \cup \mathbb{Q}_{<0}$. We know that $\mathbb{Q}_{>0}$ is countable by Theorem 29.5. Also, $\mathbb{Q}_{<0}$ is countable since there is an easy bijection $\mathbb{Q}_{>0} \to \mathbb{Q}_{<0}$. Also, $\{0\}$ is finite, hence countable. By Theorem 29.1, the union of two countable sets is countable, so $\mathbb{Q}_{>0} \cup \{0\}$ is countable. Applying Theorem 29.1 again, we find that $\mathbb{Q} = (\mathbb{Q}_{>0} \cup \{0\}) \cup \mathbb{Q}_{<0}$ is countable. It is also infinite, hence countably infinite. □

We finish with one more example of how to show a set is countably infinite.

Example 29.8. Let $S = \{(i, j) \in \mathbb{N} \times \mathbb{N} : i \geq j\}$. This set is pictured below. We will prove that S is countably infinite.

$$
\begin{array}{ccccccc}
(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & \cdots \\
(2,1) & (2,2) & (2,3) & (2,4) & (2,5) & \cdots \\
(3,1) & (3,2) & (3,3) & (3,4) & (3,5) & \cdots \\
(4,1) & (4,2) & (4,3) & (4,4) & (4,5) & \cdots \\
(5,1) & (5,2) & (5,3) & (5,4) & (5,5) & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}
$$

First, the set S is infinite, since the left column is infinite. Since $S \subseteq \mathbb{N} \times \mathbb{N}$ and $\mathbb{N} \times \mathbb{N}$ is countably infinite, we know that S is countably infinite by Theorem 28.14.

Alternatively, we can list the elements of S by using the "up arrow" argument from earlier. (We can't list the elements of S by going down columns, but could we list the elements of S by traveling across the successive rows?) △

29.D Exercises

Exercise 29.1. Finish the proof of Theorem 29.1.
(Hint: There are two unfinished cases: (a) both A and B' are finite, or (b) one of them is finite and the other infinite.)

Exercise 29.2. Prove that $\{0,1\} \times \mathbb{N}$ is countably infinite. (Hint: Use theorems in the section.)

Exercise 29.3. Let A and B be countable sets. Prove that $A \times B$ is countable. (How is this different from what was proved in Theorem 29.3?)

Exercise 29.4. Let $A_1, A_2, \ldots$ be arbitrary sets. Let $B_1 = A_1$, and for each $n \geq 1$ let $B_{n+1} = B_n \times A_{n+1}$. Hence, we have

$$
\begin{array}{ll}
B_1 = A_1, & B_2 = A_1 \times A_2, \\
B_3 = (A_1 \times A_2) \times A_3, & B_4 = ((A_1 \times A_2) \times A_3) \times A_4,
\end{array}
$$

and so forth. Do the following:

(a) Prove that if all of the A_i are countable, then B_n is countable for each $n \in \mathbb{N}$. (Hint: Induction.)

(b) Recall that the set $A_1 \times A_2 \times A_3$ (without parentheses) consists of *ordered triples* of the form (a_1, a_2, a_3) where each $a_i \in A_i$. On the other hand, the set $B_3 = (A_1 \times A_2) \times A_3$ (with parentheses) consists of elements of the form $((a_1, a_2), a_3)$ where each $a_i \in A_i$. These elements are ordered pairs where the first coordinate is also an ordered pair.
Describe a bijective function $f\colon A_1 \times A_2 \times A_3 \to (A_1 \times A_2) \times A_3 = B_3$, and prove its bijectivity.

(c) Prove that if A_1, A_2, A_3 are countable, then the set $A_1 \times A_2 \times A_3$ is countable.

(The motivated student might attempt generalizing parts (b) and (c) above, by proving that there is a bijection from the set of *ordered n-tuples* $A_1 \times A_2 \times \cdots \times A_n$ to the set B_n. This shows that if $A_1, A_2, \ldots, A_n$ are countable, then the set of ordered n-tuples $A_1 \times A_2 \times \cdots \times A_n$ is countable.)

Exercise 29.5. Prove that $|\mathbb{Z} \times \mathbb{N}| = |\mathbb{Q}|$.

Exercise 29.6. Prove that if $A_1, A_2, \ldots$ are pairwise disjoint, countably infinite sets, then $\bigcup_{i=1}^{\infty} A_i$ is countably infinite. (Hint: Not induction. Think about "up arrow" arguments. Alternatively, you could construct a bijection from $\mathbb{N} \times \mathbb{N}$ to $\bigcup_{i=1}^{\infty} A_i$.)

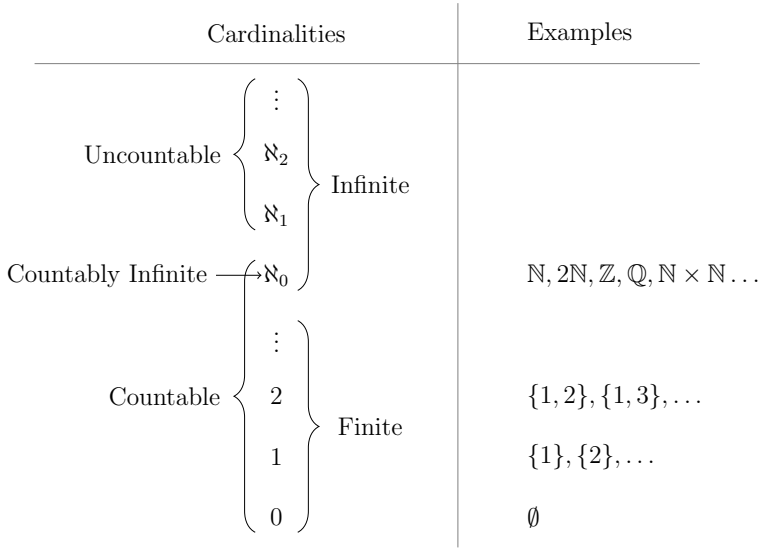
Exercise 29.7. Prove that the set of all *finite* subsets of $\mathbb{N}$ is countably infinite.

30 Uncountable sets

The results of this section will be centered around the following definition.

Definition 30.1. Any set A that is not countable is said to be *uncountable*.

We can think of the uncountable sets as those sets which are *bigger* than the countably infinite sets, as in the following figure.



As is evident from this diagram, we still don't have any examples of uncountable sets. In this section we will see that there are many examples.

30.A How big is $\mathbb{R}$?

Before we can talk about how big the set of real numbers is, we need to explain more precisely *what* a real number is. We usually think of real numbers as infinite decimal expansions. For instance:

$$\begin{aligned} 1 &= 1.00000\ldots \\ \sqrt{2} &= 1.41421\ldots \\ -\frac{\pi}{13} &= -0.24166\ldots \\ e^{-6} - \frac{24}{7} &= -3.42609\ldots \end{aligned}$$

However, real numbers do not always have unique infinite decimal expansions. If a

number ends in repeating 9's, we can shift up and end in repeating 0's. For example,

$$\begin{aligned} 0.99999\ldots &= 1.00000\ldots \\ 8.3929999\ldots &= 8.3930000\ldots \\ -3928.83829999\ldots &= -3928.83830000\ldots \end{aligned}$$

To avoid nonuniqueness issues, we will always avoid writing decimal expansions which end in repeating 9's.

Our goal now is to show that $\mathbb{R}$ is uncountable. From a previous homework problem we know that $|(0,1)| = |\mathbb{R}|$, so it suffices to show that $(0,1)$ is uncountable. (This set is easier to work with.) We know that $(0,1)$ is infinite, so to prove that it is uncountable we must show that there does *not exist* any bijection $f\colon \mathbb{N} \to (0,1)$. Cantor's trick to do this is to show that *every* function $f\colon \mathbb{N} \to (0,1)$ is not surjective, using what is now commonly called a "diagonalization argument." Before we give the technical proof, we demonstrate the idea with an example.

Suppose $f\colon \mathbb{N} \to (0,1)$ is the function

$$\begin{aligned} f(1) &= 0.29838293\ldots \\ f(2) &= 0.43828183\ldots \\ f(3) &= 0.73826261\ldots \\ f(4) &= 0.20030000\ldots \\ f(5) &= 0.73724892\ldots \\ &\vdots \end{aligned}$$

Our goal is to prove that f is not surjective. Thus, we must find some element $x \in (0,1)$ that is not in the image of f. We will construct x digit by digit, so that it doesn't match any of the numbers on our list.

First, we want x to be different from $f(1) = 0.29838293\ldots$. We can make sure this is true by guaranteeing that the first digit (past the decimal point) of x is different from the first digit of $f(1)$. So, let's change that first 2 to a 4, and put

$$x = 0.4\ldots.$$

Notice that no matter what we do with the rest of the digits of x, it will *not* match $f(1)$.

Second, we want x to be different from $f(2) = 0.43828183\ldots$. They do match on their first digit, but we can make their second digits different by changing the 3 to a 4. So we put

$$x = 0.44\ldots$$

and it will not equal $f(1)$ or $f(2)$.

Third, we want x to be different from $f(3) = 0.73826261\ldots$. It already is different because of our choice of the first two digits, but we probably should continue the pattern we've already come up with, to make sure that the *third* digit is different. So we change the 8 to a 4, and put

$$x = 0.444\ldots$$

and we have $x \neq f(1), f(2), f(3)$.

For the fourth number, $f(4) = 0.20030000\ldots$, we change the 3 to a 4, and put

$$x = 0.4444\ldots$$

and we have $x \neq f(1), f(2), f(3), f(4)$.

For the fifth number, $f(5) = 0.73724892\ldots$, we need to change the fifth digit 4, so we change it to 7. Put

$$x = 0.44447\ldots.$$

In general, we look at the nth digit of $f(n)$, and change it to a 4, unless it is already a 4, in which case we change it to a 7. We place that new digit in the appropriate place in x. After all of these changes, x cannot match any number on our list, so f is not surjective.

Remark 30.2. If we start with a different list of numbers, the number x we construct will be different (depending on that list). ▲

To make this work more easily, define the *digit change* function

$$\operatorname{dig}\colon \{0,1,2,3,4,5,6,7,8,9\} \to \{4,7\}$$

by the rule

$$\operatorname{dig}(i) = \begin{cases} 4 & \text{if } i \neq 4 \\ 7 & \text{if } i = 4. \end{cases}$$

Note that because the digit change function does not use 9's, we don't need to worry about x ending in repeating 9's.

Warning 30.3. There are many other digit change functions we could have used. This is just one option. Because there are so many different options, you should *always* tell your reader which digit change function you are using by giving the definition.

We are now prepared to give the formal proof that $(0,1)$ is uncountable. As discussed above, the technique used in this proof is known as *Cantor's diagonalization argument.*

Theorem 30.4. *The set* $(0,1)$ *is uncountable.*

Proof. Let $f\colon \mathbb{N} \to (0,1)$ be *any* function. We will show that f is not surjective.

For each $n \in \mathbb{N}$, write $f(n)$ using a decimal expansion $f(n) = 0.d_{1,n}d_{2,n}d_{3,n}\ldots$ (which doesn't end in repeating 9's). Fix $x \in (0,1)$ to be the number with decimal expansion $x = 0.x_1x_2x_3\ldots$ where $x_n = \operatorname{dig}(d_{n,n})$. In other words, the nth digit of x is the digit change of the nth digit of $f(n)$. Hence $x \neq f(n)$ for each $n \in \mathbb{N}$. Therefore f is not surjective, as x is not in the image. □

Corollary 30.5. *The set $\mathbb{R}$ is uncountable.*

The cardinality of $\mathbb{R}$ is called the *continuum*, and we write $|\mathbb{R}| = \mathfrak{c}$. You might ask: Where does $\mathfrak{c}$ fit in the chain of cardinalities? Is it just one step up from $\aleph_0$?

The answer is strange. It depends on the axioms you use! Some mathematicians do assume $\mathfrak{c} = \aleph_1$; this assumption is called the *continuum hypothesis*. Most mathematicians simply do not worry about this question.

We have already seen that $|(0,1)| = |\mathbb{R}|$, so $(0,1)$ also has continuum cardinality. Here are some more examples of sets with continuum cardinality.

(1) Any open interval (a,b) with $a,b \in \mathbb{R}$. (We can also replace a with $-\infty$, or b with ∞.)
(2) Any half-open interval $[a,b)$ with $a,b \in \mathbb{R}$. (We can replace b with ∞.)
(3) Any half-open interval $(a,b]$ with $a,b \in \mathbb{R}$. (We can replace a with $-\infty$.)
(4) Any closed interval $[a,b]$ with $a,b \in \mathbb{R}$.

To give the idea behind how to prove these facts, we will show:

Proposition 30.6. *The half open interval $(0,1]$ has the same cardinality as the open interval $(0,1)$, and hence it has continuum cardinality.*

Proof. We define a bijection $f\colon (0,1) \to (0,1]$, as follows. Fix

$$S = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots\right\} = \left\{\frac{1}{2^n} : n \in \mathbb{N}\right\} \subsetneq (0,1),$$

and fix $T = S \cup \{1\} \subsetneq (0,1]$. Now, define f as a piecewise function by the rule

$$f(x) = \begin{cases} x & \text{if } x \notin S \\ 2x & \text{if } x \in S. \end{cases}$$

The first piece of this function is a bijection from $(0,1) - S$ to $(0,1] - T$, and the second piece is a bijection from S to T. By pasting together, f is a bijection from $(0,1)$ to $(0,1]$. □

We end this section with one last result which can be used to tell whether a set is uncountable.

Theorem 30.7. *Let A and B be sets, with $A \subseteq B$. If A is uncountable, then B is uncountable.*

Proof. This is the contrapositive of Theorem 28.14, after noting that A and B must be infinite. □

Example 30.8. Any subset $S \subseteq \mathbb{R}$ such that $(0,1) \subseteq S$ is uncountable, by the previous theorem. In a later section we will show that any such subset has continuum cardinality.

There are lots of subsets of $\mathbb{R}$ which are not uncountable. Can you list some? △

30.B Exercises

Exercise 30.1. Let $a, b \in \mathbb{R}$ with $a < b$. Construct a bijection $f\colon (0,1) \to (a,b)$, and prove it is a bijection. (This shows that bounded open intervals all have the same cardinality.)

Exercise 30.2. Prove that the interval $[0,1)$ has continuum cardinality, by creating a bijection $[0,1) \to (0,1)$.

Exercise 30.3. Prove that the interval $[0,1]$ has continuum cardinality.

Exercise 30.4. Prove that the set of irrational numbers is uncountable. (Hint: Theorem 29.1 may be useful, along with contradiction.) Find a subset of the irrational numbers that is countably infinite.

Exercise 30.5. Prove that the set $\mathbb{C}$ of complex numbers is uncountable. (With more work one can prove that $\mathbb{C}$ has continuum cardinality, but this is not obvious.)

Exercise 30.6. We defined a product of two sets A and B to be the collection of ordered pairs from A and B.

Let $A_1, A_2, A_3, \ldots$ be sets. Define the product $\prod_{i=1}^{\infty} A_i = A_1 \times A_2 \times A_3 \times \cdots$ to be the set of ordered sequences (i.e., infinite tuples)

$$\{(a_1, a_2, a_3, \ldots) \,:\, a_i \in A_i \text{ for each integer } i \geq 1\}.$$

We showed previously that a *finite* product of countable sets is countable. Show that the countable product $\prod_{i=1}^{\infty} \{0,1\} = \{0,1\} \times \{0,1\} \times \{0,1\} \times \cdots$ is not countable. (Hint: (1) This product is the set of infinite sequences of 0's and 1's. (2) Use Cantor's diagonalization argument.)

31 Injections and cardinalities

In the previous section we proved the amazing fact that $\mathbb{R}$ is not a countable set. Thus, we might expect that $|\mathbb{N}| < |\mathbb{R}|$. We have only defined when cardinalities are *equal*. In this section we give a method for determining inequalities between cardinalities.

31.A Injections vs. bijections

Let A and B be arbitrary sets. We would like to think of A as "smaller" than B if we can fit A inside B. However, consider the set $\mathbb{N}$ which sits properly inside $\mathbb{Z}$. These sets have the same cardinality! Thus, we have to be extra careful about whether cardinalities are strictly smaller or not.

One way to think about fitting A inside another set B is to use an injection. This motivates the following definitions (which will only be fully justified in the next section).

Definition 31.1. Let A and B be sets. If there is an injection $f\colon A \to B$, we write $|A| \leq |B|$. If there is an injection but no bijection from A to B, then we write $|A| < |B|$.

The following are some examples of these definitions in action.

Example 31.2. (1) What is the relation between the sets $A = \{1, 3, 5\}$ and $B = \{2, 4, 6, 8\}$? There is an injection from A to B but no bijection (since A has three elements, but B has four elements). Thus $|A| < |B|$.

(2) What is the relation between the sets $2\mathbb{N}$ and $\mathbb{N}$? The function $f\colon 2\mathbb{N} \to \mathbb{N}$ defined by the rule $f(x) = x$ is injective. Hence, $|2\mathbb{N}| \leq |\mathbb{N}|$. However, both sets are countably infinite, so in fact there is a (different) bijective function between the sets, so we have $|2\mathbb{N}| = |\mathbb{N}|$.

(3) What is the relation between the sets $\mathbb{N}$ and $\mathbb{R}$? The function $\mathbb{N} \to \mathbb{R}$ sending each natural number to itself is an injective function. Hence $|\mathbb{N}| \leq |\mathbb{R}|$. In the previous section we proved that $\mathbb{R}$ is uncountable, so there is no bijection between these sets. Hence, we have the strict inequality $|\mathbb{N}| < |\mathbb{R}|$. △

Advice 31.3. We can think of injections as giving only "half" of the information needed to construct a bijection, which is why we only get an inequality $\leq$.

You might recall that in our tower of cardinalities (found at the beginning of the previous section) we had an infinite list of infinite cardinalities $\aleph_0 < \aleph_1 < \aleph_2 < \ldots$. But so far, we have only found two types of infinite cardinalities; the countably infinite sets, and the sets of continuum size. In our next theorem we will prove that for any set A we have $|A| < |\mathscr{P}(A)|$. Thus, we have an infinite chain of increasing infinite cardinalities

$$|\mathbb{N}| < |\mathscr{P}(\mathbb{N})| < |\mathscr{P}(\mathscr{P}(\mathbb{N}))| < \ldots.$$

When A is a finite set, say $|A| = n$, then we know $|A| < |\mathscr{P}(A)|$ because $n < 2^n$. But how does this process work when A is an infinite set? In that situation we cannot simply count elements. Rather, we must prove that there is no bijection $g\colon A \to \mathscr{P}(A)$. Our approach will be similar to how we showed $\mathbb{R}$ is not countable. Start with an arbitrary function $g\colon A \to \mathscr{P}(A)$, and show that g is not surjective by finding some set $B \in \mathscr{P}(A)$ which is not in the image of g. Note that B will be a subset of A, since $B \in \mathscr{P}(A)$. The hardest part is constructing B. We will give an explicit example (using finite sets), and then give the formal proof for arbitrary sets.

Fix $A = \{1, 2, 3\}$. Hence

$$\mathscr{P}(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

Consider the function $g\colon A \to \mathscr{P}(A)$ where $g(1) = \{2\}$, $g(2) = \emptyset$, and $g(3) = \{1, 2, 3\}$. We want to find a set $B \in \mathscr{P}(A)$ that we can prove is not equal to $g(1)$, $g(2)$, or $g(3)$. Of course, we could just pick one of the other five sets not in the image of g in this case; but we want to come up with a method that will work for any set A.

So, we ask the question: Is $x \in g(x)$?

- Is $1 \in g(1) = \{2\}$? The answer is no.
- Is $2 \in g(2) = \emptyset$? The answer is no.
- Is $3 \in g(3) = \{1, 2, 3\}$? The answer is yes.

We construct $B \in \mathscr{P}(A)$ by the following rule: If the answer to the question "Is $x \in g(x)$?" is no then we put $x \in B$, but if the answer is yes then we leave x out of B. Using the answers we had above, we see that $B = \{1, 2\}$.

Notice that B will not equal $g(x)$ because if $x \in g(x)$ then $x \notin B$, and vice versa. Indeed, we see that

- $1 \notin g(1)$ but $1 \in B$.
- $2 \notin g(2)$ but $2 \in B$.
- $3 \in g(3)$ but $3 \notin B$.

This forces B to be different from any element in the image.

One more example is in order, to test our understanding. Suppose that $A = \{1, 2, 3\}$ as above, and suppose $h\colon A \to \mathscr{P}(A)$ is the function defined by the rule $h(1) = \{2, 3\}$, $h(2) = \{2\}$, $h(3) = \{2, 3\}$. If we follow the same pattern as above, asking the question "Is $x \in h(x)$?" and using the answers to define B, what set B do we get? (Before looking at the answer, try this construction yourself.)

Answer: The set is $\{1\}$.

Remark 31.4. The set B is sometimes called the *barber set*. This is because there is some connection with the following paradox: There lives a barber in a small town who always obeys the rule that he will shave everyone in town who doesn't shave themselves, but if they shave themselves he will not shave them.

Does the barber shave himself? If he does, then he cannot shave himself by his own rule. But if he doesn't, then he must shave himself by his rule.

One way to resolve the paradox is to assume instead the barber does not live in the town. This corresponds, roughly, to the fact that B is not in the image of the function. ▲

We are now ready to prove the theorem in general.

Theorem 31.5. *If A is a set, then $|A| < |\mathscr{P}(A)|$.*

Proof. Let A be any set. First, we prove that $|A| \leq |\mathscr{P}(A)|$, so we need to find an injective function $f\colon A \to \mathscr{P}(A)$. Define f by the rule $f(a) = \{a\}$. To prove that f is injective, let $a_1, a_2 \in A$ and assume $f(a_1) = f(a_2)$. Hence $\{a_1\} = \{a_2\}$. Therefore $a_1 = a_2$, since sets are equal exactly when they have the same elements.

Next, let $g\colon A \to \mathscr{P}(A)$ be an arbitrary function. We will show that g is never surjective, and hence there is no bijection between A and $\mathscr{P}(A)$. Define the (barber) set as

$$B = \{x \in A \,:\, x \notin g(x)\}.$$

This is a subset of A, hence an element of $\mathscr{P}(A)$. We will show that B is not in the image of g.

Let $a \in A$ be arbitrary. There are two cases.

Case 1: Assume $a \in g(a)$. In this case $a \notin B$, hence $g(a) \neq B$.

Case 2: Assume $a \notin g(a)$. Then $a \in B$, hence $g(a) \neq B$.

In every case $B \neq g(a)$. Since $a \in A$ was arbitrary, this means B cannot be in the image of g (since it does not equal any element of the image of g). □

31.B How big is $\mathscr{P}(\mathbb{N})$?

We now know that $|\mathbb{N}| < |\mathscr{P}(\mathbb{N})|$, but exactly how big is $\mathscr{P}(\mathbb{N})$? In the next section we will prove it has continuum cardinality. However, we currently have the tools to find another set with the same cardinality. For the rest of this section, given a set A we let $\mathcal{F}(A)$ be the set of all functions from A to $\{0, 1\}$.

Theorem 31.6. *If A is any set, then $|\mathscr{P}(A)| = |\mathcal{F}(A)|$.*

Proof. We must construct a bijection between the two sets $\mathscr{P}(A)$ and $\mathcal{F}(A)$. The rule is this: send a subset $S \subseteq A$ to its characteristic function, $\chi_S\colon A \to \{0, 1\}$. In other words, we define $f\colon \mathscr{P}(A) \to \mathcal{F}(A)$ by the rule

$$f(S) = \chi_S.$$

We first show that f is injective. Let $S, T \in \mathscr{P}(A)$ and assume $f(S) = f(T)$. We then have $\chi_S = \chi_T$. Plugging in an arbitrary element $a \in A$, we have $\chi_S(a) = \chi_T(a)$. The left-hand side is 1 when $a \in S$ and 0 otherwise, and similarly for the right-hand side. Thus, $a \in S$ if and only if $a \in T$. In other words $S = T$.

Finally, we show that f is surjective. Let $\varphi\colon A \to \{0, 1\}$ be any function. Put $S = \{a \in A \,:\, \varphi(a) = 1\}$. We then check directly that $\varphi = \chi_S$. (They have the same domain and codomain, and the same rule.) Hence $\varphi = f(S)$, so f is surjective. □

Corollary 31.7. *The set $\mathcal{F}(\mathbb{N})$ is uncountable.*

Proof. The equality $|\mathscr{P}(\mathbb{N})| = |\mathcal{F}(\mathbb{N})|$ follows from the previous theorem. We also know $|\mathbb{N}| < |\mathscr{P}(\mathbb{N})|$, hence $\mathcal{F}(\mathbb{N})$ is uncountable. □

Remark 31.8. In Exercise 30.6, you proved that $\prod_{i=1}^{\infty}\{0,1\}$ is uncountable. The corollary above gives an easier way to see this, since functions $\mathbb{N} \to \{0,1\}$ can be thought of as infinite sequences of 0's and 1's. ▲

31.C Exercises

Exercise 31.1. Answer each of the following true or false problems, proving your answer.

(a) Every uncountable set has the same cardinality as $(0,1)$.
(b) Let A and B be sets. If $A \subseteq B$, then $|A| \leq |B|$.
(c) For sets A and B, if $A \subsetneq B$, then $|A| < |B|$.
(d) Given sets A, B, and C, if $A \subseteq B \subseteq C$ and both A and C are countably infinite, then B is countably infinite.
(e) No subset of $\mathbb{R}$ has smaller cardinality than $\mathbb{R}$.
(f) For sets A and B, if $|A| < |B|$ and A is finite, then B is infinite.
(g) For sets A and B, if $|A| < |B|$ and A is countable, then B is uncountable.
(h) For sets A and B, if $|A| < |B|$ and A is countably infinite, then B is uncountable.
(i) For any set A, there exists another set B such that $|A| < |B|$.

Exercise 31.2. Let $A = \{a,b,c,d,e\}$ and let $g\colon A \to \mathscr{P}(A)$ be defined by the rule $g(a) = \{b,d\}$, $g(b) = \{a,c,e\}$, $g(c) = \{a,c,d,e\}$, $g(d) = \emptyset$, $g(e) = \{e\}$. List the elements of the barber set $B = \{x \in A : x \notin g(x)\}$. Why is it not in the image of g?

Exercise 31.3. Find a set with cardinality bigger than that of $\mathbb{R}$. Then find a set with cardinality bigger than that.

Exercise 31.4. Theorem 27.5 says that for *finite* sets A and B, if $|A| = |B|$ and $f\colon A \to B$ is a function, then f is injective if and only if f is surjective. Prove that this fails for *infinite* sets, by proving the following:

(a) There is an infinite set S and a function $f\colon S \to S$ that is injective but not surjective.
(b) There is an infinite set S and a function $g\colon S \to S$ that is surjective but not injective.

In both parts prove that the function you construct has the required properties.

Exercise 31.5. Let A and B be sets with $f\colon A \to B$ a bijection. Define a new function $g\colon \mathscr{P}(A) \to \mathscr{P}(B)$ by the rule $g(S) = \{f(s) : s \in S\}$, where $S \subseteq A$ is an arbitrary element of $\mathscr{P}(A)$. Prove that g is a bijection.

Conclude that if $|A| = |B|$ then $|\mathscr{P}(A)| = |\mathscr{P}(B)|$.

Exercise 31.6. Define a function $f\colon \mathbb{R} \to \mathscr{P}(\mathbb{Q})$ by the rule

$$f(x) = \{q \in \mathbb{Q} \,:\, q \leq x\}.$$

Prove that f is injective. (Hint: For any two real numbers $x < y$, there is a rational number strictly between them. See Exercise 11.6.)

Using this injection, in conjunction with the previous exercise, derive the inequality $|\mathbb{R}| \leq |\mathscr{P}(\mathbb{N})|$.

Exercise 31.7. Let A and B be nonempty sets. Prove that there exists an injection $f\colon A \to B$ if and only if there exists a surjection $g\colon B \to A$. (Hint: For the backwards direction, given a surjection $g\colon B \to A$ define a function $f\colon A \to B$ by the rule $f(a) =$ one of the elements which mapped to a.)

32 The Schröder–Bernstein Theorem

Let $a, b \in \mathbb{R}$. If we know $a \leq b$ and $b \leq a$, then we must have $a = b$. In other words, the "less than or equal to" relation on the real numbers is antisymmetric.

We have used a similar notation for cardinalities, and it is natural to ask whether or not this relation is antisymmetric. In other words, given sets A and B, if we know $|A| \leq |B|$ and $|B| \leq |A|$, must we have $|A| = |B|$? The answer is yes, and this result is called the Schröder–Bernstein Theorem.

Theorem 32.1 (Schröder–Bernstein Theorem)**.** *Let A and B be arbitrary sets. If $|A| \leq |B|$ and $|B| \leq |A|$, then $|A| = |B|$.*

Remark 32.2. The story of how this theorem came to be proved is long and somewhat convoluted. Cantor was the first to state the theorem, but apparently he had no proof. The first proof (that we know of) was found by Dedekind, but he did not publish his work at the time.

Schröder announced a proof, which was later shown to have an error. Finally, in 1897, Bernstein (who was only 19 years old, and a student of Cantor) presented a proof. At nearly the same time Schröder independently found an error free proof as well. Hence, these two mathematicians have their names attached to the theorem. ▲

32.A Sketching the proof using genealogy

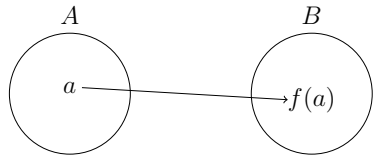
Here we will sketch the main ideas of the proof. The proof is a direct one. Given sets A and B, assume that $|A| \leq |B|$ and that $|B| \leq |A|$. In other words, we assume that we have two injective functions $f\colon A \to B$ and $g\colon B \to A$. Our goal is to show that $|A| = |B|$, in other words we need to create a bijection $h\colon A \to B$.

Of course, if f is already a bijection, we are done. So we may as well assume that f is not surjective. Similarly, we also assume g is not surjective (since if it is bijective, then we are done).

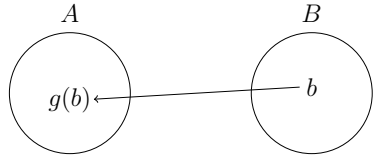
The main idea of the proof is that we partition both A and B into two pieces, and create a bijection between those pieces which, by the Pasting Together Theorem, will give us the bijection h:

$$\begin{array}{ccc} A & = & A_1 \cup A_2 \\ \downarrow{\scriptstyle h} & \downarrow{\scriptstyle h_1} & \downarrow{\scriptstyle h_2} \\ B & = & B_1 \cup B_2. \end{array}$$

The only information we have available comes from the two functions f and g that we have given to us. We must somehow use the functions f and g to make any progress on this problem. We might ask how these functions behave. Fix some element $a \in A$. Applying f, we have a new element $f(a) \in B$. We call a the *parent* of $f(a)$, and we call $f(a)$ the *child* of a. See the picture below.



We can also talk about elements of B being parents of children in A via the function g. The parent-child relationship in this case appears below.



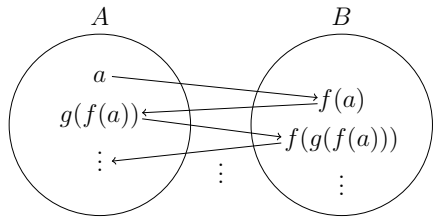
There are some very important facts we need to know about these parent-child relationships.

(1) Every element in A is the parent of exactly one child in B, because f is a function. Similarly, every element in B is the parent of exactly one child in A.
(2) Every element in A has at most one parent in B, because f is injective. Similarly, every element in B has at most one parent in A, because g is injective.
(3) Some element in A has no parent in B, because g is not surjective. Similarly, some element in B has no parent in A, because f is not surjective.

We can pass from a parent to a child, but then that child has its own child. In this way, we have a chain of *descendants*. By fact (1), there are no siblings in the list of descendants; there is always exactly one child per generation. For instance, if $a \in A$, its chain of descendants is

$$f(a), g(f(a)), f(g(f(a))), \ldots,$$

which is pictured below.



It is possible that this chain of descendants could loop around and repeat itself (for instance, we might have $g(f(a)) = a$, so that a is its own grandparent), but this situation will cause us no difficulty.

Just as we can consider descendants, we can also consider *ancestors*. The ancestors of an element are its parent, the parent of its parent, and so forth. By the important fact (2) above, there are never multiple parents in a generation. There is at most

one parent, and occasionally, by fact (3), there are no parents. Thus, sometimes the ancestry terminates, and sometimes it goes backwards forever (including when it loops).

The *genealogy* of an element x (of either A or B) is the chain consisting of its ancestors, descendants, and itself. If an element of a genealogy has no parent, we say it is the *beginning* of the genealogy. (Note that there is at most one beginning to any genealogy.) Genealogy chains can exhibit many different behaviors, but there is one important property that will help us in our proof. We say that genealogies come in two types:

(1) The genealogy does not begin in B. (This consists of the cases where the genealogy begins in A, or when the chain of ancestors never ends, including when it loops.)
(2) The genealogy does begin in B.

A key fact is that a parent and child are in the same chain, and so they have the same genealogy. Their genealogies thus have the same type.

We are now ready to describe a partition of A. We put $A = A_1 \cup A_2$ where

$$A_i = \{a \in A \,:\, \text{the genealogy of } a \text{ is of type } i\}.$$

Similarly, we put $B = B_1 \cup B_2$ where

$$B_i = \{b \in B \,:\, \text{the genealogy of } b \text{ is of type } i\}.$$

Now that we have a partition of A and a partition of B, to construct a bijection $h\colon A \to B$ it suffices by the Pasting Together Theorem to construct bijections $h_i\colon A_i \to B_i$ for each $i = 1, 2$. Each of these functions will be defined slightly differently.

Define $h_1\colon A_1 \to B_1$ by the rule $h_1(x) = f(x)$. To show that h_1 is a bijection requires us to prove three facts.

(1) We first show that h_1 really is a function taking elements of A_1 to B_1. To that end, let $x \in A_1$. This means that the genealogy of x is of type 1. The child of x is $f(x) \in B$, which has the same type of genealogy. Thus $h_1(x) = f(x) \in B_1$.
(2) We next show that h_1 is injective. Let $x_1, x_2 \in A_1$. Assume that $h_1(x_1) = h_1(x_2)$. From the definition of h_1, we have $f(x_1) = f(x_2)$. Since f is injective, $x_1 = x_2$. So we see that h_1 is injective.
(3) Finally, we show that h_1 is surjective. Let $y \in B_1$. We then know (from the definition of B_1) that y has a type 1 genealogy. In particular, y must have a parent $x \in A$, since its genealogy cannot begin at y in B. The parent, x, also has a type 1 genealogy, and so $x \in A_1$. Now, $h_1(x) = f(x) = y$, where that last equality comes from the fact that x is the parent of y. So we see that h_1 is surjective.

Having now dealt with h_1, we will next define h_2. To do this, we first define a function $j\colon B_2 \to A_2$ by the rule $j(y) = g(y)$. Mimicking the three parts above, it is straightforward (and left as Exercise 32.5) to show that j is a bijective function from B_2 to A_2. We let $h_2\colon A_2 \to B_2$ be its inverse, which is also a bijective function by Theorem 26.20(2).

This completes the sketched proof of the Schröder–Bernstein Theorem.

32.B Examples

The Schröder–Bernstein Theorem is not only beautiful symbolically, but also quite useful because it is sometimes very easy to describe injections back-and-forth between two sets A and B, yet it may be difficult to describe a bijection. Here are some standard examples.

Example 32.3. We will prove that the closed interval $[3, 10]$ has the same cardinality as $(0, 1)$.

Define $f\colon [3, 10] \to (0, 1)$ by the rule $f(x) = (x - 2)/10$. This is a linear function with $f(3) = 1/10$ and $f(10) = 8/10$. So it maps $[3, 10]$ into the interval $[1/10, 8/10] \subseteq (0, 1)$ injectively.

On the other hand, the function $g\colon (0, 1) \to [3, 10]$ given by $g(x) = x + 3$ is also an injection.

By the Schröder–Bernstein Theorem, we are done. △

This next example is so important that we will call it a theorem.

Theorem 32.4. $|\mathscr{P}(\mathbb{N})| = |\mathbb{R}|$.

Proof. In Exercise 31.6, we proved that $|\mathbb{R}| \leq |\mathscr{P}(\mathbb{N})|$. (For an alternative proof of this inequality, see Exercise 32.6 below.) By the Schröder–Bernstein Theorem, it suffices to now prove $|\mathscr{P}(\mathbb{N})| \leq |\mathbb{R}|$.

Define $f\colon \mathscr{P}(\mathbb{N}) \to \mathbb{R}$ by the rule $f(S) = 0.\chi_S(1)\chi_S(2)\chi_S(3)\ldots$. (For instance, if $S = \{1, 3, 4, 7, 9, \ldots\}$ then $f(S) = 0.101100101\ldots \in \mathbb{R}$.) It just remains to show that this function is injective. Let $S, T \subseteq \mathbb{N}$ be arbitrary, and assume $f(S) = f(T)$. Thus

$$0.\chi_S(1)\chi_S(2)\chi_S(3)\ldots = 0.\chi_T(1)\chi_T(2)\chi_T(3)\ldots.$$

Since neither decimal expansion involves repeating 9's, the two expansions are equal. Hence $\chi_S(n) = \chi_T(n)$ for each $n \in \mathbb{N}$. This means that S and T have exactly the same elements so $S = T$, which finishes showing that f is an injective function. □

32.C Exercises

Exercise 32.1. Let X, Y, and Z be sets. Prove that if $X \subseteq Y \subseteq Z$ and $|X| = |Z|$, then $|X| = |Y|$ as well.

Exercise 32.2. Prove that $A = [-5, 16)$ and $B = (0, \infty)$ have the same cardinality. (Hint: Find injections $A \to B$ and $B \to A$, and then use the Schröder–Bernstein Theorem. Alternatively, use Exercise 32.1 twice, with $X = (0, 1)$, $Z = \mathbb{R}$, and $Y = A$ or $Y = B$.)

Exercise 32.3. Prove that $A = \mathbb{Q} \cup [-2, -1] \cup \{\sqrt{2}\}$ and $B = [3, 6) \cup (15, 17)$ have the same cardinality.

Exercise 32.4. Given sets A and B, prove that if there is an injection $f\colon A \to B$ and a surjection $g\colon A \to B$, then $|A| = |B|$. (Hint: A previous homework exercise might be useful.)

Exercise 32.5. Prove the claim from Subsection 32.A that the function $j\colon B_2 \to A_2$ given by the rule $j(y) = g(y)$ is a bijective function. (Hint: Modify the three steps used to show that h_1 is a bijective function.)

Exercise 32.6. In Exercise 31.6 we showed that $|\mathbb{R}| \leq |\mathscr{P}(\mathbb{N})|$. Here is another way to do that.

Define a function $f\colon (0,1) \to \mathscr{P}(\mathbb{N})$, by sending (the decimal expansion of) a real number $0.a_1a_2a_3\ldots$ (not ending in repeating 9's) to the set

$$\{a_1, 10a_2, 100a_3, \ldots\} - \{0\} \subseteq \mathbb{N}.$$

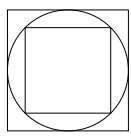
(For instance, $0.03193\ldots$ maps to $\{0, 30, 100, 9000, 30000, \ldots\} - \{0\}$.) Prove that this is an injective function.

Chapter IX

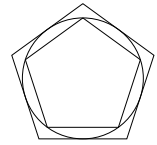
Introduction to Analysis

The only way to discover the limits of the possible is to go beyond them into the impossible. Arthur C. Clarke

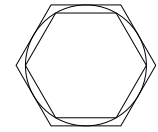
In the third century BC, the Greek mathematician Archimedes used the "method of exhaustion" to estimate the circumference of a circle of diameter 1, and thus estimate the value of π. His method involved inscribing a regular n-gon inside the circle, circumscribing the circle by a regular n-gon, and bounding the circumference of the circle between the perimeters of the two n-gons. For example, taking $n = 4$, $n = 5$, and $n = 6$, we get the following approximations.



$2.8284 < \pi < 4.0000$



$2.9389 < \pi < 3.6327$



$3.0000 < \pi < 3.464$

As n gets larger the approximations get better; for $n = 100$ we get $3.141 < \pi < 3.143$, and for $n = 1000$ we get $3.141587 < \pi < 3.141603$. This computation was among the first uses in antiquity of the idea of a limit; however, it would be nearly 2,000 years before the concept of limit was formally defined and given a logical foundation.

Newton and Leibniz used a concept of limit in the development of calculus, but it was not until around 1820 that Bolzano and Cauchy formalized the definition of limit. It was even later when it was finally written in the way most mathematicians now use limits.

33 Sequences

The infinite list of numbers

$$\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \dots$$

is an example of what we shall call a sequence. Sequences arise naturally in many contexts. For instance, you might measure the speed of a race car every second and produce a list of speeds. Or you could measure, over time, the temperature of heated metal. Maybe your list of numbers is the total population in a bacterial culture, measured every morning in the lab.

In all these cases, the numbers give us a brief glimpse at a process that could continue on forever. Extrapolating from the data, we might make a guess about how the sequence behaves, or perhaps fit it to a nice function. For instance, you might guess that the list of numbers above comes from the function $1/n$, as n ranges over the natural numbers (and you'd be right!).

A fundamentally important question we can ask is: Where are these numbers headed? Scientists use the mathematical theory developed in this section to determine the eventual behavior, or *limit*, of such sequences.

33.A What are sequences, exactly?

The formal definition of a sequence is as follows.

Definition 33.1. A *sequence* is a function $f\colon \mathbb{N} \to \mathbb{R}$. The image $f(n)$ of $n \in \mathbb{N}$ is called the nth *term* of the sequence. Often we will write the terms of a sequence with subscripts; i.e., we write them as

$$a_1, a_2, a_3, \dots$$

where $a_n = f(n)$.

Example 33.2. It is important to be able to pass back and forth between a list of numbers and the function that defines the list.

For example, define a function $f\colon \mathbb{N} \to \mathbb{R}$ by the rule $f(n) = 2n$. Thus, the nth term of our sequence is $a_n = 2n$, and the first few terms are given as follows: $a_1 = 2,\ a_2 = 4,\ a_3 = 6,\ a_4 = 8,\ a_5 = 10, \dots$.

On the other hand, if you are given the list of numbers $2, 4, 8, 16, \dots$, then you might guess that this sequence arises from the function $g\colon \mathbb{N} \to \mathbb{R}$ given by the rule $g(n) = 2^n$. △

Example 33.3. Try to figure out the rule for the following sequences:

(1) $-1, 1, -1, 1, -1, 1, \dots$,
(2) $1, 3, 5, 7, \dots$,
(3) $1, 0, 0, 0, 0, \dots$,
(4) $-10, 18397863, 2, 939, -10383, \dots$.

Here are some possible answers:
(1) $a_n = (-1)^n$,
(2) $a_n = 2n - 1$,
(3) $a_n = 0^{n-1}$ (or you could give a piecewise definition),
(4) Not enough information to find a pattern! △

Definition 33.4. If we wish to refer to a sequence $a_1, a_2, a_3, \ldots$, we may use the notation

$$(a_n)_{n\in\mathbb{N}}.$$

Example 33.5. Let $a_n = 2^n$. If we wish to refer to the sequence

$$a_1, a_2, a_3, \ldots$$

we could write $(a_n)_{n\in\mathbb{N}}$ or $(2^n)_{n\in\mathbb{N}}$. Both notations would refer to the sequence

$$2, 4, 8, \ldots. \qquad \triangle$$

33.B Arithmetic sequences

Each of the following sequences have something in common. See if you can discover what it is.

$$1, 3, 5, 7, 9, \ldots$$
$$13, 8, 3, -2, -7, \ldots$$
$$6, 9, 12, 15, 18, \ldots$$

Answer: Each term in the sequence is a fixed distance from the previous term. In the first sequence, the terms jump by adding 2, in the second sequence they jump by adding -5, and in the last sequence the terms jump by adding 3.

These types of sequences are so common we give them a special name.

Definition 33.6. Let $c, d \in \mathbb{R}$. For $n \in \mathbb{N}$, the sequence given by the formula

$$a_n = c + (n-1)d$$

is called an *arithmetic sequence* with first term c and common difference d.

Example 33.7. The arithmetic sequence with first term 2 and common difference 2 has terms

$$a_1 = 2,\ a_2 = 4,\ a_3 = 6,\ a_4 = 8,\ a_5 = 10, \ldots$$

The nth term is given by the formula $a_n = 2 + (n-1)2 = 2n$. △

Can you find the first six terms of the arithmetic sequence with first term π and common difference $-e$?

33.C Geometric sequences

There is a second very common type of sequence. Look to see if you can find what the following sequences have in common.

$$1, 2, 4, 8, 16, 32, \dots$$
$$1, 1/2, 1/4, 1/8, 1/16, \dots$$
$$3, -6, 12, -24, 48, \dots$$

Answer: Each term in the sequence is a fixed multiple of the previous term. In the first sequence we multiply each term by 2 to get the next term, in the second sequence we multiply by $1/2$, and in the third sequence we multiply by -2.

Definition 33.8. Let $c, r \in \mathbb{R}$. A sequence given by the formula

$$a_n = c \cdot r^{n-1}$$

for $n \in \mathbb{N}$ is called a *geometric sequence* with first term c and common ratio r.

Example 33.9. The geometric sequence with first term 4 and common ratio $1/10$ is given by the formula

$$a_n = 4\left(\frac{1}{10}\right)^{n-1}$$

and the first few terms are

$$4, 4/10, 4/100, \dots$$ △

33.D Sequences and their limits

In this subsection we will discuss what it means for a sequence to approach a limit. To give an example, consider the sequence $a_n = \frac{1}{n}$ given at the beginning of this section. This sequence begins

$$\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \dots$$

It appears that the limit should be 0; the terms are getting closer and closer to 0.

We will now give the formal definition of what it means for a sequence to approach a limit. This definition is quite complicated, so we will explain the true *meaning* behind the symbols afterwards.

Definition 33.10. Given a sequence $(a_n)_{n\in\mathbb{N}}$, we say that it *converges to* $L \in \mathbb{R}$ if the following holds:

$$\forall \varepsilon \in \mathbb{R}_{>0},\ \exists N \in \mathbb{R},\ \forall n \in \mathbb{N},\ n > N \Rightarrow |a_n - L| < \varepsilon. \tag{33.11}$$

In this case, we say that L is a *limit* of the sequence.

If a sequence $(a_n)_{n\in\mathbb{N}}$ converges to some $L \in \mathbb{R}$, we say that the sequence *converges* or that it is a *convergent sequence*. We write $\lim_{n\to\infty} a_n = L$.

If a sequence does not converge to any limit, we say that the sequence *diverges*, or that it is a *divergent sequence*.

We will spend the rest of this section studying this definition and coming to an understanding of what it means. We begin by peeling off each of the quantifiers.

What is ε? The first quantifier is "$\forall \varepsilon \in \mathbb{R}_{>0}$." The variable ε is used to help us measure how close the sequence gets to the limit L. We want to be able to prove that our sequence can get *arbitrarily* close to the limit L. Thus, we want to prove that our sequence eventually gets within a distance of $1/100$ of the limit, but also eventually within a distance of $1/1000$ of the limit, and eventually within $1/10000$, and so forth. Thus, we do not just take $\varepsilon = 1/100$, we allow it to be *any* positive constant.

What is N? The second quantifier is "$\exists N \in \mathbb{R}$." The variable N helps us tell how far along the sequence we must go, so that after that point the sequence stays within a distance of ε from the limit.

For instance, again consider the sequence $1, 1/2, 1/3, \ldots$. When $\varepsilon = 1/2$, how far along the sequence do we need to travel until it stays within a distance $1/2$ of the limit $L = 0$? We see that by the time we reach the second term, all of the rest of the terms are within a distance of $1/2$ from 0. When $\varepsilon = 1/100$, we now must take $N = 100$. For even smaller values of ε, we have to take larger values of N so that the sequence stays that close to the limit.

Note that the second quantifier is *existential*. This means that you must *fix* a specific value of N (depending on ε) which will satisfy the definition of limit. This value of N will usually be found in scratch work, outside the proof. This is often the hardest part of a limit proof, and we will show how this is to be done shortly.

What is n? The third and final quantifier is "$\forall n \in \mathbb{N}$." The variable n is just one of the subscripts in our sequence.

What does the premise of the implication, $n > N$, say? The condition $n > N$ just tells us that we will only look at the terms in the sequence past N. When looking at limits, we really only care about what *eventually* happens.

What does the conclusion of the implication, $|a_n - L| < \varepsilon$, say? The condition $|a_n - L| < \varepsilon$ is just an easy way of saying that the nth term of our sequence is within a distance of ε from the limit L. Equivalently, by removing the absolute value signs, we may write $L - \varepsilon < a_n < L + \varepsilon$.

Proving Limits. Every proof of a limit for a sequence will look essentially the same. First you must deal with each of the quantified variables. The universal (for all) variables must be left arbitrary. The existential variables must be fixed, but *only*

in terms of previous variables. The implication is usually dealt with directly.

Proof outline.
Let $\varepsilon > 0$.
Fix $N =$ found from scratch work $\in \mathbb{R}$.
Let $n \in \mathbb{N}$.
Assume $n > N$.
Do some work (usually by reversing the scratch work).
Conclude $|a_n - L| < \varepsilon$. □

We will show how this is accomplished, by proving the following result.

Theorem 33.12.
$$\lim_{n\to\infty} \frac{1}{n} = 0.$$

Scratch. This work should *not* appear on your homework, or in your proof.

Start with the conclusion $|a_n - L| < \varepsilon$. We know that $a_n = 1/n$ and $L = 0$. So the inequality becomes $|1/n - 0| < \varepsilon$. In other words $|1/n| < \varepsilon$. Since $1/n$ is positive, the absolute value signs disappear, and we have $1/n < \varepsilon$, or in other words $n > 1/\varepsilon$. This will be our value for N. ★

We are now ready to do the proof.

Proof. Let $\varepsilon > 0$. Fix $N = 1/\varepsilon \in \mathbb{R}$. Let $n \in \mathbb{N}$. Assume $n > N$. Thus, $n > 1/\varepsilon$. Taking reciprocals (noting that both sides of the inequality are positive), we get $1/n < \varepsilon$. Hence
$$|a_n - L| = |1/n - 0| = 1/n < \varepsilon$$
as desired. □

Let's do another example.

Proposition 33.13. *Given the sequence* $(a_n)_{n\in\mathbb{N}}$ *defined by* $a_n = 1 - \frac{3}{n}$, *then*
$$\lim_{n\to\infty} a_n = 1.$$

Scratch. Start with $|a_n - L| < \varepsilon$. Plugging in the values for a_n and L, we have
$$\left|\left(1 - \frac{3}{n}\right) - 1\right| < \varepsilon.$$
Simplifying we have $|-3/n| < \varepsilon$. Since $n > 0$ we have $|-3/n| = 3/n$. Hence, we may write $3/n < \varepsilon$. Solving for n, we get $n > 3/\varepsilon$. This is our value for N. ★

With the scratch work done, the proof is straightforward.

Proof. Let $\varepsilon > 0$. Fix $N = 3/\varepsilon \in \mathbb{R}$. Let $n \in \mathbb{N}$. Assume $n > N$. Thus $n > 3/\varepsilon$. Since ε and n are positive, we get $3/n < \varepsilon$. Thus

$$|a_n - L| = \left|\left(1 - \frac{3}{n}\right) - 1\right| = \left|-\frac{3}{n}\right| = \frac{3}{n} < \varepsilon$$

as desired. □

Warning 33.14. In the previous examples, each term of the sequence is closer to the limit L than the previous term. It is tempting to think that this is a valid definition of a limit; i.e., the terms get closer and closer to L without ever actually reaching L or getting farther away. In the next example, we will see that this is not true.

Proposition 33.15. *Define a sequence $(a_n)_{n\in\mathbb{N}}$ by the formula*

$$a_n = \begin{cases} \frac{1}{n} & \text{if } n \text{ is odd,} \\ 0 & \text{if } n \text{ is even.} \end{cases}$$

We have

$$\lim_{n\to\infty} a_n = 0.$$

This example is interesting because the sequence actually reaches the limit (every even numbered term is equal to the limit) and moves away from the limit infinitely often (each odd numbered term is farther from the limit than the previous term). However, it does *not* move too far away from the limit.

Scratch. Start with $|a_n - L| < \varepsilon$. There are two cases.

Case 1: Suppose n is odd. Then $a_n = 1/n$ so our inequality becomes $|1/n - 0| < \varepsilon$. Solving for n, as before, we reach $n > 1/\varepsilon$.

Case 2: Suppose n is even. Then $a_n = 0$ and our inequality becomes $|0 - 0| < \varepsilon$. This is true no matter the value of n, so in this case any value of N will work.

We must use an N that works in every case. Thus $N = 1/\varepsilon$ should suffice. ★

With our scratch work completed, the proof now follows.

Proof. Let $\varepsilon > 0$. Fix $N = 1/\varepsilon$. Let $n \in \mathbb{N}$. Assume $n > N$. We have two cases to consider.

Case 1: Suppose n is odd. In this case we have

$$|a_n - L| = |1/n - 0| = 1/n < 1/N = \varepsilon.$$

Case 2: Suppose n is even. In this case we have

$$|a_n - L| = |0 - 0| = 0 < \varepsilon.$$

Thus, in every case we have $|a_n - L| < \varepsilon$. □

33.E Divergence

Remember that for a sequence $a_1, a_2, a_3, \ldots$ to converge, we have:

$$\exists L \in \mathbb{R},\ \forall \varepsilon \in \mathbb{R}_{>0},\ \exists N \in \mathbb{R},\ \forall n \in \mathbb{N},\ n > N \Rightarrow |a_n - L| < \varepsilon.$$

Thus, to prove that the sequence is divergent we need to prove:

$$\forall L \in \mathbb{R},\ \exists \varepsilon \in \mathbb{R}_{>0},\ \forall N \in \mathbb{R},\ \exists n \in \mathbb{N},\ n > N \wedge |a_n - L| \geq \varepsilon.$$

As you can see, each of the four different quantifiers has been changed.

We start a divergence proof by letting $L \in \mathbb{R}$ be arbitrary. Subsequently, we must find some ε so that our sequence will continue to have some terms at least ε away from L. (Thus, ε usually depends on L.) We let $N \in \mathbb{R}$ be arbitrary, and must find a subscript n, past N, such that a_n has distance more than ε from L.

We will do one example, where ε does not depend on L.

Proposition 33.16. *Given the sequence $(a_n)_{n\in\mathbb{N}}$ defined by $a_n = (-1)^n$, then a_n is divergent.*

Proof. Let $L \in \mathbb{R}$. Fix $\varepsilon = 1/2 \in \mathbb{R}_{>0}$. Let $N \in \mathbb{R}$. To find a term in our sequence that is at least distance $1/2$ from L we consider two cases.

Case 1: Suppose $L \geq 0$. Fix $n \in \mathbb{N}$ to be the smallest odd number with $n > N$. Since n is odd, $a_n = -1$. We find

$$|a_n - L| = |(-1) - L| = L + 1 > 1/2 = \varepsilon.$$

Case 2: Suppose $L < 0$. In this case we fix $n \in \mathbb{N}$ to be the smallest even number with $n > N$. Since n is even, $a_n = 1$. We find

$$|a_n - L| = |1 - L| = 1 - L > 1/2 = \varepsilon$$

(since $L < 0$, we know $-L > 0$, hence $1 - L > 1$).

In every case, we fixed some $n \in \mathbb{N}$ satisfying both of the inequalities $n > N$ and $|a_n - L| > \varepsilon$. □

33.F One more limit trick

Sometimes when proving that a limit exists it will be useful to use one of the following functions to specify the value of N.

Definition 33.17. Let $a, b \in \mathbb{R}$. We define the *maximum* and *minimum* of these two numbers to be

$$\max(a,b) = \begin{cases} a & \text{if } a \geq b, \\ b & \text{if } a < b \end{cases} \quad \text{and} \quad \min(a,b) = \begin{cases} a & \text{if } a \leq b, \\ b & \text{if } a > b. \end{cases}$$

Example 33.18. Let $a = 5$ and $b = 6$. Then $\max(a, b) = 6$ and $\min(a, b) = 5$. Let $x \in \mathbb{R}$ and let $y = \max(x, 1)$ and $z = \min(x, 1)$. Then $y \geq 1$ and $z \leq 1$. △

Here is an example of a limit proof in which it is useful to specify N as the maximum of two numbers.

Proposition 33.19. *Define a sequence* $(a_n)_{n\in\mathbb{N}}$ *by the formula*

$$a_n = \frac{n+3}{2n-21}.$$

Then

$$\lim_{n\to\infty} a_n = \frac{1}{2}.$$

Scratch. Once again, we start our scratch work by considering $|a_n - L| < \varepsilon$, and try to solve for n. Thus, we want

$$\left|\frac{n+3}{2n-21} - \frac{1}{2}\right| < \varepsilon.$$

Finding a common denominator, we obtain

$$\left|\frac{2(n+3) - (2n-21)}{2(2n-21)}\right| < \varepsilon,$$

or in other words $27/|4n - 42| < \varepsilon$. Thus, we reduce to $|4n - 42| > 27/\varepsilon$.

There are two possibilities here, depending on whether $4n - 42$ is positive or negative. (Note: It is never zero since $n \in \mathbb{N}$.) In the positive case we want

$$4n - 42 > \frac{27}{\varepsilon},$$

which reduces (after some algebra) to

$$n > \frac{27 + 42\varepsilon}{4\varepsilon}. \tag{33.20}$$

In the case when $4n - 42$ is negative, we want

$$42 - 4n > \frac{27}{\varepsilon},$$

which reduces to

$$n < \frac{42\varepsilon - 27}{4\varepsilon}. \tag{33.21}$$

Inequality (33.21) does not help us because we want to find a *lower* bound on n. Thus, we must guarantee that case never happens, or in other words, we must guarantee that $4n - 42$ is positive. This means we want to take $4n - 42 > 0$, or in other words $n > 21/2$. Combining this with (33.20), we see that we want $N = \max(21/2, (27 + 42\varepsilon)/(4\varepsilon))$, so that if $n > N$ then n is greater than both $21/2$ and $(27 + 42\varepsilon)/(4\varepsilon)$. ★

We are now ready to begin our proof.

Proof. Let $\varepsilon > 0$. Fix $N = \max(21/2, (27 + 42\varepsilon)/4\varepsilon)$. Let $n \in \mathbb{N}$. Assume $n > N$. Using the fact that $N \geq 21/2$, we have $4n - 42 > 4N - 42 \geq 0$. We also find

$$4n - 42 > 4N - 42 \geq 4\frac{27 + 42\varepsilon}{4\varepsilon} - 42 = \frac{27}{\varepsilon}.$$

Hence, we find that

$$|a_n - L| = \left|\frac{n+3}{2n-21} - \frac{1}{2}\right| = \left|\frac{27}{4n-42}\right| = \frac{27}{4n-42} < \frac{27}{27/\varepsilon} = \varepsilon. \qquad \square$$

The trick used in the proof above is that we can use maximums to guarantee our sequence is far enough along so that it behaves in a certain way (in this case, we needed it to stay positive).

33.G Exercises

Exercise 33.1. Write the first six terms, and determine the nth term a_n, for each of the following sequences.

(a) An arithmetic sequence with first term 5 and common difference -3.
(b) A geometric sequence with first term 4 and common ratio 2.
(c) An arithmetic sequence with first term $1/2$ and common difference $3/4$.
(d) A geometric sequence with first term $3/5$ and common ratio $2/3$.

Exercise 33.2. Translate the following phrases into symbolic logic. (Your answer should include any necessary quantifications on the variables ε, N, and n.)

(a) The sequence $(a_n)_{n\in\mathbb{N}}$ defined by $a_n = 3 - 4/n$ converges to $L = 3$.
(b) The sequence $(a_n)_{n\in\mathbb{N}}$ defined by $a_n = 6$ does not converge to $L = 3$. (Note: This sequence does converge to $L = 6$.)

Exercise 33.3. Let $a, b, x \in \mathbb{R}$. Prove the following.

(a) $\max(a, b) \geq a$ and $\max(a, b) \geq b$.
(b) $\min(a, b) \leq a$ and $\min(a, b) \leq b$.
(c) If $x > \max(a, b)$ then $x > a$ and $x > b$.

Exercise 33.4. Prove that

$$\lim_{n\to\infty} \frac{2}{n^2} = 0.$$

Exercise 33.5. Prove that

$$\lim_{n\to\infty} \frac{3n-5}{2n+4} = \frac{3}{2}.$$

(Hint: When $n \in \mathbb{N}$, then $2n + 4$ is always positive, so you don't have to worry about when it is negative.)

Exercise 33.6. Prove or disprove: The sequence $(a_n)_{n\in\mathbb{N}}$ defined by $a_n = (n + 1)/n$ converges. (Hint: On scratch paper, write out the first ten terms to see if the sequence is going somewhere.)

Exercise 33.7. Let $(a_n)_{n\in\mathbb{N}}$ be an arithmetic sequence with first term c and common difference d. Prove that if $d = 0$, the sequence $(a_n)_{n\in\mathbb{N}}$ converges to c. (In other words, prove that the constant sequence $c, c, c, \ldots$ converges to c.)

Exercise 33.8. Prove that the sequence $(a_n)_{n\in\mathbb{N}}$ defined by $a_n = n$ does not converge to $L = 3$.

Exercise 33.9. Prove that $\lim_{n\to\infty}(\sqrt{n^2+1} - n) = 0$.

Exercise 33.10. Let $(a_n)_{n\in\mathbb{N}}$ be a geometric sequence with first term c and common ratio r. Prove the following statements.

(a) If $|r| < 1$, then a_n converges to 0.
(b) If $c \neq 0$ and a_n converges to 0, then $|r| < 1$.
(c) If $c > 0$ and $r > 1$, then a_n diverges.

(Feel free to use laws of logarithms, and especially the fact that if $a, b \in \mathbb{R}$ with $0 < a < b$, then $\ln(a) < \ln(b)$. In particular, for $0 < r < 1$ we have $\ln(r) < 0$, and for $r > 1$ we have $\ln(r) > 0$.)

34 Series

In this section we introduce the idea of adding infinitely many objects together. There are many applications for these ideas, which lead naturally into the development of the integral in calculus.

34.A What is a series?

We write $\sum_{n=1}^{\infty} a_n$ to denote the *series* $a_1 + a_2 + a_3 + \cdots$. But what does this really mean?

The numbers a_n form a sequence, which we call the *terms* or the *summands* of the series. From those terms, we can form an entirely *new* sequence

$$\begin{aligned} s_1 &= a_1 \\ s_2 &= a_1 + a_2 \\ s_3 &= a_1 + a_2 + a_3 \\ s_4 &= a_1 + a_2 + a_3 + a_4 \\ &\vdots \end{aligned}$$

which is called the sequence of *partial sums*. These sums are only the beginning portion of the series, which is why we call them partial sums. We can also define the partial sums using the more compact formula

$$s_{n+1} = s_n + a_{n+1} \tag{34.1}$$

which we sometimes will find useful.

We say that the series $\sum_{n=1}^{\infty} a_n$ *converges* to a sum S if $\lim_{n\to\infty} s_n = S$. In other words, the series converges when the sequence of partial sums converges. Thus, the methods of the previous section apply.

When first working with series it is important to note that the sequence of terms $a_1, a_2, a_3, \ldots$ is very different from the sequence of partial sums $s_1, s_2, s_3, \ldots$. For instance, it is possible that the terms converge but the partial sums diverge.

In the next two examples, we will investigate how to describe the partial sums when given a sequence of summands.

Example 34.2. Let $(a_n)_{n\in\mathbb{N}}$ be defined by $a_n = 1$. In other words, a_n is the constant sequence of 1's. We find that the partial sums for this sequence are

$$\begin{aligned} s_1 &= a_1 = 1 \\ s_2 &= a_1 + a_2 = 1 + 1 = 2 \\ s_3 &= a_1 + a_2 + a_3 = 1 + 1 + 1 = 3 \\ s_4 &= a_1 + a_2 + a_3 + a_4 = 1 + 1 + 1 + 1 = 4 \\ &\vdots \end{aligned}$$

We would guess that, in general, $s_n = n$. We will prove it by induction.

Proof. **Base case:** When $n = 1$, we have $s_1 = 1$.

Inductive step: Assume $s_k = k$ for some $k \geq 1$, with $k \in \mathbb{N}$. We want to show $s_{k+1} = k + 1$. Using (34.1), we have

$$s_{k+1} = s_k + a_{k+1} = s_k + 1 = k + 1,$$

since $s_k = k$ by the inductive hypothesis. □

In this example the sequence of terms $a_1, a_2, a_3, \ldots$ converges to 1. However, the sequence of partial sums $s_1, s_2, s_3, \ldots$ diverges. △

Example 34.3. Consider the sequence $(b_n)_{n\in\mathbb{N}}$ defined by the rule

$$b_n = \frac{1}{n} - \frac{1}{n+1}.$$

The first few terms in this sequence are given by

$$\begin{aligned} b_1 &= \frac{1}{1} - \frac{1}{2} = \frac{1}{2} \\ b_2 &= \frac{1}{2} - \frac{1}{3} = \frac{1}{6} \\ b_3 &= \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \\ b_4 &= \frac{1}{4} - \frac{1}{5} = \frac{1}{20} \\ &\vdots \end{aligned}$$

and it appears that this sequence is converging (fairly quickly) to 0.

Now consider the sequence of partial sums. The first few terms are computed to equal

$$\begin{aligned} s_1 &= b_1 = \frac{1}{2} \\ s_2 &= b_1 + b_2 = \frac{2}{3} \\ s_3 &= b_1 + b_2 + b_3 = \frac{3}{4} \\ s_4 &= b_1 + b_2 + b_3 + b_4 = \frac{4}{5} \\ &\vdots \end{aligned}$$

which appears to be converging to 1. Before we can prove convergence, we need more information about the sequence of partial sums.

It appears that the partial sums are given by the rule $s_n = 1 - \frac{1}{n+1}$. (Try it for the first four terms above.) We will now prove that this is correct, by induction.

Proof. **Base case:** When $n = 1$, then $s_1 = 1/2 = 1 - 1/2$.

Inductive step: Assume $s_k = 1 - 1/(k+1)$ for some $k \geq 1$, with $k \in \mathbb{N}$. We want to show $s_{k+1} = 1 - 1/(k+2)$. Using (34.1), we have

$$s_{k+1} = s_k + b_{k+1} = \left(1 - \frac{1}{k+1}\right) + \left(\frac{1}{k+1} - \frac{1}{k+2}\right) = 1 - \frac{1}{k+2}$$

as desired. □

Now that we know $s_n = 1 - 1/(n+1)$, we can prove $\lim_{n\to\infty} s_n = 1$. We will not include the scratch work, but here is the proof.

Proof. Let $\varepsilon > 0$. Fix $N = \frac{1}{\varepsilon} - 1 \in \mathbb{R}$. Let $n \in \mathbb{N}$. Assume $n > N$. We find

$$|s_n - S| = \left|1 - \frac{1}{n+1} - 1\right| = \frac{1}{n+1} < \frac{1}{N+1} = \varepsilon$$

as desired. □

This finishes our proof that $\sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1}\right) = 1$. Those who have taken calculus might recognize that this is a telescoping sum, which makes it much easier to simplify. △

Advice 34.4. To prove that a series converges try the following steps:
(1) Compute a few partial sums.
(2) Conjecture a general formula for the partial sums.
(3) Prove that formula by induction.
(4) Using your formula for the partial sums, find the limit.
(5) Finally, prove that the partial sums converge to that limit.

We will demonstrate how to prove convergence of series with one more example, leaving most of the work as Exercise 34.3.

Example 34.5. Consider the series $\sum_{n=1}^{\infty} \frac{1}{2^n}$. The first few partial sums work out to be $1/2, 3/4, 7/8, 15/16, \ldots$. By induction, we can prove that

$$s_n = 1 - \frac{1}{2^n}.$$

These partial sums converge to 1. △

We now prove a powerful result that can often be used to show that a series does not converge. We first state and prove the result in terms of convergence, and give the contrapositive (in terms of divergence) as a corollary. Before reading this proof, it might be helpful to review the triangle inequality (Theorem 8.21).

Theorem 34.6. *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real numbers. If the series*

$$\sum_{i=1}^{\infty} a_i$$

converges, then

$$\lim_{n\to\infty} a_n = 0.$$

Proof. We will write

$$s_n = \sum_{i=1}^{n} a_i$$

for the partial sums of the series. The theorem asserts that if $\lim_{n\to\infty} s_n$ exists, then $\lim_{n\to\infty} a_n = 0$. Note that for any $n > 1$, we have $s_n - s_{n-1} = a_n$.

Assume that the series converges. Then the sequence s_n of partial sums converges to some limit L.

Let $\varepsilon > 0$ be arbitrary. We wish to find an $N \in \mathbb{R}$ such that for all natural numbers $n > N$, we have $|a_n| < \varepsilon$.

Since $\varepsilon > 0$, we also have that $\varepsilon/2 > 0$. Hence, since the sequence s_n converges, there is some $M \in \mathbb{R}$ such that for all natural numbers $n > M$, we have $|s_n - L| < \varepsilon/2$. Taking $N = M + 1$, we see that if $n > N$, then both n and $n - 1$ are greater than M. Hence, for any $n > N$,

$$|s_n - L| < \varepsilon/2 \quad \text{and} \quad |s_{n-1} - L| < \varepsilon/2.$$

Therefore, for any $n > N$,

$$\begin{aligned}
|a_n| &= |s_n - s_{n-1}| \\
&= |(s_n - L) + (L - s_{n-1})| \\
&\leq |s_n - L| + |L - s_{n-1}| && \text{(by the triangle inequality)} \\
&= |s_n - L| + |s_{n-1} - L| \\
&< \varepsilon/2 + \varepsilon/2 \\
&= \varepsilon.
\end{aligned}$$

We thus see that for this value of N, it is true that for all $n > N$ we have $|a_n - 0| < \varepsilon$. Therefore

$$\lim_{n\to\infty} a_n = 0.$$

□

Corollary 34.7. *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real numbers. If*

$$\lim_{n\to\infty} a_n \neq 0$$

then

$$\sum_{i=1}^{\infty} a_i$$

does not converge.

Example 34.8. The series

$$\sum_{i=1}^{\infty} \frac{i+3}{2i-1}$$

does not converge, because

$$\lim_{n\to\infty} \frac{n+3}{2n-1} = \frac{1}{2} \neq 0$$

by Proposition 33.19. △

Example 34.9. The series

$$\sum_{i=1}^{\infty} (-1)^i$$

does not converge, since

$$\lim_{n\to\infty} (-1)^n$$

does not exist (see Proposition 33.16), and hence does not equal 0. △

Note that the converse of Theorem 34.6 does not hold; it is possible for

$$\lim_{n\to\infty} a_n = 0$$

to hold while

$$\sum_{i=1}^{\infty} a_i$$

does not converge. See Exercise 34.5 for an example.

34.B Exercises

Exercise 34.1. Consider the sequence $(a_n)_{n\in\mathbb{N}}$ given by the rule $a_n = n$. Find the first 6 terms of the sequence of partial sums s_n. Conjecture a simple formula for s_n and prove it.

Exercise 34.2. Let $c, d \in \mathbb{R}$ and let $(a_n)_{n\in\mathbb{N}}$ be the arithmetic sequence defined by $a_n = c+(n-1)d$ (i.e., the arithmetic sequence with first term c and common difference d). Find a formula for the nth partial sum $s_n = \sum_{k=1}^{n} a_k$ and prove it.

Exercise 34.3. Give a complete proof that $\sum_{n=1}^{\infty} 1/2^n = 1$, by filling in the missing details from Example 34.5. This should include:

(a) Giving a proof (by induction) that $s_n = 1 - 1/2^n$.

(b) Giving a proof that $\lim_{n\to\infty} s_n = 1$.

Exercise 34.4. Prove or disprove: The series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{3^n}$ converges.

Exercise 34.5. In this exercise we will show that the harmonic series $\sum_{k=1}^{\infty} \frac{1}{k}$ does not converge. Throughout the exercise, let $s_n = \sum_{k=1}^{n} \frac{1}{k}$ be the nth partial sum, for each integer $n \geq 1$.

(a) The main idea for showing that the harmonic series diverges is to break the series into pieces

$$1 + \underbrace{\boxed{\frac{1}{2}}}_{t_1} + \underbrace{\boxed{\frac{1}{3}+\frac{1}{4}}}_{t_2} + \underbrace{\boxed{\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}}}_{t_3} + \underbrace{\boxed{\frac{1}{9}+\frac{1}{10}+\cdots+\frac{1}{15}+\frac{1}{16}}}_{t_4} + \cdots,$$

where each box contains twice as many terms as the previous box.
Notice that

$$\begin{aligned} t_1 &= \frac{1}{2} \geq \frac{1}{2}, \\ t_2 &= \frac{1}{3}+\frac{1}{4} \geq \frac{1}{4}+\frac{1}{4} = 2\cdot\frac{1}{4} = \frac{1}{2}, \text{ and} \\ t_3 &= \frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8} \geq \frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8} = 4\cdot\frac{1}{8} = \frac{1}{2}, \end{aligned}$$

where the inequalities come from replacing terms with possibly smaller fractions. In general, for each $n \geq 1$ define

$$t_n = \sum_{k=2^{n-1}+1}^{2^{n-1}+2^{n-1}} \frac{1}{k} = \sum_{k=2^{n-1}+1}^{2^n} \frac{1}{k} = \frac{1}{2^{n-1}+1} + \cdots + \frac{1}{2^n}.$$

Prove that $t_n \geq \frac{1}{2}$ for each $n \geq 1$. (Hint: How many terms are being added? What is the smallest one?)

(b) Show that $s_{2^n} \geq 1 + \frac{n}{2}$, for each $n \geq 0$, by induction.
(Hint: For each $n \geq 0$ we have $s_{2^{n+1}} = s_{2^n} + t_{n+1}$.)

(c) Now show that the harmonic series does not converge.

35 Limits of functions

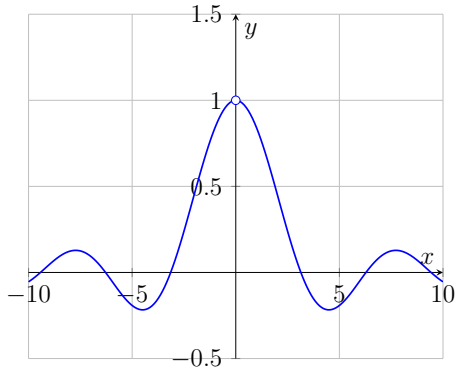
Let $f\colon \mathbb{R} - \{0\} \to \mathbb{R}$ be defined by

$$f(x) = \frac{\sin x}{x}.$$

If we evaluate this function for some values of x near 0, we find the following interesting behavior.

x	0.1	0.01	0.001
$f(x)$	0.998334	0.999983	0.999999

It appears that as x gets close to 0, the value of $f(x)$ gets close to 1. Note that we cannot just plug $x = 0$ into the function, since that would result in division by 0. A graph of the function $f(x)$ seems to confirm this behavior.



However, graphs can sometimes be misleading. Thus, in this section we will formalize what we mean when we say that a function approaches a value. This will become our definition of the limit of a function.

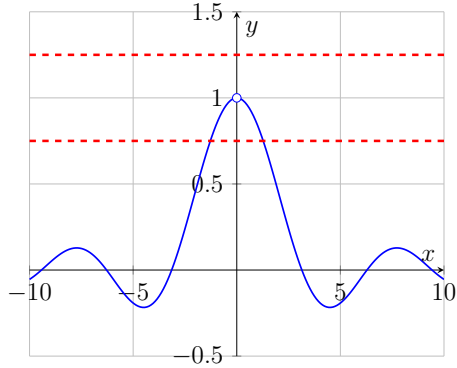
35.A Windows

In the example above, there are two related quantities we focus upon: the input x of the function, and the output $f(x)$. We are interested in the behavior of the outputs $f(x)$ as x approaches some *fixed* constant, which we might call the *point of interest*. In the previous example that point of interest is 0, but more generally we use the letter $a \in \mathbb{R}$ to describe the place where x is headed. Sometimes we write $x \to a$ as shorthand for the sentence "as x goes towards the point of interest a."

Similarly, we use the letter L to denote the limiting value that $f(x)$ approaches (if any) as $x \to a$. We might write $f(x) \to L$ to mean that "$f(x)$ is approaching the limit L."

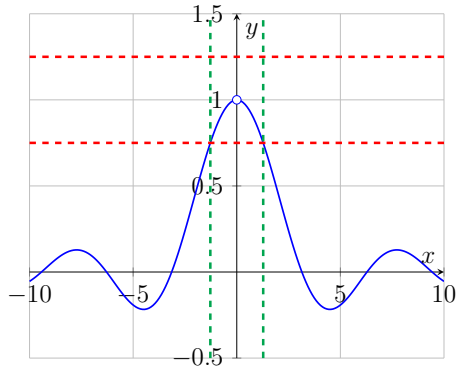
To formalize all of this, we start (just as with sequences) by letting $\varepsilon > 0$ denote some (positive, arbitrarily small) quantity that tells us how close our function should

be to the limit L. For instance, in the example above, when $\varepsilon = 1/4$ we want our function to stay between the two dashed lines in the graph below.



The top dashed line is $y = L + \varepsilon$ and the bottom dashed line is $y = L - \varepsilon$. We want the graph of $y = f(x)$ to stay between these two lines, but of course, as you can plainly see, the function does *not* stay between those two lines everywhere. However, it does stay between those two lines near the point of interest $a = 0$.

The next question we must answer is: How close must the input x get to the point of interest a so that the output $f(x)$ stays between those two lines? For example, let's add two (green) vertical dashed lines to our previous graph, which tell us exactly how far away from $a = 0$ the values of x can range before the function takes on values outside the band bordered by the (red) horizontal dashed lines.



Those vertical lines are (approximately) $x = 1.275698$ and $x = -1.275698$. Thus we see that x can vary as much as 1.275698 away from the point of interest, and our function stays within $\varepsilon = 1/4$ of the limit. If we take a smaller value of ε, then we have less room to vary.

Those two horizontal lines along with the two vertical lines give us a *window* (the rectangle in the middle). As we take smaller values of ε, that window should squeeze

in on the limit L. Notice that if the graph of our function has points between the two green lines that are also either above the top red line or below the bottom red line (i.e., directly above or below the window), then we have not placed the vertical lines correctly.

We let $\delta > 0$ be a variable which measures (given some ε) how far x can vary away from the point of interest, and still guarantee that $f(x)$ stays within a distance of ε from the limit value L. In other words, δ is some constant small enough that the vertical lines $x = a + \delta$ and $x = a - \delta$ together with the horizontal lines $y = L + \varepsilon$ and $y = L - \varepsilon$ produce a window for our function.

35.B Limit definition

We begin by defining certain subsets of $\mathbb{R}$.

Definition 35.1. A *deleted neighborhood* of $a \in \mathbb{R}$ is a subset of $\mathbb{R}$ of the form $(c, d) - \{a\}$, where $c < d$ are real, and $a \in (c, d) \subseteq \mathbb{R}$.

We can think of a deleted neighborhood as a set of real numbers that contains all the points "close to" a (but does not contain a itself).

Example 35.2. Let $a = 2$. Some deleted neighborhoods of a would include

$$(-10, 2.1) - \{2\}, \quad (1, 3) - \{2\}, \quad (1.9999999, 2.00000001) - \{2\}. \qquad \triangle$$

We are now equipped to give the formal definition of what we mean by a limit for functions. Afterwards, we will explain all of the notation.

Definition 35.3. Suppose we are given a point of interest $a \in \mathbb{R}$, and a possible limit value $L \in \mathbb{R}$, along with a function $f\colon S \to \mathbb{R}$, where S is a subset of $\mathbb{R}$ that includes some deleted neighborhood of a. We write

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

to mean that

$$\forall \varepsilon \in \mathbb{R}_{>0},\ \exists \delta \in \mathbb{R}_{>0},\ \forall x \in S,\ 0 < |x - a| < \delta \Rightarrow |f(x) - L| < \varepsilon.$$

In this case we say that the limit of the function f, as x approaches the point of interest a, is the real number L.

What is ε? As before, ε measures how close the function gets to the limit L. It is allowed to get arbitrarily small.

What is δ? The second quantifier is "$\exists \delta \in \mathbb{R}_{>0}$". The variable δ measures how close x must be to the point a in order for our function to stay within ε of the limit. It must be fixed in terms of ε.

What is S? The domain of the function f is S; it is the set of real numbers on which f is defined. We want $f(x)$ to be defined for all real numbers x that are "close to" a, but not necessarily at a. We do this by requiring that some deleted neighborhood of a is a subset of S. This reflects the idea that a limit depends on what happens near (but not at) a.

What is x? It is the variable we are plugging into our function f.

What does the premise of the implication, $0 < |x-a| < \delta$, say? It says two different things. First, the inequality $0 < |x - a|$ tells us that $x \neq a$. When defining limits, we do not care what $f(x)$ actually does at $x = a$, only what it does nearby. Second, the inequality $|x - a| < \delta$ tells us that x is allowed to range only a distance of δ away from the point of interest.

What does the conclusion of the implication, $|f(x) - L| < \varepsilon$, say? This condition is saying that the function value $f(x)$ is within a distance of ε from the limit L. Equivalently, by removing the absolute value signs, we may write $-\varepsilon < f(x) - L < \varepsilon$, or in other words, $L - \varepsilon < f(x) < L + \varepsilon$.

When handling limits, each proof should look nearly the same. We give the general outline of such a proof below. The reader should note how each quantifier is dealt with in turn, and the implication will be proved directly.

Proof outline.

Let $\varepsilon > 0$.
Fix $\delta = \boxed{\text{found in scratch work}} > 0$.
Let $x \in S$.
Assume $0 < |x - a| < \delta$.
Reverse the steps in scratch work.
Conclude that $|f(x) - L| < \varepsilon$. □

35.C Examples of limits

Some of the easiest limits are for linear functions.

Example 35.4. Let $f(x) = 3x + 2$, let $a = 2$, and let $L = 8$. (Note: If we plug $x = 2$ into f, we get $f(2) = 8$. In this case we do believe that the limit will also equal the function's value at the point of interest.) We will show that

$$\lim_{x \to 2} 3x + 2 = 8.$$

Scratch. Since this is the scratch work, we start with the conclusion $|f(x) - L| < \varepsilon$, and try to eventually get information about the quantity $|x - a| = |x - 2|$. We have

$$|f(x) - L| = |(3x + 2) - 8| = |3x - 6| = 3|x - 2|.$$

Thus we want $3|x - 2| < \varepsilon$, or in other words $|x - 2| < \varepsilon/3$. This tells us what value for δ we should use. ★

With the scratch work done, we can now give the formal proof.

Proof. Let $\varepsilon > 0$. Fix $\delta = \varepsilon/3 > 0$. Let $x \in \mathbb{R}$. Assume $0 < |x-2| < \delta$. We then find

$$|f(x) - L| = |(3x+2) - 8| = |3x - 6| = 3|x-2| < 3\delta = 3(\varepsilon/3) = \varepsilon$$

as desired. □

Limits for other linear polynomials will work similarly. △

Example 35.5. We find $\lim_{x\to -1} -2x + 3$.

Set $f(x) = -2x + 3$. Plugging in the point of interest $a = -1$, we find $f(-1) = (-2)(-1)+3 = 5$. Thus we would guess the limit is $L = 5$. We now prove it (without including our scratch work).

Proof. Let $\varepsilon > 0$. Fix $\delta = \varepsilon/2 > 0$. Let $x \in \mathbb{R}$. Assume $0 < |x-(-1)| < \delta$. We find

$$|f(x) - L| = |(-2x+3) - 5| = |-2x - 2| = 2|x+1| < 2\delta = 2(\varepsilon/2) = \varepsilon$$

as desired. □

When working with a linear function $f(x) = cx + d$ (for some constants $c, d \in \mathbb{R}$ with $c \neq 0$) the best value for δ should be $\delta = \varepsilon/|c|$. In this example we have $c = -2$ and indeed, the value for δ was $\varepsilon/|c|$. △

There is one last trick which will help us to evaluate limits. In limit proofs we have *one* assumption, namely

$$0 < |x - a| < \delta.$$

However, we also have some control on which δ we fix. If we guarantee that $\delta \leq 1$, then our assumption gives us $|x - a| < 1$, or in other words

$$a - 1 < x < a + 1.$$

This allows us to limit x a lot. (If we need an even smaller interval for x, we can take δ even smaller.) We will use this idea to find the limit of a quadratic function.

Proposition 35.6. $\lim_{x\to 1} x^2 + x = 2$.

We will include the scratch work, so you can see how this is done. (Normally, it should not appear in your proof.)

Scratch. Start with $|f(x) - L| < \varepsilon$. We have $f(x) = x^2 + x$ and $L = 2$, and so plugging in those values we want $|x^2 + x - 2| < \varepsilon$. If we factor the left side of the inequality, we get

$$|x - 1| \cdot |x + 2| < \varepsilon. \tag{35.7}$$

At this point, one might want to take $\delta = \varepsilon/|x+2|$, but δ must not depend on x. (Why?) To handle this issue, we bound $|x+2|$ as follows. Assuming $\delta \leq 1$, we get $|x-1| < \delta \leq 1$, and hence $-1 < x - 1 < 1$. Thus, by adding 3 throughout, we have $2 < x + 2 < 4$. Now $2 < |x+2| < 4$. Using this bound in (35.7), we see that we need $|x-1| < \varepsilon/4$. ★

We are now ready for the proof.

Proof. Let $\varepsilon > 0$. Fix $\delta = \min(1, \varepsilon/4) > 0$. Let $x \in \mathbb{R}$. Assume that we have $0 < |x-1| < \delta$. First, since $\delta \leq 1$, this tells us $|x-1| < 1$, so $-1 < x-1 < 1$. Therefore $2 < x+2 < 4$, and taking absolute values we get $|x+2| < 4$. Second, since $\delta \leq \varepsilon/4$ we find

$$|f(x) - L| = |(x^2 + x) - 2| = |x-1||x+2| < \delta \cdot 4 \leq (\varepsilon/4)4 = \varepsilon$$

as desired. □

We end this section with one final example of a limit proof.

Proposition 35.8.

$$\lim_{x \to 2} \frac{2x+1}{3x+2} = \frac{5}{8}.$$

Before reading the proof below, try to do the scratch work yourself, and see if it helps you figure out the choices made in the proof.

Proof. Let $\varepsilon > 0$. Fix $\delta = \min(1, 40\varepsilon) > 0$. Let $x \in \mathbb{R} - \{-2/3\}$. (Notice that we have to avoid $x = -2/3$ since the function is not defined there.) Assume $0 < |x-2| < \delta$. First, since $\delta \leq 1$, this tells us $|x-2| < 1$ hence $-1 < x-2 < 1$. Adding 2 we get $1 < x < 3$. Multiplying by 3 and adding 2, we get $5 < 3x+2 < 11$. Taking reciprocals, we find that $1/11 < 1/(3x+2) < 1/5$. Then taking absolute values, we have $1/|3x+2| < 1/5$. Second, since $\delta \leq 40\varepsilon$ we find

$$|f(x) - L| = \left|\frac{2x+1}{3x+2} - \frac{5}{8}\right| = \left|\frac{x-2}{8(3x+2)}\right| = \frac{|x-2|}{8 \cdot |3x+2|} < \frac{\delta}{8 \cdot 5} \leq \frac{40\varepsilon}{40} = \varepsilon. \qquad \square$$

Advice 35.9. When simplifying $|f(x) - L|$ you should expect that $|x-a|$ is one of the factors. That can help you simplify the expression (and also gives a quick double-check that you have not made an algebra mistake).

35.D Exercises

Exercise 35.1. Prove that

$$\lim_{x \to 4} 2x + 3 = 11.$$

Exercise 35.2. Let $a, c, d \in \mathbb{R}$. Prove that

$$\lim_{x \to a} cx + d = ca + d.$$

(Hint: Consider the cases in which $c = 0$ and $c \neq 0$, separately. Be sure that your proof properly handles the case where $c < 0$.)

Exercise 35.3. Prove that

$$\lim_{x\to 5} x^2 + 3x + 3 = 43.$$

Exercise 35.4. Prove that

$$\lim_{x\to 2} \frac{7x+4}{4x+1} = 2.$$

Exercise 35.5. Prove that

$$\lim_{x\to 3} x^3 + x^2 + 2 = 38.$$

36 Continuity

36.A Defining continuity

Intuitively, continuity for a function $f\colon \mathbb{R} \to \mathbb{R}$ means that there are no holes or jumps in the function. Put another way, if we focus on a point of interest $a \in \mathbb{R}$, we need $f(a)$ to be defined, and for x near a we want $f(x)$ to be near $f(a)$. Thus, there are two separate conditions that combine to give the formal definition of continuity.

Definition 36.1. Let $S \subseteq \mathbb{R}$. Given a function $f\colon S \to \mathbb{R}$ and a point of interest $a \in S$, we say that f is *continuous at* a if
(1) S includes an open interval around a and
(2) $\lim\limits_{x\to a} f(x) = f(a)$.
If a function is continuous at all points in its domain, we say that the function is *continuous* (everywhere that it is defined).

Example 36.2. Let $f\colon \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = cx + d$, where $c, d \in \mathbb{R}$. By Exercise 35.2, for any $a \in \mathbb{R}$ we have

$$\lim_{x\to a} f(x) = ca + d = f(a).$$

Hence, f is continuous at a for each $a \in \mathbb{R}$. Since f is continuous at all points in its domain, we can thus say that f is continuous. △

Example 36.3. Let $f\colon \mathbb{R} \to \mathbb{R}$ be the characteristic function of the set $[0, 1]$. Then

$$f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

One can check that f is continuous for all $a \in \mathbb{R} - \{0, 1\}$. We will show that f is not continuous at $x = 1$. (Determining continuity at other points is left as Exercise 36.3.)

Note that $f(1) = 1$. Then we need to show that

$$\lim_{x\to 1} f(x) \neq 1.$$

Hence, we need to show that

$$\exists \varepsilon > 0, \forall \delta > 0, \exists x \in \mathbb{R},\ (0 < |x - 1| < \delta) \wedge (|f(x) - 1| \ge \varepsilon).$$

Choose $\varepsilon = 1/2$. Let $\delta > 0$ be arbitrary, and fix $x = 1 + \delta/2$. We notice that $0 < |x - 1| = \delta/2 < \delta$ and, since $x > 1$,

$$|f(x) - 1| = |0 - 1| = 1 > \varepsilon.$$

Hence, f is not continuous at 1. △

Example 36.4. Let $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ be defined by $f(x) = x$. Note that 0 is in the domain of f, but the domain of f does not include any open interval around 0. Hence, f is not continuous at 0. Since 0 belongs to the domain of f, we see that f is a discontinuous function. (Some texts would say this function is continuous at 0, using an alternative definition of continuity for endpoints of intervals. This illustrates that one must carefully check definitions when verifying statements.) △

36.B Building new functions from old

We will now discuss several different ways to combine functions to form new functions.

Definition 36.5. Let $A, B \subseteq \mathbb{R}$, and let $f\colon A \to \mathbb{R}$ and $g\colon B \to \mathbb{R}$ be functions. We define a new function $f + g\colon A \cap B \to \mathbb{R}$ by the rule

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

We may also define a new function $fg\colon A \cap B \to \mathbb{R}$ by

$$(fg)(x) = f(x)g(x).$$

If we let $B' = \{x \in B : g(x) \neq 0\}$, we may define a function $f/g\colon A \cap B' \to \mathbb{R}$ by

$$(f/g)(x) = f(x)/g(x).$$

Example 36.6. Let $f(x) = x^2$ and let $g(x) = 2x + 1$ be functions defined on all real numbers. Then $(f + g)(x) = x^2 + 2x + 1$ and $(fg)(x) = x^2(2x + 1)$ are functions defined on all real numbers. Note that $g(x) = 0$ when $x = -1/2$. Hence, $(f/g)(x) = x^2/(2x + 1)$ is a function defined on $\mathbb{R} - \{-1/2\}$. △

36.C Limit laws

When we build a function by adding or multiplying two known functions, it turns out that we can prove that the limits of the functions interact nicely.

Theorem 36.7. *Let f and g be real-valued functions defined on a deleted neighborhood S of $a \in \mathbb{R}$. Suppose that $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$. Then*

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

Proof. Assume that $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$. We wish to show that $\lim_{x\to a}(f + g)(x) = L + M$. We will do this by choosing an arbitrary ε and finding a δ such that for all $x \in S$,

$$0 < |x - a| < \delta \Rightarrow |(f + g)(x) - (L + M)| < \varepsilon.$$

Let $\varepsilon > 0$. Then $\varepsilon/2 > 0$. Hence, since $\lim_{x\to a} f(x) = L$, there is some $\delta_1 > 0$ such that for all $x \in S$, we have the implication

$$0 < |x - a| < \delta_1 \Rightarrow |f(x) - L| < \varepsilon/2. \tag{36.8}$$

Similarly, since $\lim_{x\to a} g(x) = M$, there is some $\delta_2 > 0$ such that for all $x \in S$, we have

$$0 < |x - a| < \delta_2 \Rightarrow |g(x) - M| < \varepsilon/2. \tag{36.9}$$

Choose $\delta = \min(\delta_1, \delta_2)$. Let $x \in S$, and assume $0 < |x - a| < \delta$. Since $\delta \leq \delta_1$, and $\delta \leq \delta_2$, from (36.8) and (36.9) we get

$$|f(x) - L| < \varepsilon/2 \text{ and } |g(x) - M| < \varepsilon/2.$$

Now, using the triangle inequality we have

$$\begin{aligned} |(f+g)(x) - (L+M)| &= |f(x) + g(x) - L - M| \\ &= |(f(x) - L) + (g(x) - M)| \\ &\leq |f(x) - L| + |g(x) - M| \\ &< \varepsilon/2 + \varepsilon/2 \\ &= \varepsilon. \end{aligned}$$

Hence,

$$\lim_{x\to a}(f+g)(x) = L + M. \qquad \square$$

Theorem 36.10. *Let f and g be real-valued functions defined on a deleted neighborhood S of $a \in \mathbb{R}$. Suppose that $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$. Then*

$$\lim_{x\to a}(fg)(x) = LM.$$

Proof. We first do the proof under the assumption that one of L and M is not zero. Without loss of generality, we assume that $L \neq 0$.

Let $\varepsilon > 0$. Then $\varepsilon/(2|L|) > 0$, so, since $\lim_{x\to a} g(x) = M$, we can find a $\delta_1 > 0$ so that for any $x \in S$, if $0 < |x - a| < \delta_1$, we have $|g(x) - M| < \varepsilon/(2|L|)$. Note that this conclusion further implies that $|g(x)| < |M| + \varepsilon/(2|L|)$.

Now

$$\frac{\varepsilon}{2\left(\varepsilon/(2|L|) + |M|\right)} > 0,$$

so there is some δ_2 so that for $0 < |x - a| < \delta_2$, we have

$$|f(x) - L| < \frac{\varepsilon}{2\left(\varepsilon/(2|L|) + |M|\right)}.$$

Choose $\delta = \min(\delta_1, \delta_2)$. Let $x \in S$.

Now, assume that $0 < |x - a| < \delta$. Then $|x - a| < \delta_1$ and $|x - a| < \delta_2$. Hence, $|g(x) - M| < \varepsilon/(2|L|)$ and

$$|f(x) - L| < \frac{\varepsilon}{2\left(\varepsilon/(2|L|) + |M|\right)}.$$

We then have

$$\begin{aligned}
|(fg)(x) - LM| &= |f(x)g(x) - LM| \\
&= |f(x)g(x) - Lg(x) + Lg(x) - LM| \\
&= |(f(x) - L)g(x) + L(g(x) - M)| \\
&\leq |(f(x) - L)g(x)| + |L(g(x) - M)| \\
&= |f(x) - L||g(x)| + |L||g(x) - M| \\
&< \frac{\varepsilon}{2\left(\varepsilon/(2|L|) + |M|\right)}(|M| + \varepsilon/(2|L|)) + |L|\varepsilon/(2|L|) \\
&= \varepsilon/2 + \varepsilon/2 \\
&= \varepsilon.
\end{aligned}$$

Hence,

$$\lim_{x \to a}(fg)(x) = LM.$$

The proof in the case that both L and M are equal to 0 is much less complicated, and is left to the reader (see Exercise 36.4). □

We can also prove that the limit of f/g is the limit of f over the limit of g (provided that the limit of g is nonzero). We state the theorem here, but do not prove it.

Theorem 36.11. *Let f and g be real-valued functions defined on a deleted neighborhood S of $a \in \mathbb{R}$. Suppose that $\lim_{x \to a} f(x) = L$ and $\lim_{x \to a} g(x) = M$ with $M \neq 0$. Then*

$$\lim_{x \to a}(f/g)(x) = L/M.$$

From the previous three theorems, we can deduce the following result about continuity.

Theorem 36.12. *Let S be a subset of $\mathbb{R}$, and let $f\colon S \to \mathbb{R}$ and $g\colon S \to \mathbb{R}$ be real-valued functions that are continuous at some $a \in S$. Then $f + g$, and fg are continuous at a. If $g(a) \neq 0$, then f/g is continuous at a.*

Proof. We prove the theorem for the sum; the proofs for the product and the quotient are similar.

Suppose that f and g are continuous at a. Then

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

Hence, $f + g$ is continuous at a. □

Remark 36.13. Note that if f and g are continuous at all points in their domain, then so is their sum and their product. The quotient f/g will be continuous at all points of S at which g is nonzero. ▲

36.D Continuity of polynomials

In Exercise 35.2, you were asked to prove that any linear function (of the form $f(x) = cx + d$) is continuous at all real numbers. This implies that constant functions are continuous (a constant function is of the form $f(x) = d$; it is a linear function in which $c = 0$).

We now generalize this exercise to prove that every polynomial function is continuous.

Theorem 36.14. *Let $f\colon \mathbb{R} \to \mathbb{R}$ be a polynomial function; that is, a function of the form*

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

with each $a_i \in \mathbb{R}$. Then f is continuous at every real number.

Proof. We prove the theorem by induction on n, the exponent of the largest power of x involved in the expression for f.

Let $P(n)$ be the open sentence

$$\text{Every function of the form } g(x) = b_n x^n + \ldots + b_1 x + b_0 \text{ is continuous.}$$

Base Case: We have proved that $P(1)$ is true in Exercise 35.2.

Inductive step: Assume that $P(k)$ is true. We will show that $P(k+1)$ now follows.

Let $f(x) = a_{k+1}x^{k+1} + a_k x^k + \cdots + a_1 x + a_0$. Then both of the (functions given by the) polynomials x^k and $a_k x^k + \cdots + a_1 x + a_0$ are continuous by the inductive hypothesis, and $a_{k+1}x$ is continuous by Exercise 35.2. Hence, the product

$$a_{k+1}x^{k+1} = (a_{k+1}x)x^k$$

is continuous by Theorem 36.12, so the sum

$$f(x) = a_{k+1}x^{k+1} + \left(a_k x^k + \cdots + a_1 x + a_0\right)$$

is also continuous, again by Theorem 36.12. Hence, $P(k+1)$ is true. □

Example 36.15. The polynomial function given by the rule

$$f(x) = 3x^4 - 6x^2 + 2x - 1$$

is continuous at every real number. △

36.E Exercises

Exercise 36.1. Prove that the function $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ given by the rule $f(x) = \sqrt{x}$ is continuous at $a = 9$. (Hint: $\sqrt{x} - 3 = (\sqrt{x} - 3) \cdot \frac{\sqrt{x}+3}{\sqrt{x}+3} = \frac{x-9}{\sqrt{x}+3}$.)

Exercise 36.2. Prove that if a limit exists, the limit is unique.

Exercise 36.3. For f equal to the characteristic function of $[0, 1]$ as described in Example 36.3, prove the following.

(a) For each $a \in \mathbb{R} - \{0, 1\}$, the function f is continuous at a.
(b) The function f is not continuous at 0.

Exercise 36.4. Prove Theorem 36.10 in the case that both L and M are equal to 0.

Exercise 36.5. Let f be a function of the form

$$f(x) = \frac{g(x)}{h(x)},$$

where $g(x)$ and $h(x)$ are polynomials. Prove that f is continuous at all points where it is defined. (You may use Theorem 36.12.)

Index